IMPROVING AUDIO-VISUAL SPEECH RECOGNITION USING DEEP NEURAL NETWORKS WITH DYNAMIC STREAM RELIABILITY ESTIMATES

Hendrik Meutzner[1], *Ning Ma*[2], *Robert Nickel*[3], *Christopher Schymura*[1], *Dorothea Kolossa*[1]

[1]Insititute of Communication Acoustics, Ruhr-Universität Bochum, Germany
[2]Department of Computer Science, University of Sheffield, Sheffield S1 4DP, UK
[3]Bucknell University, Lewisburg, PA, USA

ABSTRACT

Audio-visual speech recognition is a promising approach to tackling the problem of reduced recognition rates under adverse acoustic conditions. However, finding an optimal mechanism for combining multi-modal information remains a challenging task. Various methods are applicable for integrating acoustic and visual information in Gaussian-mixture-model-based speech recognition, e.g., via dynamic stream weighting. The recent advances of deep neural network (DNN)-based speech recognition promise improved performance when using audio-visual information. However, the question of how to optimally integrate acoustic and visual information remains.

In this paper, we propose a state-based integration scheme that uses dynamic stream weights in DNN-based audio-visual speech recognition. The dynamic weights are obtained from a time-variant reliability estimate that is derived from the audio signal. We show that this state-based integration is superior to early integration of multi-modal features, even if early integration also includes the proposed reliability estimate. Furthermore, the proposed adaptive mechanism is able to outperform a fixed weighting approach that exploits oracle knowledge of the true signal-to-noise ratio.

Index Terms— audio-visual speech recognition, deep neural networks, feature fusion, dynamic stream weighting

1. INTRODUCTION

Despite many recent advances in the field of automatic speech recognition, there is still room for improvement regarding the robustness against non-stationary environmental noise, which is often present in real world applications, e.g., in distant talking scenarios. Audio-visual speech recognition (AVSR) ameliorates this problem by including additional visual information to preserve or even improve the recognition performance under harsh environmental conditions.

A common approach to integrating acoustic and visual information is to combine the likelihood estimates of both modalities in a weighted fashion. However, finding an optimal weighting scheme that appropriately considers the reliability of each modality is a challenging task.

The approach of dynamic stream weight estimation has proven to be advantageous for combining the acoustic and visual information in purely statistical systems that are based on hidden Markov models (HMMs) modeling the state output densities using Gaussian mixture models (GMMs) [1, 2]. Recent studies have shown that deep neural networks (DNNs) can yield significant improvements for AVSR compared to generative statistical models. However, most studies only utilize the audio and video information at hand and do not consider additional stream reliability measures to inform the multi-modal integration process [3, 4, 5, 6].

In this paper, we analyze the efficacy of using an additional reliability measure, i.e., dynamically estimated signal-to-noise ratios (SNR) of the acoustic modality, to better integrate the acoustic and visual features in a DNN-based recognition system. We compare the performance of two different integration methods, namely fusing the audio, video and reliability features at the DNN input layer (early integration) versus a state-based integration scheme that uses dynamic stream weights based on the same reliability estimate.

2. RELATED WORK

Thangthai et al. [3] report that a HMM-DNN hybrid system is superior to an HMM-GMM-based recognizer when using a simple concatenation of audio and video features without additional weighting of the individual modalities. Noda et al. [4] and Ninomiya et al. [5] utilize multi-stream HMMs for integrating audio-visual features that have been derived from deep learning architectures, where the integration of streams is based on manually optimized weights. Huang and Kingsbury [6] investigate the fusion of mid-level features by concatenating the hidden representations of single-modality deep belief networks (DBNs) and using the result as the input to an audio-visual DBN. Heckmann et al. [7] compare different criteria, i.e., entropy, dispersion and voicing index, for combining the audio and video a-posteriori probabilities estimated by an artificial neural network on a number recognition task.

This project was supported by the German research foundation DFG (project KO3434/4-1), the EU FET grant TWO!EARS (ICT-618075), and by the DFG Research Training Group GRK 1817/1.

