Multimodal Integration for Large-Vocabulary 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*—For many small- and medium-vocabulary tasks, audio-visual speech recognition can significantly improve the recognition rates compared to audio-only systems. However, there is still an ongoing debate regarding the best combination strategy for multi-modal information, which should allow for the translation of these gains to large-vocabulary recognition. While an integration at the level of state-posterior probabilities, using dynamic stream weighting, is almost universally helpful for small-vocabulary systems, in large-vocabulary speech recognition, the recognition accuracy remains difficult to improve. In the following, we specifically consider the large-vocabulary task of the LRS2 database, and we investigate a broad range of integration strategies, comparing early integration and end-to-end learning with many versions of hybrid recognition and dynamic stream weighting. One aspect, which is shown to provide much benefit here, is the use of dynamic stream reliability indicators, which allow for hybrid architectures to strongly profit from the inclusion of visual information whenever the audio channel is distorted even slightly.**

***Index Terms*—Audiovisual Speech Recognition, Multi-modal Integration, Dynamic Stream Weighting**

I. Introduction

Large Vocabulary Continuous Speech Recognition (LVCSR) remains difficult as a lipreading task, because many pairs of phonemes correspond to a single viseme, making many pairs of words almost indistinguishable to a vision-only system, as for example "do" and "to". Due to this intrinsic difficulty, an integration of lipreading into speech recognition becomes difficult in large- or open-vocabulary applications [1]. Nonetheless, lip-reading gives a great benefit to human listening [2]. In this work, we use an exemplary large-vocabulary dataset - the LRS2 corpus described in [3], to test whether and how similar benefits are attainable for automatic systems.

Many studies have shown that video information can dramatically improve small-vocabulary speech recognition performance, when the audio signals are recorded in a noisy environment. Often, stream weighting proved to be an effective method to combine audio and video information. As in [4], separate models are then trained for each of the modalities, and possibly, for each of the feature streams per modality. Stream weighting is realized through a weighted combination of the DNN state posteriors of each modality

$$\log \widetilde{p}(s|\mathbf{o}_t) = \sum_i \lambda_t^i \cdot \log\, p(s|\mathbf{o}_t^i), \tag{1}$$

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

where $\log\, p(s|\mathbf{o}_t^i)$ is the log-posterior of state s in stream i and $\log \widetilde{p}(s|\mathbf{o}_t)$ is its estimated combined log-posterior. The stream weights λ_t^i are typically predicted by appropriate reliability measures. In most of the state-based multi-modal integration studies, only two streams with few reliabilities are used. For example, in [5] the weights are only estimated from an entropy estimate. In this work, we consider a broad range of possible reliability metrics and apply them to fuse the information of three models, one acoustic and two visual.

Overall, we compare the performance of three different integration methods. The paper is organized as follows: Section II discusses the differences between end-to-end and hybrid speech recognition models. Early integration and state-based integration are discussed in Section III. Different reliability indicators are introduced in Section IV. Section V explains the experimental setup, while Section VI shows the results. Finally, in Section VII, we discuss the overall performance of all systems and give an outlook on future work.

II. End-to-end vs. hybrid models

In recent years, end-to-end acoustic speech recognition has quickly gained widespread popularity. In its original form, this model predicts character sequences directly from the audio signal. Different from the end-to-end model, in a hybrid unimodal speech recognition system, an acoustic model is trained to calculate log-pseudo-posteriors $\log\, p(s|\mathbf{o}_t)$. A decoder then uses these pseudo-posteriors to obtain the best word sequence by graph search through a language model [6].

While the hybrid model has the disadvantage of higher complexity compared to end-to-end learning, many recent papers have still shown superior performance of hybrid ASR compared to end-to-end recognition, for example in [7].

For our application of multi-modal recognition, hybrid frameworks are advantageous for multi-modal fusion, because they allow for an integration at the level of the pseudo-posteriors, and for using reliability information, which has proven beneficial for multimodal integration in many studies, see e.g. [8]–[10]. End-to-end models, in contrast, typically use an attention mechanism rather than reliabilities for the multi-modal fusion, which is also the case in the baseline models that we consider [11], [12]. Recently, [13] has extended the work of [11] and improved the performance by using a loss function that explicitly considers facial action units.

