Using Twin-HMM-Based Audio-Visual Speech Enhancement as a Front-End for Robust Audio-Visual Speech Recognition

Ahmed Hussen Abdelaziz, Steffen Zeiler, Dorothea Kolossa

Institute of Communication Acoustics, Digital Signal Processing Group,
Ruhr-Universität Bochum, Germany
{Ahmed.HussenAbdelAziz, Steffen.Zeiler, Dorothea.Kolossa}@rub.de

10.21437/Interspeech.2013-257

Abstract

In this paper we propose the use of the recently introduced twin-HMM-based audio-visual speech enhancement algorithm as a front-end for audio-visual speech recognition systems. This algorithm determines the clean speech statistics in the recognition domain based on the audio-visual observations and transforms these statistics to the synthesis domain through the so-called twin HMMs. The adopted front-end is used together with back-end methods like the conventional maximum likelihood decoding or the newly introduced significance decoding. The proposed combination of the front- and back-end is applied to acoustically corrupted signals of the Grid audio-visual corpus and results in statistically significant improvements of the audio-visual recognition accuracy compared to using the ETSI advanced front-end.

Index Terms: Multimodal speech processing, audiovisual speech recognition, model-based speech enhancement

1. Introduction

Enhancing noisy speech signals is a crucial preprocessing stage in automatic speech recognition (ASR) systems. Most speech enhancement algorithms depend on estimating the hidden clean spectrum of speech from only the observable corrupted acoustical input. This makes it difficult to extract the clean spectrum from the corrupted one in low signal-to-noise ratios (SNRs). If, however, an additional source of information, independent of any acoustical noise, is involved in the clean spectrum estimation process, the estimation accuracy can be greatly improved. Visual observations have been introduced, e.g. in [1–4], as an attractive candidate of such a noise-independent information source. Recently, we have proposed the so-called twin-HMM-based audio-visual speech enhancement (THMMB-AV-SE) algorithm [5] that successfully exploits the visual information to enhance speech signals for human listening purposes. In this paper, we employ the THMMB-AV-SE algorithm as a preprocessing stage for audio-visual ASR systems.

The THMMB-AV-SE algorithm can compute not only the required clean spectrum estimate but also its estimation uncertainty, which was not exploitable for applications targeting human listeners. However, this uncertainty represents a further information source that can be employed in the recognition process using the so-called uncertainty-of-observation decoding approaches [6–13]. In this work, we dynamically compensate the acoustical observations and the statistical model parameters using the estimated uncertainties to increase the targeted match between them. The dynamic compensation is achieved here using the newest member of the uncertainty-of-observation decoding rules, namely the significance decoding (SD) approach [14].

As will be shown in the following, combining the THMMB-AV-SE algorithm in the front end and the SD approach in the back-end improves the audio-visual ASR performance in various noisy environments.

The remaining paper is organized as follows. First, the THMMB-AV-SE framework is briefly described in Section 2. Then, the new significance decoding rule is discussed in Section 3. In Section 4, the combination of the proposed front- and back-end methods is evaluated using the GRID audio-visual corpus [15]. Finally, the proposed approach are evaluated and the experimental results are discussed in Section 5.

2. Twin-HMM-based audio-visual speech enhancement

The main idea of the twin HMM (THMM) is to have *one* underlying state sequence that describes the temporal evolution of speech with *two* associated output density functions. One of the two density functions is trained with audio recognition (REC) features and is used for recognition purposes while the other one is trained with so-called synthesis (SYN) features and is dedicated to speech synthesis. This idea of the THMM is illustrated in Figure 2.

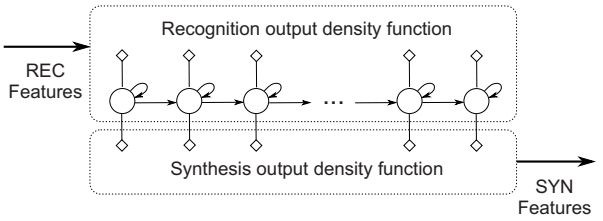

Figure 2: Concept of twin-HMM for model-based speech enhancement

Independence, saliency and robustness against noise and reverberation are the characteristics that have to be taken into account when choosing the audio REC features. On the other hand, the SYN features should be chosen so that they contain all needed information required for resynthesizing enhanced speech signals.