3. SYSTEM OVERVIEW

We utilize a hybrid system based on HMMs for modeling the temporal structure of the audio-visual speech signals. The HMM state observation probabilities are estimated by a feed-forward DNN using the observed feature sequence. The speech recognizer has been implemented using the Kaldi speech recognition toolkit [8], which was extended to support stream weighting for combining multiple modalities.

We start by training a conventional HMM-GMM-based triphone recognizer using feature space maximum likelihood linear regression (fMLLR) [9] and speaker adaptive training on top of LDA[1]-transformed features [10], where we follow the standard recipe provided by the Kaldi baseline scripts for the CHiME-2 [11] data. The input dimension of the LDA transform depends on the dimension of utilized features. When using a context window length of 7 frames (centered around the current frame) the output dimension of the LDA transform is 40 dimensions.

Next, we gradually build a feed-forward DNN to replace the generative model of the previous recognition system. The DNN layers are first initialized by means of restricted Boltzmann machines that are pre-trained using the contrastive divergence algorithm [12]. The pre-trained layers are then stacked and fine-tuned by minimizing the per-frame cross-entropy. The fine-tuning makes use of the state alignments derived from a forced alignment by using an HMM-GMM system that was trained on clean data (cf. Sec. 4).

As a final step, we improve the DNNs by conducting several iterations of sequence-discriminative training using the state-level minimum Bayes risk (sMBR) criterion [13].

The input features of all DNN systems are temporally spliced versions of the LDA-fMLLR-transformed features used for the initial HMM-GMM system with a context window length of 11 frames, corresponding to an input layer size of 440 dimensions. The networks consist of 6 hidden layers, each using a sigmoid activation function and the number of hidden units in each layer is 2048. The output layers use a soft-max function and consist of 1453 units that correspond to the individual triphone states in the HMM, which have been determined by decision tree clustering using the clean audio-only HMM-GMM system.

3.1. Feature types and noise estimation

We compare various feature types and feature combinations: 23-dimensional Mel filterbank features, 13-dimensional Mel frequency cepstral coefficients (MFCC), 32-dimensional ratemap features [14], and 13-dimensional Gammatone frequency cepstral coefficients (GFCC) [15]. The ratemap and GFCC features are motivated by the auditory system, as they encode a spectro-temporal representation of the auditory nerve firing rate that stems from the mechanotransduction process in the cochlea. This provides an interesting comparison with the more commonly used features such as MFCCs.

[1]Linear discriminant analysis.

The 63-dimensional video features are obtained from a discrete cosine transform of the gray-scale images that contain the mouth regions, determined by the Viola-Jones algorithm, as in [16].

Our goal is to combine the acoustic and visual information in such a way as to reliably obtain better recognition rates than the best of the two single-modality systems. For this purpose, here, we utilize a time-variant stream reliability measure, and we evaluate two different approaches for its use—either an early integration of different modalities, including reliabilities, at the input layer of the DNN, or a weighted combination of the DNN posterior outputs, with the weighting controlled by the reliability measure. For this purpose, we measure the degradation—and thus the reliability—of the acoustic modality using the Improved Minima Controlled Recursive Averaging (IMCRA) approach [17] that was shown to perform well under highly non-stationary noise conditions [18].

The IMCRA algorithm provides various time-frequency estimates, where we make use of the estimated a-priori SNR $\Xi_{t,f}$ as well as of the enhanced power spectrum

$$\tilde{X}_{t,f} = \frac{\Xi_{t,f}^2}{(\Xi_{t,f}+1)^2} X_{t,f} = G_{t,f}\, X_{t,f}, \qquad (1)$$

where $X_{t,f}$ represents the linear power spectrum of the noisy observation signal. We regard $\Xi_{t,f}$ and $\tilde{X}_{t,f}$ as the reliability feature, with the latter encoding the SNR indirectly in terms of the Wiener gain function $G_{t,f}$.

The control parameters of the IMCRA algorithm were chosen as in [17], with the difference that we set the window length to one as we found that small windows sizes result in higher recognition rates. The number of frequency components of the spectral quantities is given by $N_f = 257$.

When used as a single or auxiliary feature, the IMCRA estimates are warped to a 23-dimensional Mel scale to reduce the number of feature vector components and to give a fair comparison to the other lower-dimensional features.