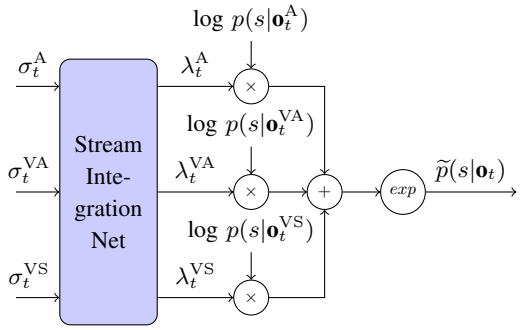

Fig. 1: Audio-visual fusion strategy for audio and two streams of different video models

III. System overview

A. System framework

As shown in Fig. 1, a set of reliability measures σ_t^{A}, σ_t^{VA}, σ_t^{VS} (described in more detail in Sec. IV) is used as the input for the stream integration net, which uses these to obtain weights λ_t^i for all streams over time. We then fuse the audio and video models through Eq. (1).

For training the stream integration net, we carry out forced alignment on the clean audio training set to obtain target state sequences $p^*(s|\mathbf{o}_t)$, in which the reference state probability is set to one and any other state probabilities are set to zero. Two state-based loss functions are employed here, the cross-entropy (CE)

$$\mathrm{CE} = -\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{S} p^*(s|\mathbf{o}_t)\cdot \log\, \widetilde{p}(s|\mathbf{o}_t), \tag{2}$$

and the mean squared error (MSE)

$$\mathrm{MSE} = \frac{1}{T\cdot S}\sum_{t=1}^{T}\sum_{s=1}^{S}(p^*(s|\mathbf{o}_t) - \widetilde{p}(s|\mathbf{o}_t))^2. \tag{3}$$

We also consider maximum mutual information (MMI) as a sequence-based criterion [14] via

$$\frac{\partial F_{MMI}}{\partial \log\, \widetilde{p}(s|\mathbf{o}_t)} = \kappa(p^*(s|\mathbf{o}_t) - \gamma_t^{DEN}(s)). \tag{4}$$

Here, $\gamma_t^{DEN}(s)$ is the posterior probability of state s at time t, computed over the denominator lattices that are obtained from the state pseudo-posteriors $\widetilde{p}(s|\mathbf{o}_t)$. κ is the acoustic scaling factor and set to 1.0.

B. Oracle weight baseline

To get an estimate of the best achievable word error rate (WER), we use convex optimization via CVX [15], [16] to optimize the cross-entropy in Eq. (2), which yields a set of oracle stream weights.

C. Early integration baseline

Early integration fuses audio and video information directly at the input of the system, using stacked feature vectors via

$$\mathbf{o}_t = [(\mathbf{o}_t^{\mathrm{A}})^T, (\mathbf{o}_t^{\mathrm{VS}})^T, (\mathbf{o}_t^{\mathrm{VA}})^T]^T \tag{5}$$

where $\mathbf{o}_t^{\mathrm{A}}$ are audio features, $\mathbf{o}_t^{\mathrm{VS}}$ are shape-based video features, and $\mathbf{o}_t^{\mathrm{VA}}$ are appearance-based video features, described in more detail in Sec. V-C, and T denotes vector transposition. As the audio and video features have different frame rates, we use a digital differential analyzer, similar to Bresenham's algorithm [17] to synchronize the video features before applying Eq. (5).

D. End-to-end baselines

In addition to the early integration baseline, we compare the performance of our suggested hybrid audiovisual ASR to end-to-end models with attention mechanisms [11], [12], which offer an alternative approach to multimodal fusion. As described in [12], both audio and video encoders are LSTM networks. The decoder is an LSTM transducer [18], which uses the encoded audio and video sequences, either via a dual attention mechanism [12], or, in [11], using a multi-head attention mechanism that is specifically optimized towards audio-visual integration.