The statistical parameters of the REC output density function are estimated using the conventional EM-Algorithm [16]. The state occupation probabilities γ extracted at the final expectation step of the EM-algorithm are then used in estimating the parameters of the SYN output density function as shown in Figure 1.

Enhancing speech signals using THMMs takes place in two stages, namely the recognition and the synthesis stage. In

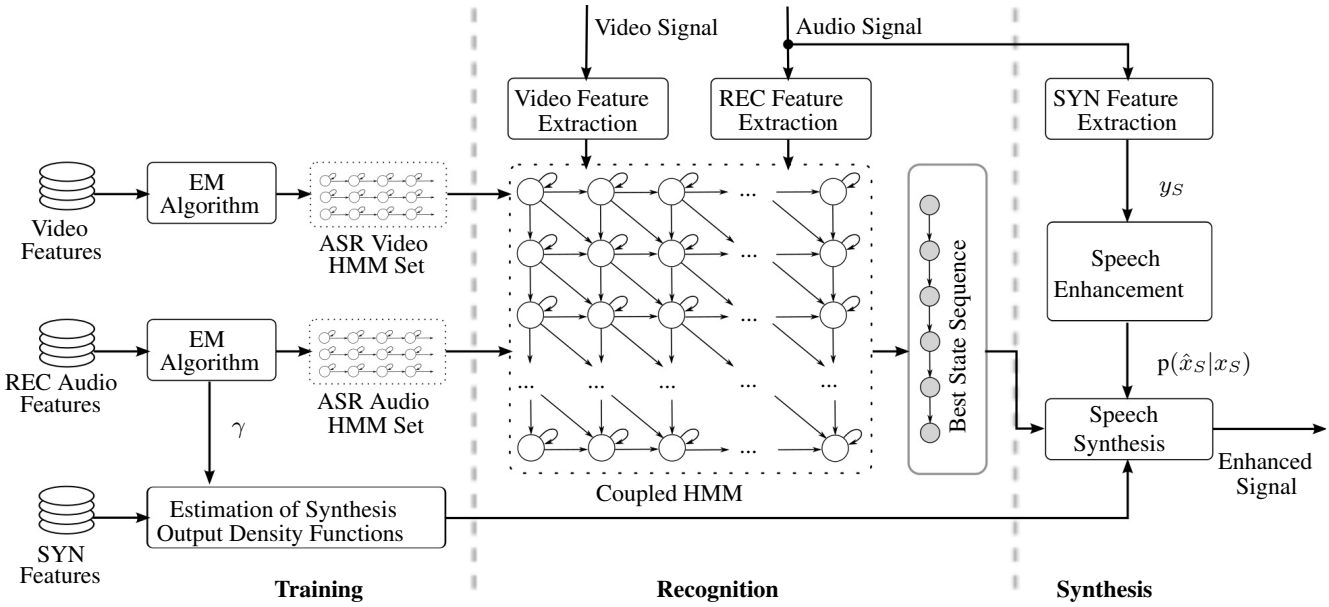

Figure 1: THMMB-AV-SE Framework.

the recognition stage, the state sequence of the THMMs that best describes the noisy speech is found. Since estimating frame/state alignment in noisy environments is a challenging task, we use the video features, which are completely independent of the acoustical environment, in conjunction with the audio features.

The fusion between the two modalities, i.e. the audio and the video modality, is done by using the so-called coupled HMM (CHMM) [17], which describes the co-evolution of the two modalities over time and allows asynchrony between them. The CHMMs are composed from the REC output distributions of the THMMs for the audio part and pre-trained video HMMs for the video part as shown in Figure 1.

After finding the best frame/state alignment of the noisy audio signal as discussed above, the enhanced signal is synthesized using: (1) the best state sequence, (2) the SYN output distributions $\mathrm{p}\left(\mathbf{x}_{S_t}|q(t)=i\right)$ of theses states and (3) unbiased estimates of the clean SYN feature vectors $\hat{\mathbf{x}}_{S_t}$.

