TOWARDS RELIABILITY-GUIDED INFORMATION INTEGRATION IN AUDIO-VISUAL SPEECH RECOGNITION

Wentao Yu, Steffen Zeiler, Dorothea Kolossa[*]

Institute of Communication Acoustics, Ruhr University Bochum, Germany
{wentao.yu, steffen.zeiler, dorothea.kolossa}@rub.de

Abstract: Audio-visual speech recognition can improve the recognition rate in many small-vocabulary tasks. But for large vocabularies, due to difficulties like unsatisfactory lipreading accuracies, improving the recognition rate over audio-only baselines remains difficult. In this work, we propose a new fusion strategy, which fuses the state posteriors of separate stream recognizers through a bidirectional LSTM network. Our proposed fusion strategy outperforms all baselines as well as oracle dynamic stream-weighting, which gives a theoretical upper bound for dynamic stream-weighting approaches. The proposed system achieves a relative word error rate reduction of 42.18% compared to the audio-only setup and 34.73% compared to the non-oracle dynamic stream-weighting baseline.

1 Introduction

Audio-visual speech recognition (AVSR) was inspired by the natural ability of humans to integrate visual information in their speech perception. People often unconsciously focus on the speaker's lips in a noisy environment, and even in clean speech, seeing the lips of the speaker influences perception, as demonstrated by the McGurk effect [1]. Many machine AVSR systems exist, which mimic this ability and show promising results for small-vocabulary tasks [2, 3, 4]. However, for large-vocabularies, some words are virtually indistinguishable to a lipreading model, e.g. "do" and "to". These lead to a somewhat depressing performance of purely visual lipreading LV models and, hence, an inherent difficulty of AVSR on large-vocabulary tasks [5, 6].

In contrast to early and late integration, the idea of decision fusion, such as dynamic stream-weighting, provides an effective method to integrate the audio and video information. Different from decision fusion, which combines the decisions of multiple classifiers into a common decision, an alternative idea is to fuse *representations*, e.g. via multi-modal attentions [7]. Another example using this idea is that of gating, e.g. in [8] or in [9], where a newly designed *Gated Multimodal Unit* (GMU) is used for dynamically fusing feature streams within each cell of a network. Alternatively, [10] suggests the use of deep feedforward sequential memory networks (DFSMN) to firstly create and secondly fuse audio and video representations.

In this work, we combine the ideas of representation fusion and decision fusion, namely, using the posterior probabilities for states $\mathbf{s}$ and observation $\mathbf{o}_t$ of $i = 1 \dots M$ single-modality hybrid models $p(\mathbf{s}|\mathbf{o}_t^i)$ as our representation of the uni-modal streams. A wide variety of reliability indicators are extracted as auxiliary inputs to boost the integration model performance. To evaluate our proposed model, we use an open source large-vocabulary English dataset—the LRS2 corpus [11]—for all experiments. Our experimental results show that our proposed

[*]This project has received funding from the German Research Foundation DFG under grant number KO3434/4-2.

fusion strategy can notably improve the recognition rate of clean audio data from low-quality video streams.

In the following, we compare our proposed fusion strategy with different baseline models, which are introduced in Section 2. Section 3 describes the new fusion approach. The set of reliability measures that it employs are described in Section 4. Section 5 explains the experimental setup. In Sections 6 and 7, we discuss the experimental results and give an outlook on future work.

2 Related work

Many different fusion strategies can improve the recognition accuracy for AVSR. Here, we give a brief introduction of the fusion strategies that are used as baseline models in this work. In the following, we consider $M = 3$ single-modality models, one acoustic and two visual.