IV. Reliability measures

As shown in Tab. I, the proposed reliability measures can be grouped into the ones that are model-based (**MB**) and the signal-based (**SB**) measures. For the model-based measures, the audio and video models are considered separately. The signal-based measures can also be subdivided into audio-based (**AB**) and video-based (**VB**) measures.

TABLE I: Proposed reliability measures

Model-based (**MB**)	Signal-based (**SB**)	
	Audio-based (**AB**)	Video-based (**VB**)
• Entropy • Dispersion • Posterior difference • Temporal divergence • Entropy and dispersion ratio	• MFCC • ΔMFCC • SNR • Signal and noise energy • Soft VAD	• IDCT • Image distortion

A. Model-based reliability measures

The ***entropy*** is a proxy for the model's uncertainty about the state s, given the current observation $\mathbf{o}_t$. It is calculated for each stream i via

$$\mathrm{H}_t^i = -\sum_{s=1}^{S^i} p(s|\mathbf{o}_t^i)\cdot \log\, p(s|\mathbf{o}_t^i), \tag{6}$$

with S^i as the number of states in stream model i.

Similarly, the ***dispersion*** is related to the decoder's discriminative power. It is computed by:

$$\mathrm{D}_t^i = \frac{2}{K(K-1)}\sum_{l=1}^{K}\sum_{m=l+1}^{K} \log\, \frac{\widehat{p}(l|\mathbf{o}_t^i)}{\widehat{p}(m|\mathbf{o}_t^i)}, \tag{7}$$

where the probabilities p are sorted in descending order to obtain $\widehat{p}$. K is set to 15.

The K-largest ***posterior difference***, defined via

$$\mathrm{Diff}_t^i = \frac{1}{K-1}\sum_{k=2}^{K} \log\, \frac{\widehat{p}(1|\mathbf{o}_t^i)}{\widehat{p}(k|\mathbf{o}_t^i)}, \tag{8}$$

is also considered, showing the average ratio between the largest posterior and the next $K-1$ values.

The ***temporal divergence*** is computed as the Kullback-Leibler divergence between two posterior vectors $p(s|\mathbf{o}_t^i)$ and $p(s|\mathbf{o}_{t+\Delta t}^i)$, i.e.

$$\mathrm{Div}_{\Delta t}^i(t) = \mathrm{D}_{KL}(p(s|\mathbf{o}_t^i)||p(s|\mathbf{o}_{t+\Delta t}^i)). \tag{9}$$

Δt is set to 250 ms. As described in [19], the mean of the temporal divergence is also an interesting measure of reliability and it is used here by averaging $\mathrm{Div}_{\Delta t}^i(t)$ over segments of 50 ms length.

The ***entropy ratio*** is described in [20]. The strongly related ***dispersion ratio*** $\omega_{D,t}^i$ is estimated based on the average dispersion $\overline{\mathrm{D}}_t$

$$\omega_{D,t}^i = \frac{\widetilde{\mathrm{D}}_t^i}{\sum_{k=\mathrm{A,VA,VS}} \widetilde{\mathrm{D}}_t^k}, \tag{10}$$

where

$$\widetilde{\mathrm{D}}_t^i = \begin{cases} \frac{1}{10000} & \mathrm{D}_t^i < \overline{\mathrm{D}}_t \\ \mathrm{D}_t^i & \mathrm{D}_t^i \geq \overline{\mathrm{D}}_t. \end{cases} \tag{11}$$

A, VA, and VS represent the audio, video appearance and video shape stream, respectively. D_t^i is obtained from Eq. (7).

B. Signal-based reliability measures

The first 5 ***MFCC*** coefficients and their temporal derivatives, Δ***MFCC***, are related to the audio quality.