The estimates $\hat{\mathbf{x}}_{S_t}$ of the clean SYN feature vectors $\mathbf{x}_{S_t}$ can be obtained from the noisy SYN feature vectors $\mathbf{y}_{S_t}$ using a speech enhancement algorithm like the Wiener filter. Such enhancement algorithms can typically compute the estimation error variances $\mathbf{\Sigma}_{\mathbf{e}_t}$, which represent the uncertainty of the observation estimate. The estimate $\hat{\mathbf{x}}_{S_t}$ and its uncertainty $\mathbf{\Sigma}_{\mathbf{e}_t}$ can be considered as the mean vector and the covariance matrix of a Gaussian[1] probability density function (PDF) $\mathrm{p}\left(\hat{\mathbf{x}}_{S_t}|\mathbf{x}_{S_t}\right)$ that describes the stochastic process of disturbing $\mathbf{x}_{S_t}$ with the estimation error and the residual noise.

The synthesizer exploits the statistics $\mathrm{p}\left(\hat{\mathbf{x}}_{S_t}|\mathbf{x}_{S_t}\right)$ and $\mathrm{p}\left(\mathbf{x}_{S_t}|q(t)=i\right)$ to compute the conditional PDF [5], [14]

$$\mathrm{p}\left(\mathbf{x}_{S_t}|\hat{\mathbf{x}}_{S_t}, q(t)=i\right) = \mathcal{N}\left(\mathbf{x}_{S_t}; \tilde{\boldsymbol{\mu}}_{S_{i,t}}, \tilde{\mathbf{\Sigma}}_{S_{i,t}}\right) \quad (1)$$

with mean vector

$$\tilde{\boldsymbol{\mu}}_{S_{i,t}} = \mathbf{\Sigma}_{e_t}\left(\mathbf{\Sigma}_{S_i}+\mathbf{\Sigma}_{e_t}\right)^{-1}\boldsymbol{\mu}_{S_i} + \mathbf{\Sigma}_{S_i}\left(\mathbf{\Sigma}_{S_i}+\mathbf{\Sigma}_{e_t}\right)^{-1}\hat{\mathbf{x}}_{S_t} \quad (2)$$

[1]The Gaussian assumption here is conventional [7].

and covariance matrix

$$\tilde{\mathbf{\Sigma}}_{S_{i,t}} = \mathbf{\Sigma}_{e_t}\left(\mathbf{\Sigma}_{e_t}+\mathbf{\Sigma}_{S_i}\right)^{-1}\mathbf{\Sigma}_{S_i}. \quad (3)$$

In (2) and (3), $\boldsymbol{\mu}_{S_i}$ and $\mathbf{\Sigma}_{S_i}$ are the mean vector and the covariance matrix of the SYN output distribution of the i^{th} recognized state q of the THMM. The mean vectors $\tilde{\boldsymbol{\mu}}_{S_{i,t}}$ can now be considered as the enhanced SYN features $\tilde{\mathbf{x}}_{S_t}$, from which the speech signal could be reconstructed using the overlap-add method.

Since this THMM framework is used here as a preprocessing stage for an ASR system, the REC features can be directly extracted from the enhanced SYN features $\tilde{\mathbf{x}}_{S_t}$ and there is no need for synthesizing the time-domain signal.

3. Decoding with uncertain data

The estimated vectors $\tilde{\mathbf{x}}_{S_t}$ computed using the THMM-AV-SE framework can be directly used in ASR systems as if they were the hidden clean SYN feature vectors $\mathbf{x}_{S_t}$, but by doing so, we would neglect an available source of information, which is the uncertainty of the estimated vectors. This uncertainty can be quantitatively represented by the computed covariance matrix $\mathbf{\Sigma}_{\tilde{\mathbf{x}}_S} = \tilde{\mathbf{\Sigma}}_{S_{i,t}}$. Good improvements of the recognition accuracy can be gained by involving this uncertainty in the recognition procedure using the so-called *uncertainty-of-observation* decoding approaches instead of the conventional maximum likelihood decoding rule [16].