2.1 Early integration baseline

Early integration fuses the audio and visual information at the level of features via

$$\mathbf{o}_t = [(\mathbf{o}_t^{\mathrm{A}})^T, (\mathbf{o}_t^{\mathrm{VS}})^T, (\mathbf{o}_t^{\mathrm{VA}})^T]^T. \tag{1}$$

Here, superscript T denotes the transpose, $\mathbf{o}_t^{\mathrm{A}}$, $\mathbf{o}_t^{\mathrm{VA}}$, and $\mathbf{o}_t^{\mathrm{VS}}$ are audio features, shape-based, and appearance-based video features, respectively, described in more detail in Section 5.2.

2.2 Dynamic stream-weighting

Due to its high effectiveness, dynamic stream-weighting is often used as the fusion strategy in AVSR. Here, we therefore employ dynamic stream-weighting as in [12] as the second multi-modal baseline, carrying out a weighted combination of stream-wise DNN state posteriors according to

$$\log \widetilde{p}(s|\mathbf{o}_t) = \sum_i \lambda_t^i \cdot \log p(s|\mathbf{o}_t^i). \tag{2}$$

Here, the stream weights λ_t^i are estimated by a feedforward network from the extracted reliability measures. $p(s|\mathbf{o}_t^{\mathrm{i}})$ is the state posterior of state s at time t. For dynamic stream-weighting, we consider the mean square error (MSE) and the cross-entropy (CE) as loss functions.

2.3 Oracle weight baseline

We also consider oracle stream-weighting as a baseline. In the decoding phase, its dynamic stream weights are computed by minimizing the cross-entropy over the ground-truth forced alignment information, so a known text transcription of the test set is the prerequisite for this method. To optimize the cross-entropy in Eq. (4), we use convex optimization via CVX [13, 14]. The obtained oracle stream weights are then used to calculate the estimated log-posterior through Eq. (2). The oracle stream-weighting baseline gives the best achievable upper bound of recognition rate for a stream-weighting-based recognition system.

3 System overview

Figure 1 shows the proposed decision fusion network (DFN). The estimated log-posterior probability vector at time t is

$$\log \widetilde{p}(\mathbf{s}|\mathbf{o}_t) = \mathrm{DFN}([p(\mathbf{s}|\mathbf{o}_t^{\mathrm{A}})^T, p(\mathbf{s}|\mathbf{o}_t^{\mathrm{VA}})^T, p(\mathbf{s}|\mathbf{o}_t^{\mathrm{VS}})^T, \mathbf{R}_t^T]^T), \tag{3}$$

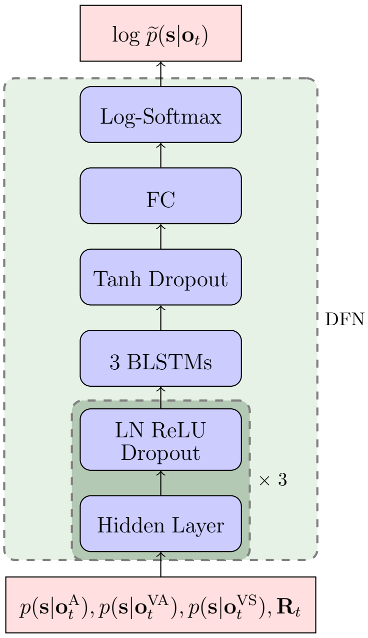
$\log \widetilde{p}(\mathbf{s}|\mathbf{o}_t)$
Log-Softmax
FC
Tanh Dropout
3 BLSTMs
LN ReLU Dropout
Hidden Layer
× 3
DFN
$p(\mathbf{s}|\mathbf{o}_t^{\mathrm{A}}), p(\mathbf{s}|\mathbf{o}_t^{\mathrm{VA}}), p(\mathbf{s}|\mathbf{o}_t^{\mathrm{VS}}), \mathbf{R}_t$

Figure 1 – Audio-visual fusion via the DFN, applied to one stream of audio and two streams of different video features

where $\mathbf{R}_t$ is the vector of reliability measures at time t, described in more detail in Section 4. The DFN architecture and training is introduced in Section 5.3.