3.2. Early integration

Let $\mathbf{x}^i = [x_0^i, \ldots, x_{D_i-1}^i]$ denote the feature vector of the i-th feature type that consists of D_i dimensions, where we have omitted the dependency on the frame time t for convenience. The extended feature vector for two different feature types is then given by their concatenated version

$$\tilde{\mathbf{x}} = [\mathbf{x}^0||\mathbf{x}^1] = [x_0^0, \ldots, x_{D_0-1}^0, x_0^1, \ldots, x_{D_1-1}^1]. \qquad (2)$$

This procedure can be extended to an arbitrary number of feature types, and we compare the performance for different feature combinations in Sec. 5.2.

3.3. State-based integration

The state-based audio-visual integration is achieved through a weighted combination of the DNN state posteriors of two different models

$$\log \tilde{p}(\mathbf{o}_t^{\mathrm{AV}}|s) = \lambda_t \log p'(\mathbf{o}_t^{\mathrm{A}}|s) + (1-\lambda_t) \log p''(\mathbf{o}_t^{\mathrm{V}}|s), \quad (3)$$

where $\lambda_t \in [0,1]$ represents the time-dependent stream weight and $\log p'(\mathbf{o}_t^{\mathrm{A}}|s)$ and $\log p''(\mathbf{o}_t^{\mathrm{V}}|s)$ denote the log-likelihood of the acoustic feature observations $\mathbf{o}_t^{\mathrm{A}}$ and the visual feature observation $\mathbf{o}_t^{\mathrm{V}}$, respectively, given the state index s. The combined log-likelihood $\log \tilde{p}(\mathbf{o}_t^{\mathrm{AV}}|s)$ of both modalities is then used for decoding the sequence of observation pairs $\mathbf{o}_t^{\mathrm{AV}} = (\mathbf{o}_t^{\mathrm{A}}, \mathbf{o}_t^{\mathrm{V}})$.

For computing the dynamic stream weights, we map the frequency-averaged a-priori SNR estimate

$$\bar{\Xi}_t = \frac{1}{N_f} \sum_{\forall f} \Xi_{t,f}, \tag{4}$$

to a suitable stream weight value by using a logistic function

$$\lambda_t = \alpha + \frac{\beta}{1 + e^{-\frac{\bar{\Xi}_t - \mu}{\sigma}}}, \tag{5}$$

where α represents an offset, β is a scaling factor, μ is the midpoint of the curve and σ defines the slope of the curve. The parameters of the mapping function have been found by fitting the logistic function to the cumulative probability density function of the IMCRA a-priori SNR estimates using all utterances and time frames of the training set, with the constraint that weight values are limited to the interval $[0.60, 0.74]$, which are based on the range of optimal weights found via stream-weight tuning (cf. Fig. 2).

In addition, we use a semi-dynamic scheme, where the weights are not adjusted frame by frame but kept constant per utterance using the temporal average of the a-priori SNR estimate. The parameters of the logistic function have been found in the same manner as for Eq. (5) by using the density functions of the temporally averaged SNR.

4. EXPERIMENTAL SETUP

We evaluate our approach using the audio data of Track 1 of the 2nd CHiME Speech Separation and Recognition Challenge [11], where the task is to recognize short command sentences that are of the form

```
<command:4><color:4><preposition:4>
        <letter:25><number:10><adverb:4>.
```

In the brackets, the number of alternatives for each word type is shown. Clean audio material from the GRID corpus [19] is used, which contains recordings from 34 speakers (18 male, 16 female). All audio signals have been processed to simulate room reverberation, small speaker movements and environmental noise. To achieve this, the signals were filtered with binaural room impulse responses followed by an additive mixing with a highly non-stationary background noise that was recorded in a family living room. The temporal placement of the clean speech within the background noise has been done in a controlled manner to yield six different SNR conditions between -6 dB and 9 dB without rescaling the speech or noise signals. The CHiME challenge data consists of three pre-segmented sets, i.e., a training set, a development set, and a test set. The original audio material is sampled at 16 kHz and contains binaural signals. In the following experiments, all signals were downmixed by taking the average of the left and the right channel before extracting audio features. The video data was also taken from the GRID corpus, which contains clean facial video recordings for each utterance.