Decoding of uncertain data using the uncertainty-of-observation techniques demands knowledge of the statistics $\mathrm{p}\left(\tilde{\mathbf{x}}_t|\mathbf{x}_t\right)$ in the recognition domain. Since we have assumed a Gaussian density function for $\mathrm{p}\left(\tilde{\mathbf{x}}_{S_t}|\mathbf{x}_{S_t}\right)$ in the synthesis domain with mean vector $\tilde{\mathbf{x}}_{S_t}$ and covariance matrix $\mathbf{\Sigma}_{\tilde{\mathbf{x}}_S}$, we can propagate this PDF through different feature extraction stages using the techniques introduced in [18] so that we get the required statistics $\mathrm{p}\left(\tilde{\mathbf{x}}_{R_t}|\mathbf{x}_{R_t}\right)$ in the recognition domain. In [18], it is shown that for the mel-cepstral features that are conventionally used in ASR, $\mathrm{p}\left(\tilde{\mathbf{x}}_{R_t}|\mathbf{x}_{R_t}\right)$ is well-modeled as Gaussian with mean vector $\tilde{\mathbf{x}}_{R_t}$ and covariance matrix $\mathbf{\Sigma}_{\tilde{\mathbf{x}}_R}$.

In this paper, we utilize the propagated PDF $\mathrm{p}\left(\tilde{\mathbf{x}}_{R_t}|\mathbf{x}_{R_t}\right)$ and the REC distribution of the THMM $\mathrm{p}\left(\mathbf{x}_{R_t}|q(t)\right)$ in the recognition process using the so-called *significance decoding* (SD) approach. The main idea of SD is based on a similar approach [19] taken in the expectation maximization (EM) algorithm, where the complete data log-likelihood is replaced by the expected value of the complete data log-likelihood, given the observed data. In the SD framework, the hidden observation likelihood[2] $\mathcal{L} = \mathrm{p}(\mathbf{x}_{R_t}|q(t))$ is replaced by the conditional expectation of the observation likelihood given all observable variables that are statistically related to $\mathbf{x}_{R_t}$. Therefore, we replace the hidden likelihood $\mathcal{L}$ by

$$\begin{aligned}\mathcal{L}^{\mathrm{SD}} &:= E\left[\mathrm{p}(\mathbf{x}_{R_t}|q(t))|\tilde{\mathbf{x}}_{R_t}, q(t)\right] \\ &= \int \mathrm{p}(\mathbf{x}_{R_t}|q(t))\,\mathrm{p}\left(\mathbf{x}_{R_t}|\tilde{\mathbf{x}}_{R_t}, q(t)\right) d\mathbf{x}_{R_t}. \end{aligned} \quad (4)$$

Conventionally, the REC output distributions are assumed to be Gaussian mixture models (GMMs) via:

$$\mathrm{p}(\mathbf{x}_{R_t}|q(t)) = \sum_{\kappa=1}^{K} \omega_{R_{i_\kappa}} \mathcal{N}\left(\mathbf{x}_{R_t}; \boldsymbol{\mu}_{R_{i_\kappa}}, \boldsymbol{\Sigma}_{R_{i_\kappa}}\right) \quad (5)$$

with $\omega_{R_{i_\kappa}}$, $\boldsymbol{\mu}_{R_{i_\kappa}}$ and $\boldsymbol{\Sigma}_{R_{i_\kappa}}$ as the weight, the mean vector and the covariance matrix of the REC output distribution of the κ^{th} mixture of the i^{th} state of the THMM. By applying the propagated PDF $\mathrm{p}\left(\tilde{\mathbf{x}}_{R_t}|\mathbf{x}_{R_t}\right)$ and (5) to (4), we get the general form of significance decoding as follows [14]:

$$\mathcal{L}^{\mathrm{SD}} = \sum_{\kappa=1}^{K}\sum_{\nu=1}^{K} \omega_{R_{i_\kappa}} \tilde{\omega}_{i_\nu} \mathcal{N}\left(\tilde{\boldsymbol{\mu}}_{i_\nu}; \boldsymbol{\mu}_{R_{i_\kappa}}, \boldsymbol{\Sigma}_{R_{i_\kappa}} + \tilde{\boldsymbol{\Sigma}}_{i_\nu}\right) \quad (6)$$