The cross-entropy is used as the objective function to train the DFN:

$$\mathscr{L}_{\mathrm{CE}} = -\frac{1}{T}\sum_{t=1}^{T} p^*(\mathbf{s}|\mathbf{o}_t)^T \cdot \log \widetilde{p}(\mathbf{s}|\mathbf{o}_t), \tag{4}$$

where $p^*(\mathbf{s}|\mathbf{o}_t)$ is the one-hot vector of target state probabilities, obtained by forced alignment on the clean acoustic training data.

4 Reliability measures

In many studies [3, 4, 15], reliability information has shown its benefit for multimodal integration. In this work, we extract a broad range of reliability measures (see Table 1). They can be grouped into model-based and signal-based measures. Most of them are computed as in [12].

Model-based	Signal-based	
	Audio-based	Video-based
• Entropy • Dispersion • Posterior difference • Temporal divergence • Entropy and dispersion ratio	• MFCC • ΔMFCC • SNR • f_0 • Δf_0 • Probability of voicing	• Confidence • IDCT • Image distortion

Table 1 – Overview of reliability measures

The model-based measures are entropy, dispersion, posterior difference, temporal divergence, entropy- and dispersion-ratio. The audio signal based reliability measures comprise the first 5 MFCC coefficients with their temporal derivatives ΔMFCC. In this work, we use the deep learning approach DeepXi [17] to estimate the frame-wise SNR. The pitch f_0 and its temporal derivative, Δf_0, as well as the probability of voicing [18] are also included as reliability indicators. Signal-based reliability measures for the video data contain the Inverse Discrete Cosine Transform (IDCT) and image distortion estimates, which comprise the lighting condition, the degree of blurring, and the head pose. Same as in [12], OpenFace [16] is used for face detection and facial landmark extraction and the confidence of its face detector in each frame is also considered as a visual-signal quality indicator.

5 Experimental Setup

5.1 Dataset

The Oxford-BBC Lip Reading Sentences (LRS2) corpus is used as the audio-visual dataset. Table 2 gives an overview of the dataset size and partitioning. In clean conditions, acoustic speech recognition is far easier than lipreading, which only comes into its own as an auxiliary information source under some acoustic noise. To evaluate these contributions, we add acoustic noise to the LRS2 database. The MUSAN corpus [19] ambient subset is used as the noise source. Seven different SNRs are considered, from -9 dB to 9 dB in steps of 3 dB. We also generate data for a far-field AVSR scenario, via convolutions with measured impulse responses, again from the MUSAN corpus. Both these two augmentations use the standard Kaldi's Voxceleb example recipe.

Subset	Utterances	Vocabulary	Duration [hh:mm]
pre-train set	96,000	41,000	196:25
training set	45,839	17,660	28:33
validation set	1,082	1,984	00:40
test set	1,243	1,698	00:35

Table 2 – Size of subsets within the LRS2 Corpus

5.2 Feature extraction

The audio model uses 40 log Mel features together with two pitch features (f_0, Δf_0) and the probability of voicing, yielding 43-dimensional feature vectors. The audio features are extracted with a 25 ms frame size and 10 ms frameshift. The video frame is 40 ms long without overlap. The video appearance model (VA) uses 43-dimensional IDCT coefficients of the grayscale mouth region of interest (ROI) as features. The video shape model (VS) is based on the 34-dimensional non-rigid shape parameters described in [16]. Since the audio and video features have different frame rates, Bresenham's algorithm [20] is used to align the audio and video features. This algorithm gives the first-order approximation, which is suitable to the audio/video feature alignment problem.