4.1. Training and evaluation

We use the development set for determining the optimal set of features for the early integration scheme as well as for finding the fixed oracle stream weights for each SNR condition. The test set is then used for evaluating our proposed methods.

All models are trained under matched conditions, i.e., the training and evaluation process only considers the noisy mixture signals provided by the corpus.

We measure the speech recognition performance in terms of the keyword accuracy (the official evaluation metric of the CHiME challenge), where a keyword is defined by the letter-number pair that occures before the adverb of each utterance.

5. RESULTS

We first provide a performance comparison between different models and look for the optimal set of features. Based on these findings we then analyze the proposed state-based integration approach using dynamic reliability estimates.

5.1. Model performance

We compare the performance between different models, i.e., the GMM-based system and the DNN-based systems at different training stages when using MFCC features in Fig. 1. We can see that DNN-based systems are able to outperform the GMM-based system for each SNR, where the largest relative improvements are seen for very low SNR conditions. The recognition accuracy further increases when using the sMBR criterion for training, where a higher number of iterations (i.e., sMBR-5 vs. sMBR-1) yields slight performance improvements on average over all SNRs. For the sake of compactness, we thus limit the ensuing evaluation to using only the strongest DNN system (DNN sMBR-5).

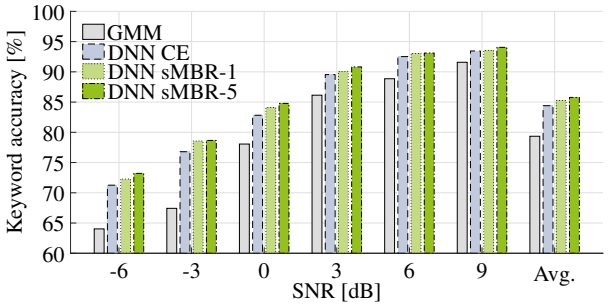

Fig. 1: Comparison of models for MFCC features (development set).

5.2. Early integration

Table 1 shows the keyword accuracies that are achieved for single feature types and their combinations using early inte-

gration. The results show that a fusion of audio and video features (i.e., FV, MV, RV, and GV) can generally improve the recognition performance at low SNR conditions below 0 dB, whereas audio-only features (i.e., F, M, R, and G) are generally superior at higher SNR conditions ($\geq$ 6 dB). A direct comparison of the isolated IMCRA-based reliability measures shows that the enhanced spectrum $\mathrm{I}_{\tilde{X}}$ outperforms the estimated a-priori SNR I_{Ξ} for each condition. When using $\mathrm{I}_{\tilde{X}}$ as an additional reliability feature for the early integration approach, the average recognition performance can be further improved as compared to the audio-visual feature combinations in most cases ($\mathrm{I}_{\tilde{X}}$FV, $\mathrm{I}_{\tilde{X}}$MV, $\mathrm{I}_{\tilde{X}}$GV). The ratemap features clearly outperform the other feature types when used in combination with the video features (i.e., RV and $\mathrm{I}_{\tilde{X}}$RV). Considering the lip-reading performance, the video features alone yield a keyword recognition accuracy of 71.34 %, which corresponds to a word accuracy of 84.81 %.

Table 1: Keyword accuracies (%) obtained on the development set for various features: Mel filterbank (F), MFCC (M), Ratemap (R), GFCC (G), Video (V), IMCRA-enhanced spectrum ($\mathrm{I}_{\tilde{X}}$), IMCRA-estimated a-priori SNR (I_{Ξ}), and selected fused combinations.