with

$$\tilde{\omega}_{i_\nu} = \frac{\omega_{R_{i_\nu}} \mathcal{N}\left(\tilde{\mathbf{x}}_{R_t}; \boldsymbol{\mu}_{R_{i_\nu}}, \boldsymbol{\Sigma}_{R_{i_\nu}} + \boldsymbol{\Sigma}_{\tilde{\mathbf{x}}_R}\right)}{\sum_{m=1}^{K} \omega_{R_{i_m}} \mathcal{N}\left(\tilde{\mathbf{x}}_{R_t}; \boldsymbol{\mu}_{R_{i_m}}, \boldsymbol{\Sigma}_{R_{i_m}} + \boldsymbol{\Sigma}_{\tilde{\mathbf{x}}_R}\right)}, \quad (7)$$

$$\begin{aligned}\tilde{\boldsymbol{\mu}}_{i_\nu} = \boldsymbol{\Sigma}_{R_{i_\nu}} \left(\boldsymbol{\Sigma}_{\tilde{\mathbf{x}}_R} + \boldsymbol{\Sigma}_{R_{i_\nu}}\right)^{-1} \tilde{\mathbf{x}}_{R_t} + \\ \boldsymbol{\Sigma}_{\tilde{\mathbf{x}}_R} \left(\boldsymbol{\Sigma}_{\tilde{\mathbf{x}}_R} + \boldsymbol{\Sigma}_{R_{i_\nu}}\right)^{-1} \boldsymbol{\mu}_{R_{i_\nu}} \end{aligned} \quad (8)$$

and

$$\tilde{\boldsymbol{\Sigma}}_{i_\nu} = \boldsymbol{\Sigma}_{\tilde{\mathbf{x}}_R} \left(\boldsymbol{\Sigma}_{\tilde{\mathbf{x}}_R} + \boldsymbol{\Sigma}_{R_{i_\nu}}\right)^{-1} \boldsymbol{\Sigma}_{R_{i_\nu}}. \quad (9)$$

A brief summary of the recognition procedure is shown in Fig. 3. First, the THMMB-AV-SE framework uses the video features to enhance the noisy audio features and generates the synthesis domain PDF $\mathrm{p}\left(\tilde{\mathbf{x}}_{S_t}|\mathbf{x}_{S_t}\right)$. The synthesis domain PDF is then propagated through the feature extraction phases so that the corresponding recognition domain PDF $\mathrm{p}\left(\tilde{\mathbf{x}}_{R_t}|\mathbf{x}_{R_t}\right)$ is estimated. Finally, the video features in conjunction with the estimated PDF $\mathrm{p}\left(\tilde{\mathbf{x}}_{R_t}|\mathbf{x}_{R_t}\right)$ are used in the audio-visual speech recognizer to transcribe the utterance, where the audio likelihoods are computed according to significance decoding rule given in (6).

[2] $\mathcal{L} = \mathrm{p}(\mathbf{x}_{R_t}|q(t))$ is hidden because of the absence of the clean REC observations $\mathbf{x}_{R_t}$ in noisy environments.

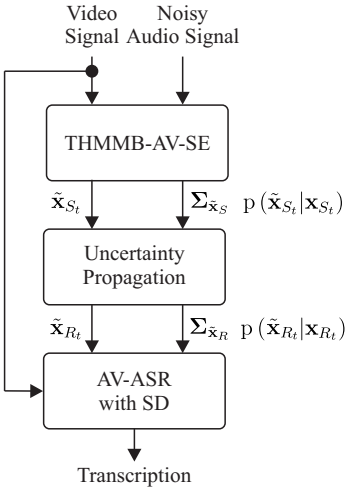

Figure 3: Using THMM-based speech enhancement as a front end for significance-decoding-based audio-visual speech recognition.

4. Experiments and results

4.1. Dataset

We have used the Grid audio-visual corpus [15] to evaluate the proposed framework shown in Figure 3. Only 33 out of 34 speakers of the Grid corpus are used because the visual data of Speaker 21 is not available. The task of the Grid corpus is to recognize sentences from a small vocabulary (51 words) with a fixed grammar of the form "command-color-preposition-letter-digit-adverb".

The speech data of each speaker are divided into a training and a test set. In order to create corrupted versions of the Grid corpus, two acoustical noise types have been artificially added to the the audio test sets at different signal-to-noise ratios (SNRs) in the range from 15 dB down to 0 dB. We have chosen speech babble and white noise as examples of non-stationary and stationary noise, respectively. The noise signals have been taken from the NOISEX database [20].