5.3 Implementation details

All our experiments are based on the Kaldi toolkit [21]. Both pre-train and training sets are used together to train the acoustic and visual models. The initial HMM-GMM training follows the standard Kaldi AMI recipe. HMM-DNN training uses the nnet2 p-norm network [22] recipe. After the Kaldi HMM-DNN training, the number of states in each stream is identical to 3,856. We extract 41 reliability indicators in total. So in Figure 1, the input dimension is $(3 \times 3856 + 41)$ =11,609. The first three hidden layers have 8,192, 4,096, and 1024 units, respectively, each using the ReLU activation function and layer normalization (LN). They feed into 3 BLSTM layers with 1024 memory cells for each direction, using tanh as the activation function. The dropout rate in Figure 1 is 0.15. Finally, a fully connected (FC) layer projects the data to the output dimension of 3,856. A log-softmax function is applied to obtain the estimated log-posteriors. We conducted two experiments with the proposed DFN strategy. One is the BLSTM-DFN, exactly as described in Figure 1. The other is an LSTM-DFN, which replaces the BLSTM layers by LSTM layers.

To avoid overfitting, we checked for early stopping every 7,900 iterations, using the validation set. The training process is stopped, if the validation loss value does not decrease for 23,700 iterations (3 times early stopping checking). The initial learning rate is set to 0.0005 and is reduced by 20% whenever the validation loss does not decrease in early stopping checking. The batch size is set to 10. The DFN training is based on the PyTorch library [23] with the ADAM optimizer. The BLSTM-DFN or LSTM-DFN model training via CUDA on a GeForce RTX 2080 Ti GPU runs for approximately 15 days.

6 Results

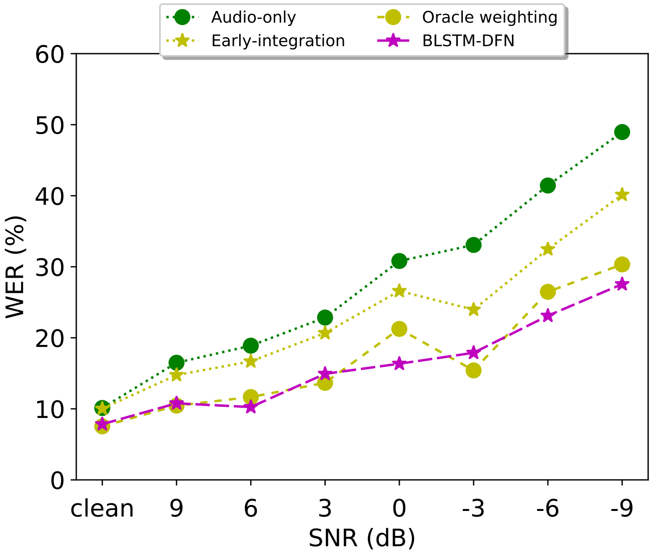

Figure 2 – WER (%) on the test set of the LRS2 corpus.

In this section, we compare the performance of different fusion strategies. Figure 2 compares the results of our proposed BLSTM-DFN with the audio-only model as well as the different baselines (more details can be found in Table 3). The proposed BLSTM-DFN can improve the WER for every SNR, even for clean acoustic data, comparing to the audio-only model. The model is also capable of achieving better results in many cases than the–realistically unachievable—oracle weighting, which can be considered as the upper limit for the dynamic stream-weighting approaches.

Table 3 lists all the results of our experiments under noisy conditions. The visual models show average WERs over 80%, which exemplifies that lip-reading is still difficult for large-vocabulary tasks. We assume that this is due to insufficient training data. We have also employed a spatio-temporal visual front-end [24] to extract high-level visual features, without seeing improvements. Early integration (EI) and dynamic stream-weighting (MSE and CE) can also improve the results. But comparing the average WER over all SNR conditions, it can be seen that the proposed BLSTM-DFN is greatly beneficial, outperforming not only all feasible stream integration approaches but even oracle weighting. As mentioned in Section 5.1, we also considered the case of far-field AVSR. The BLSTM-DFN still outperforms the other fusion strategies, but it is not as close to the OW. We suspect the reason is an insufficient amount of reverberant acoustic training signals. Overall, in this work, the BLSTM-DFN shows a relative WER reduction of 42.18 % compared to the audio-only system, while the LSTM-DFN yields a relative WER improvement of 27.09 %.