The estimated Signal-to-Noise Ratio (***SNR***) also represents the quality of the audio signal; it is computed in each frame

$$\mathrm{SNR_t} = 10 \log \left(\frac{\mathrm{S}_t}{\mathrm{N}_t} \right). \tag{12}$$

The ***signal energy*** S_t is estimated as the sum of squared amplitudes of the Hamming-windowed frame t. The ***noise energy*** N_t is estimated by a variant of the minima-controlled recursive averaging algorithm (MCRA-2) [21].

The ratio between the energy of the speech band and the total energy of each frame is used as a soft voice-activity detection (***soft VAD***) cue.

The first 5 Inverse Discrete Cosine Transform (***IDCT***) coefficients of the mouth region represent low-level image properties.

The ***image distortion*** measures comprise the lighting condition, the degree of blurring and the head pose, all computed as in [22]. The lighting condition represents the mean brightness of the image. The degree of blurring is estimated as the variance of the image after high-pass filtering. To obtain an indicator for head rotation and tilt, the cross-correlation between the original image and its horizontally mirrored version is computed.

V. Experimental setup

A. Dataset

Our experiments are based on the LRS2 corpus. The training set contains 45,839 spoken sentences and 17,660 words, with a test set of 1,243 sentences and 1,698 words. To analyze the performance in different acoustic noise conditions, we have artificially created noisy versions of the LRS2 database. The augmentation recipe from Kaldi's Voxceleb example is employed for this purpose, using the MUSAN corpus [23] as the noise dataset. It contains 3 different kinds of noise, i.e. ambient noise, music and babble noise. Seven different SNRs are selected, from -9 dB to 9 dB in steps of 3 dB. Each audio signal is augmented with these three noise types and the SNR is randomly chosen from all SNRs.

B. Implementation details

All hybrid recognition experiments are carried out using the Kaldi toolkit [24], with the training set of the LRS2 corpus employed in the training of the acoustic and visual models. The initial HMM-GMM training follows the standard Kaldi AMI recipe; subsequent HMM-DNN training uses the nnet2 recipe. The output dimension of all three models is 3784. For performance reasons, the audio model alignments are also used for the HMM-DNN training of both video models. The integration model has 5 hidden layers, with 43, 25, 17, 10 and 3 units, respectively, each using ReLU activation functions. The output is the predicted weight of each stream, λ_t^i. A sigmoid function is used to limit the weights to values between 0 and 1. Finally, we normalize the multi-modal posterior probabilities to one at each time t. To avoid overfitting, early stopping is used if the training loss does not improve for 1200 iterations. The stream integration model is trained on the development set and performance is tested on the evaluation set.

C. Feature extraction

The audio model uses 13-dimensional MFCCs as features. MFCCs are extracted with 25 ms frame size and 10 ms frame shift. The video frame is 40 ms long without overlap. The mouth region is detected via OpenFace [25]. The video appearance model (VA) uses 43-dimensional IDCT coefficients of the mouth region in the grayscale image as features. The video shape model (VS) uses the 34-dimensional non-rigid shape parameters [25] as features.

VI. Results

Here, we provide the performance comparisons between our proposed hybrid model and the end-to-end AVSR models, which are described in [12] and [11].

Fig. 2 shows the performance of all considered baseline hybrid models (more details in Tab. II). The audio-only model (**AO**) has a much better performance than the video-shape (**VS**) and video-appearance (**VA**) models alone. Early integration (**EI**) can already improve the WER at lower SNR conditions ($\leqslant$ 0 dB), but there is no improvement, if we compare the average WER over all SNRs. As the oracle weight

model (**OW**) shows, there is much room for improvement through optimal stream integration.