4.2. Experimental setup

The REC features used in the enhancement stage as well as in the recognition stage are the 13 static mel-frequency cepstral coefficients (MFCCs) [21] extracted as described in the ETSI advanced front-end (AFE) [22] together with the 26 corresponding first and second derivatives.

For the SYN features, we have used the short-time spectral amplitude with 129 dimensions. Since the sampling frequency of the ETSI-AFE is $f_s = 8$ kHz, we have downsampled the speech data of the Grid corpus from its original sampling frequency $f_s = 25$ kHz to the recommended one. However, we have used 150 overlapped samples instead of the 120 samples recommended in the ETSI-AFE to conform with the overlap-add constraints for synthesis.

In order to compute reliable video features, the Viola-Jones [23] algorithm has been used to detect the speaker's mouth region in the image. From this region, 64-dimensional DCT coefficient vectors encoding the appearance and shape of the speaker's mouth have then been extracted.

Both the audio and video features have been post processed by linear discriminant analysis (LDA) [24] in order to reduce

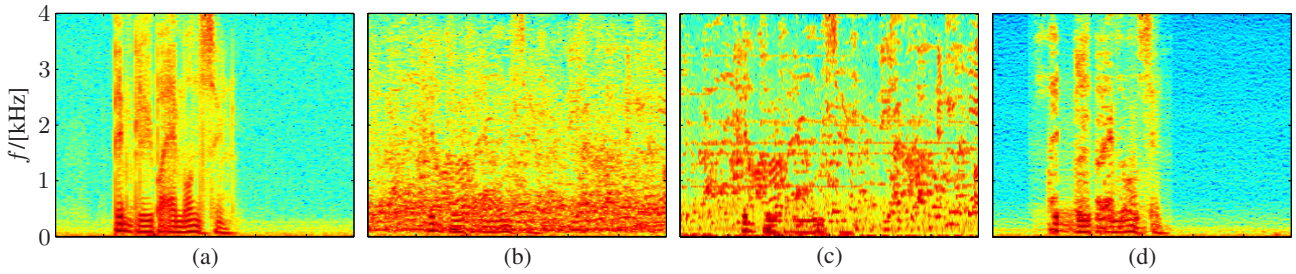

Figure 4: (a) Spectrum of the three seconds long Grid sentence "BIN BLUE BY M ONE SOON" uttered in clean conditions. (b) The same sentence with added babble noise at 0 dB SNR. (c) The enhanced spectrum of the Wiener filter used in the ETSI-AFE. (d) The filtered noisy spectrum after THMMB-AV-SE.

the dimensions of the audio as well as the video features to 31 and improve the recognition performance.

All models have been generatively trained using the REC and SYN features extracted from clean signals. The trained models are linear, whole-word and speaker-dependent, with the number of states depending on the number of phones per word. The REC and the SYN output distributions of each state of the twin HMMs are modeled, respectively, by 3- and 1-component diagonal covariance GMMs. For the video model, 4-component diagonal covariance GMMs have been used.

The mean and variance of a Wiener filter have been used as the parameters of the Gaussian PDF $\mathrm{p}\left(\hat{\mathbf{x}}_{S_t}|\mathbf{x}_{S_t}\right)$ [25] and the noise power has been estimated using the improved minima controlled recursive averaging (IMCRA) [26].

4.3. Results

Figure 4.a and 4.b depicts the spectrograms of a speech signal in clean and noisy conditions (babble noise at 0 dB), respectively. In Figure 4.c and 4.d, we compare the enhanced spectrograms generated when the ETSI-AFE and the THMMB-AV-SE are applied to the corrupted signal. It is clear that the improvements of the speech quality gained by applying the THMMB-AV-SE are much more than those gained by applying the ETSI-AFE.

As an objective quality measure of the processed signals, we have used the performance of the audio-only ASR in terms of word accuracy. The results exhibited in Table 1 shows that utilizing the video information in enhancing corrupted speech signals using THMMs is very successful. The signals processed by THMM-AV-SE provide great improvements in audio-only ASR accuracy compared to those processed by the conventional methods like the ETSI-AFE in different noise types and levels.