SNR	-9	-6	-3	0	3	6	9	clean	avg.	reverb.
AO	48.96	41.44	33.07	30.81	22.85	18.89	16.49	10.12	27.83	23.61
VA	85.83	87.00	85.26	88.10	87.03	88.44	88.25	88.10	87.25	88.10
VS	88.11	90.27	87.29	88.88	85.88	85.33	88.58	87.10	87.68	87.10
EI	40.14	32.47	23.96	26.59	20.67	16.68	14.76	10.02	23.16	19.15
MSE	46.48	37.79	27.45	27.47	19.52	16.58	15.09	9.42	24.98	19.54
CE	45.79	37.14	26.32	28.03	19.40	16.68	14.76	9.42	24.65	19.44
OW	30.33	26.47	**15.41**	21.25	**13.66**	11.66	**10.45**	**7.54**	17.10	**12.70**
LSTM-DFN	33.30	27.22	21.26	21.25	19.17	13.97	15.84	10.32	20.29	15.67
BLSTM-DFN	**27.55**	**23.11**	17.89	**16.35**	14.93	**10.25**	10.78	7.84	**16.09**	15.28

Table 3 – WER (%) on the LRS2 test set under additive noise. AO: audio-only model; VA: video appearance model; VS: video shape model; EI: early integration model; MSE: dynamic stream-weighting with mean square error as objective function; CE: dynamic stream-weighting with cross-entropy as objective function; OW: oracle stream-weighting baseline; LSTM-DFN: proposed model with LSTM layers; BLSTM-DFN: proposed model with bidirectional LSTM layers; reverb.: far-field AVSR results.

7 Conclusion

In this paper, we propose a new fusion strategy for audio-visual speech recognition, namely the decision fusion net (DFN), which considers the state posteriors of different streams as a representation for fusion. It uses reliability indicators to help in the estimation of the optimal combined state-posteriors. There are two flavors, a BLSTM-DFN with optimal performance, as well as an LSTM-DFN, which provides the option of real-time decoding. In experimental results on noisy as well as on reverberant data, our proposed model shows significant improvements, with the BLSTM version achieving a relative word-error-rate reduction of 42.18% over audio-only recognition and outperforming all baseline models. It is worth mentioning that the DFN is even superior to the oracle stream-weighting on average, which provides a theoretical upper bound for instantaneous stream-weighting approaches. The next goal of our work is to focus on end-to-end audio-visual speech recognition models.

References

[1] McGurk, H. and J. MacDonald: *Hearing lips and seeing voices*. *Nature*, 264(5588), pp. 746–748, 1976.

[2] Wand, M. and J. Schmidhuber: *Improving speaker-independent lipreading with domain-adversarial training*. In *Proc. Interspeech*. 2017.

[3] Meutzner, H., N. Ma, R. Nickel, C. Schymura, and D. Kolossa: *Improving audio-visual speech recognition using deep neural networks with dynamic stream reliability estimates*. In *Proc. ICASSP*, pp. 5320–5324. 2017.

[4] Gurban, M., J. Thiran, T. Drugman, and T. Dutoit: *Dynamic modality weighting for multi-stream HMMs in audio-visual speech recognition.* In *Proc. ICMI*, pp. 237–240. 2008.

[5] Thangthai, K. and R. Harvey: *Building large-vocabulary speaker-independent lipreading systems.* In *Proc. Interspeech*. 2018.

[6] Sterpu, G., C. Saam, and N. Harte: *How to teach DNNs to pay attention to the visual modality in speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28, pp. 1052–1064, 2020.

[7] Zhou, P., W. Yang, W. Chen, Y. Wang, and J. Jia: *Modality attention for end-to-end audio-visual speech recognition.* In *Proc. ICASSP*. 2019. URL `http://arxiv.org/abs/1811.05250`. `1811.05250`.