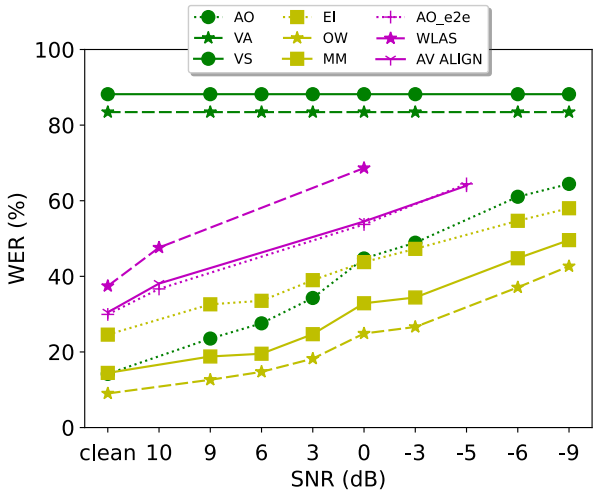

Fig. 2: Word error rate on LRS2 corpus

Fig. 2 also depicts the word error rate results of different end-to-end models, which can be found in the appendix of [11]. The Watch Listen Attend and Spell model [12] (**WLAS**) and the proposed fusion strategy in [11] (**AV Align**) can not improve the WER on the LRS2 dataset. We also find that the hybrid audio-only model (**AO**) offers a better performance than the end-to-end audio-only model (**AO_e2e**) from [11].

TABLE II: WER (%) on LRS2.

SNR	-9	-6	-3	0	3	6	9	clean	avg.
AO	64.45	61.05	48.90	44.73	34.28	27.57	23.54	14.20	39.84
VA	83.44	83.44	83.44	83.44	83.44	83.44	83.44	83.44	83.44
VS	88.18	88.18	88.18	88.18	88.18	88.18	88.18	88.18	88.19
EI	58.01	54.69	47.27	43.77	38.99	33.53	32.60	24.58	41.68
OW	42.67	37.08	26.61	24.88	18.22	14.74	12.64	9.02	23.23
MSE	50.66	47.10	35.81	33.56	25.20	19.93	**18.29**	13.60	30.52
CE	50.90	48.40	36.17	33.94	25.47	19.86	18.36	**13.45**	30.82
MMI	51.69	48.74	36.53	33.75	25.85	20.11	18.87	13.60	31.14
MM	**49.58**	**44.78**	**34.41**	**32.88**	**24.70**	**19.53**	18.80	14.46	29.89

Tab. II summarizes all results of the Kaldi experiments: the first 3 rows show the performance of all single-modality models. The performance metrics of the video appearance and shape models are far from satisfying. We have also employed the pre-trained spatio-temporal visual front-end from [26] to extract high-level visual features, without seeing improvements. We hypothesize that the unsatisfying video model performance is due to an insufficient amount of training data.

The 4th and 5th row show the results of the early-integration baseline, and of the oracle weighting that gives an upper bound of achievable performance for the considered hybrid architecture. The final 4 rows show the WERs for our proposed experiments, using all reliability indicators. Comparing the different loss functions for training the stream integration net, cf. Sec. III-A, the mean squared error (**MSE**) has the best performance at lower SNR conditions ($\leqslant 3$ dB), which can be carried over into the sequence-based optimization by adding an MSE-based fine-tuning after pre-training with the original **MMI** loss (**MM**). Comparing the best performance between our proposed hybrid audio-visual model (**MM**) and the end-to-end audio-visual model (**AV ALIGN**) in Fig. 2, we find that the hybrid model offers clear performance benefits, and that, in contrast to the end-to-end integration mechanism, it is indeed able to profit strongly from the inclusion of the visual modality under all noisy conditions.

To compare the value of the different types of reliability indicators for multi-modal integration, we have repeated this experiment with different reliability sets, using the best combination of loss functions, MMI training with MSE fine-tuning (**MM**). Tab. III gives the average WER over all SNR conditions. Here, we find that audio-based reliabilities have a slightly higher benefit, but using all reliabilities simultaneously achieves the best performance overall, corresponding to a relative word-error-rate reduction of 24.97% over the best audio-only model.