Feat.	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	Avg.
F	72.11	78.06	86.14	90.65	92.86	93.11	85.49
M	73.21	78.66	84.78	90.82	93.11	94.05	85.77
R	73.13	**80.36**	**86.48**	89.54	92.77	93.37	85.94
G	**73.98**	78.57	**86.48**	**89.71**	**92.86**	**94.47**	**86.01**
V	71.34	71.34	71.34	71.34	71.34	71.34	71.34
I_{Ξ}	62.07	66.84	75.68	80.44	85.29	87.93	76.38
$\mathrm{I}_{\tilde{X}}$	71.60	78.49	83.50	89.12	91.75	93.11	84.59
FV	**87.76**	**88.52**	88.69	89.88	90.90	90.82	89.43
MV	82.14	83.42	84.01	86.31	87.41	87.33	85.10
RV	**87.76**	88.18	**89.88**	**91.50**	**92.26**	**92.26**	**90.31**
GV	83.84	84.69	85.29	86.90	88.27	88.52	86.25
$\mathrm{I}_{\tilde{X}}$FV	**86.56**	**89.03**	**89.37**	91.33	91.84	92.43	90.09
$\mathrm{I}_{\tilde{X}}$MV	84.44	87.24	88.78	89.80	90.65	91.07	88.66
$\mathrm{I}_{\tilde{X}}$RV	85.97	88.10	**89.37**	**91.75**	**92.94**	**93.11**	**90.21**
$\mathrm{I}_{\tilde{X}}$GV	84.69	87.16	88.86	90.90	91.84	91.58	89.17

5.3. State-based integration

The analysis of the state-based integration approach is done using the ratemap acoustic features as they have achieved the best performance for the early integration approach. Figure 2 shows the optimal fixed stream weights λ_{opt} that were found via parameter search on the development set. We can see that stream weights are roughly increasing with increasing SNR and their values are within the interval [0.60, 0.74].

The final results obtained on the test set using the ratemap acoustic features are given by Tab. 2. We can see that the early integration (RV, $\mathrm{I}_{\tilde{X}}$RV) outperforms the fixed stream weights (Fixed) on average. However, the reliability features do not improve the performance for the early integration approach ($\mathrm{I}_{\tilde{X}}$RV vs. RV). The best performance is achieved by

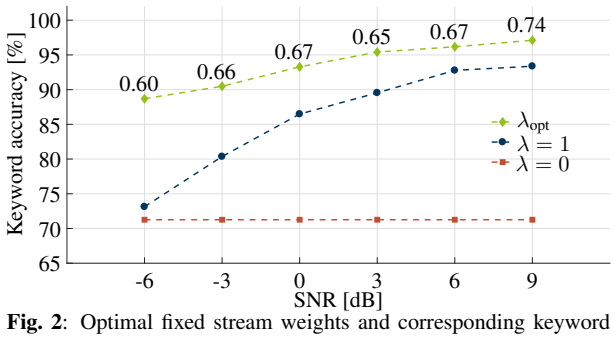

Fig. 2: Optimal fixed stream weights and corresponding keyword accuracies for ratemap acoustic features (development set).

the dynamic stream weights, where the frame-wise weighting (SNR-F) outperforms the utterance-wise weighting (SNR-U) for most SNRs above -3 dB and on average.

Table 2: Keyword accuracies (%) for early versus state-based integration using the ratemap acoustic features. All scores are based on the test set. The methods R, V, RV, and $\mathrm{I}_{\tilde{X}}$RV correspond to the features of Tab. 1. *Fixed* corresponds to the optimal fixed stream weights tuned on the development set. *SNR-U* and *SNR-F* denote the utterance and frame-wise dynamic stream weights estimated without oracle SNR knowledge.

Method	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	Avg.
R	73.88	78.44	85.48	88.49	92.01	93.56	85.31
V	70.96	70.96	70.96	70.96	70.96	70.96	70.96
RV	**87.20**	**87.63**	**89.60**	**90.64**	**91.58**	91.58	**89.70**
$\mathrm{I}_{\tilde{X}}$RV	84.36	86.00	88.06	90.38	90.89	**92.53**	88.70
Fixed	71.48	79.21	89.35	94.76	94.85	93.56	87.20
SNR-U	**88.32**	**90.21**	93.04	94.67	**95.02**	96.74	93.00
SNR-F	87.97	90.12	**93.47**	**95.02**	94.76	**96.82**	**93.03**

6. CONCLUSIONS

We have analyzed the effect of an additional stream reliability estimate for improving the integration of acoustic and visual data in DNN-based AVSR. Specifically, we have compared its value in early integration and in state-based integration. All experiments have clearly shown that state-based integration, with stream weighting based on the reliabilities, is superior to using reliabilities as additional features in early integration, calling into question our original idea of letting the DNN learn the optimal integration from data alone. Under all conditions, the introduced dynamic weighting mechanism has also outperformed a strong baseline setting using fixed stream weights that exploit oracle knowledge of the current SNR.