[8] Yu, J., S. Zhang, J. Wu, S. Ghorbani, B. Wu, S. Kang, S. Liu, X. Liu, H. Meng, and D. Yu: *Audio-visual recognition of overlapped speech for the lrs2 dataset.* In *Proc. ICASSP*, pp. 6984–6988. 2020.

[9] Arevalo, J., T. Solorio, M. Montes-y Gomez, and F. González: *Gated multimodal networks. Neural Computing and Applications*, pp. 1–20, 2020.

[10] Zhang, S., M. Lei, B. Ma, and L. Xie: *Robust audio-visual speech recognition using bimodal DFSMN with multi-condition training and dropout regularization.* In *Proc. ICASSP*, pp. 6570–6574. 2019.

[11] Afouras, T., J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman: *Deep audio-visual speech recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[12] Yu, W., S. Zeiler, and D. Kolossa: *Multimodal integration for large-vocabulary audio-visual speech recognition.* In *arXiv:2007.14223v1*. 2020. `arXiv:2007.14223`.

[13] Grant, M. and S. Boyd: *CVX: Matlab Software for Disciplined Convex Programming, version 2.1.* `http://cvxr.com/cvx`, 2014.

[14] Grant, M. and S. Boyd: *Graph implementations for nonsmooth convex programs.* In V. Blondel, S. Boyd, and H. Kimura (eds.), *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pp. 95–110. Springer-Verlag Limited, 2008. `http://stanford.edu/~boyd/graph_dcp.html`.

[15] Hermansky, H.: *Multistream recognition of speech: Dealing with unknown unknowns. Proceedings of the IEEE*, 101(5), pp. 1076–1088, 2013.

[16] Amos, B., L. Bartosz, and M. Satyanarayanan: *OpenFace: A general-purpose face recognition library with mobile applications.* Tech. Rep., CMU-CS-16-118, CMU School of Computer Science, 2016.

[17] Nicolson, A. and K. Paliwal: *Deep learning for minimum mean-square error approaches to speech enhancement. Speech Communication*, 111, pp. 44–55, 2019.

[18] Ghahremani, P., B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur: *A pitch extraction algorithm tuned for automatic speech recognition.* In *Proc. ICASSP*, pp. 2494–2498. 2014.

[19] SNYDER, D., G. CHEN, and D. POVEY: *MUSAN: A Music, Speech, and Noise Corpus*. 2015. ArXiv:1510.08484v1, 1510.08484.

[20] SPROULL, R.: *Using program transformations to derive line-drawing algorithms*. *ACM Transactions on Graphics (TOG)*, 1(4), pp. 259–273, 1982.

[21] POVEY, D., A. GHOSHAL, G. BOULIANNE, L. BURGET, O. GLEMBEK, N. GOEL, M. HANNEMANN, P. MOTLICEK, Y. QIAN, P. SCHWARZ ET AL.: *The Kaldi speech recognition toolkit*. In *Proc. IEEE*. IEEE Signal Processing Society, 2011.

[22] ZHANG, X., J. TRMAL, D. POVEY, and S. KHUDANPUR: *Improving deep neural network acoustic models using generalized maxout networks*. In *Proc. ICASSP*, pp. 215–219. 2014.

[23] PASZKE, A., S. GROSS, F. MASSA, A. LERER, J. BRADBURY, G. CHANAN, T. KILLEEN, Z. LIN, N. GIMELSHEIN, L. ANTIGA ET AL.: *Pytorch: An imperative style, high-performance deep learning library*. In *Advances in neural information processing systems*, pp. 8026–8037. 2019.

[24] STAFYLAKIS, T. and G. TZIMIROPOULOS: *Combining residual networks with LSTMs for lipreading*. *arXiv preprint arXiv:1703.04105*, 2017.