TABLE III: WER (%) of different reliability sets, abbreviations as introduced in Tab. I

	MB	SB	AB	VB	All
MM	31.40	31.33	31.03	31.76	**29.89**

VII. Conclusion

Improving the performance of large-vocabulary speech recognition through the inclusion of video data has remained challenging despite much progress in deep learning models for speech recognition and image processing. In this paper, we address this issue by learning an explicit stream integration network for audio-visual speech recognition. This network utilizes stream reliability indicators to optimize stream fusion time-frame by time-frame, ultimately providing a discriminatively optimized fusion of state-posteriors for hybrid speech recognition. We have compared the performance of this learned integration model to that of early integration as well as to a baseline end-to-end model. All experiments show that the proposed model dramatically outperforms both of these baseline systems and that it is capable of providing clear improvements in accuracy compared to audio-only recognition, as hoped.

However, experiments based on oracle knowledge for stream fusion also point at the possibility of significant further gains. Achieving similar improvements without oracle information will be the natural next goal of our work, where we will focus on both the topology and loss function of the fusion network as well as on the integration of deeper, pre-trained image recognition models.

References

[1] K. Thangthai and R. Harvey, "Building large-vocabulary speaker-independent lipreading systems," in *Interspeech*, 2018.

[2] M. J. Crosse, G. M. Di Liberto, and E. C. Lalor, "Eye can hear clearly now: inverse effectiveness in natural audiovisual speech processing relies on long-term crossmodal temporal integration," *Journal of Neuroscience*, vol. 36, no. 38, pp. 9888–9895, 2016.

[3] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," in *arXiv:1809.02108*, 2018.

[4] H. Meutzner, 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*, 2017, pp. 5320–5324.

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

[6] H. Bourlard and N. Morgan, *Connectionist speech recognition: a hybrid approach*, vol. 247, Springer Science & Business Media, 2012.

[7] C. Lüscher, E. Beck, K. Irie, M. Kitza, W. Michel, A. Zeyer, R. Schlüter, and H. Ney, "RWTH ASR systems for LibriSpeech: Hybrid vs Attention," *Interspeech, Graz, Austria*, 2019.

[8] V. Estellers, M. Gurban, and J. Thiran, "On dynamic stream weighting for audio-visual speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1145–1157, 2011.

[9] A.H. Abdelaziz, S. Zeiler, and D. Kolossa, "Learning dynamic stream weights for coupled-HMM-based audio-visual speech recognition," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 5, pp. 863–876, May 2015.

[10] S. Zeiler, R. Nickel, N. Ma, G. Brown, and D. Kolossa, "Robust audiovisual speech recognition using noise-adaptive linear discriminant analysis," in *Proc. ICASSP*, 2016.

[11] G. Sterpu, C. Saam, and N. Harte, "Attention-based audio-visual fusion for robust automatic speech recognition," in *Proc. ICMI*, 2018, pp. 111–115.

[12] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6447–6456.

[13] G Sterpu, 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*, vol. 28, pp. 1052–1064, 2020.

[14] K. Veselỳ, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Interspeech*, 2013, vol. 2013, pp. 2345–2349.

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

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

[17] R. Sproull, "Using program transformations to derive line-drawing algorithms," *ACM Transactions on Graphics (TOG)*, vol. 1, no. 4, pp. 259–273, 1982.

[18] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *Proc. ICLR*, 2014.

[19] H. Hermansky, "Multistream recognition of speech: Dealing with unknown unknowns," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1076–1088, 2013.

[20] H. Misra, H. Bourlard, and V. Tyagi, "New entropy based combination rules in HMM/ANN multi-stream ASR," in *Proc. ICASSP*. IEEE, 2003, vol. 2, pp. II–741.

[21] S. Rangachari and P. Loizou, "A noise-estimation algorithm for highly non-stationary environments," *Speech communication*, vol. 48, no. 2, pp. 220–231, 2006.

[22] L. Schönherr, D. Orth, M. Heckmann, and D. Kolossa, "Environmentally robust audio-visual speaker identification," in *Proc. SLT*. IEEE, 2016, pp. 312–318.

[23] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," 2015, arXiv:1510.08484v1.

[24] D. Povey, 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.

[25] B. Amos, 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.

[26] T Stafylakis and G Tzimiropoulos, "Combining residual networks with LSTMs for lipreading," *arXiv preprint arXiv:1703.04105*, 2017.