Furthermore, we have compared a range of acoustic feature types for use in DNN-based AVSR, finding that the auditory-inspired ratemap features show superior performance compared to other feature types that are more commonly used for speech recognition applications (such as Mel filterbank features).

7. REFERENCES

[1] Ahmed Hussen Abdelaziz and Dorothea Kolossa, "Dynamic Stream Weight Estimation in Coupled-HMM-based Audio-visual Speech Recognition using Multilayer Perceptrons," in *Proc. INTERSPEECH*, 2014, pp. 1144–1148.

[2] Ahmed Hussen Abdelaziz, Steffen Zeiler, and Dorothea 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.

[3] Kwanchiva Thangthai, Richard Harvey, Stephen Cox, and Barry-John Theobald, "Improving Lip-reading Performance for Robust Audiovisual Speech Recognition using DNNs," in *Proc. Joint Conference on Facial Analysis, Animation and Audio–Visual Speech Processing*, 2015.

[4] Kuniaki Noda, Yuki Yamaguchi, Kazuhiro Nakadai, Hiroshi G. Okuno, and Tetsuya Ogata, "Audio-visual Speech Recognition Using Deep Learning," *Applied Intelligence*, vol. 42, no. 4, pp. 722–737, June 2015.

[5] Hiroshi Ninomiya, Norihide Kitaoka, Satoshi Tamura, Yurie Iribe, and Kazuya Takeda, "Integration of Deep Bottleneck Features for Audio-visual Speech Recognition," in *Proc. INTERSPEECH*, 2015.

[6] Jing Huang and Brian Kingsbury, "Audio-visual Deep Learning for Noise Robust Speech Recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 7596–7599.

[7] Martin Heckmann, Frédéric Berthommier, and Kristian Kroschel, "Noise Adaptive Stream Weighting in Audio-visual Speech Recognition," *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 1, pp. 1260–1273, Jan. 2002.

[8] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The Kaldi Speech Recognition Toolkit," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, Dec. 2011.

[9] Daniel Povey and George Saon, "Feature and Model Space Speaker Adaptation with Full Covariance Gaussians," in *Proc. INTERSPEECH*, 2006.

[10] Shakti P. Rath, Daniel Povey, Karel Veselý, and Jan Cernocký, "Improved Feature Processing for Deep Neural Networks," in *Proc. INTERSPEECH*, 2013, pp. 109–113.

[11] Emmanuel Vincent, Jon Barker, Shinji Watanabe, Jonathan Le Roux, Francesco Nesta, and Marco Matassoni, "The second 'CHiME' Speech Separation and Recognition Challenge: Datasets, tasks and baselines," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2013, pp. 126–130.

[12] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh, "A Fast Learning Algorithm for Deep Belief Nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, July 2006.

[13] Karel Veselý, Arnab Ghoshal, Lukás Burget, and Daniel Povey, "Sequence-discriminative Training of Deep Neural Networks," in *Proc. INTERSPEECH*, 2013, pp. 2345–2349.

[14] Guy J. Brown and Martin Cooke, "Computational auditory scene analysis," *Computer Speech & Language*, vol. 8, no. 4, pp. 297 – 336, 1994.

[15] Y. Shao, Z. Jin, D. Wang, and S. Srinivasan, "An auditory-based feature for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2009, pp. 4625–4628.

[16] Ahmed Hussen Abdelaziz, Steffen Zeiler, and Dorothea Kolossa, "Twin-HMM-based Audio-visual Speech Enhancement," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 3726–3730.

[17] Israel Cohen, "Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, Sept 2003.

[18] Sundarrajan Rangachari and Philipos C Loizou, "A Noise-estimation Algorithm for Highly Non-stationary Environments," *Speech communication*, vol. 48, no. 2, pp. 220–231, 2006.

[19] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao, "An Audio-visual Corpus for Speech Perception and Automatic Speech Recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.