Table 1: Word accuracy of audio-only ASR when speech signals are preprocessed by the ETSI-AFE and the THMMB-AV-SE

Noise Type	SNR/ dB	ETSI-AFE	THMMB-AV-SE
Clean	-	99.11	**99.13**
Babble	15	92.42	**97.13**
	10	78.73	**93.31**
	5	52.70	**84.38**
	0	30.49	**67.45**
White	15	87.56	**96.17**
	10	72.75	**92.83**
	5	53.48	**86.70**
	0	34.22	**75.87**

However, the main goal we are after is not to improve the audio-only ASR performance but rather the audio-visual ASR performance. Thus, in Table 2 we compare the audio-visual word accuracies that result when using the THMMB-AV-SE with those achieved using the ETSI-AFE. Moreover, Table 2 compares the results of THMMB-AV-SE for both the maximum likelihood (ML) and the significance decoding rule.

The THMMB-AV-SE front-end in conjunction with significance decoding achieves the best performance in all noisy cases. The results with asterisks in Table 2 are significantly better compared to the ETSI-AFE results, according to Fisher's exact test applied at $\mathrm{p} = 0.05$.

Table 2: Word accuracy of audio-visual ASR for different preprocessing algorithms (ETSI-AFE and THMMB-AV-SE) and different decoding rules (ML and SD).

Noise Type	SNR/ dB	ETSI-AFE ML	THMMB-AV-SE ML	THMMB-AV-SE SD
Clean	-	**98.49**	98.48	98.42
Babble	15	96.40	96.58	**96.64**
	10	93.56	94.17*	**94.50***
	5	87.12	88.32*	**89.32***
	0	75.72	77.77*	**80.88***
White	15	96.11	96.24	**96.29**
	10	93.40	93.69	**93.89***
	5	88.28	88.72	**89.45***
	0	80.12	81.06*	**82.32***
Average	-	89.91	90.56*	**91.30***

5. Conclusions

Twin-HMM-based audio-visual speech enhancement has recently shown great success in de-noising speech signals even at very low SNRs. In this paper, we show that using this algorithm as a front-end for audio-visual ASR systems improves their performance compared to the conventional ETSI-AFE front-end. It can also be used as a front-end for standard audio-only ASR, in which case the recognition rate is improved drastically.

The recognizer performance can be enhanced further by combining this multi-modal speech enhancement with a back-end method that takes into account the observation uncertainty, like the recently introduced significance decoding rule.

6. Acknowledgements

This work has been supported by the Ministry of Economic Affairs and Energy of the State of North Rhine- Westphalia, Grant IV.5-43-02/2-005-WFBO-009.

7. References

[1] I. Almajai and B. Milner, "Visually derived Wiener filters for speech enhancement," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 19, no. 6, pp. 1642–1651, 2011.

[2] G. Potamianos, C. Neti, and S. Deligne, "Joint audio-visual speech processing for recognition and enhancement," in *Proc. AVSP*, 2003.

[3] F. Berthommier, "Characterization and extraction of mouth opening parameters available for audiovisual speech enhancement," in *Proc. ICASSP*, Montreal, Quebec, Canada, May 2004.

[4] S. Deligne, G. Potamianos, and C. Neti, "Audio-visual speech enhancement with AVCDCN (audio-visual codebook dependent cepstral normalization)," in *Proc. ICSLP*, 2002.

[5] A. H. Abdelaziz, S. Zeiler, and D. Kolossa, "Twin-HMM-based audio-visual speech enhancement," *Accepted for ICASSP*, 2013.

[6] J. A. Arrowood and M. A. Clements, "Using observation uncertainty in HMM decoding," in *Proc. International Conference on Spoken Language Processing*, Denver, Colorado, Sepember 2002.

[7] L. Deng, J. Droppo, and A. Acero, "Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 3, pp. 412–421, 2005.

[8] J. Droppo, A. Acero, and L. Deng, "Uncertainty decoding with SPLICE for noise robust speech recognition," *in Proc. ICASSP*, vol. I, pp. 57–60, May 2002.

[9] H. Liao and M. J. F. Gales, "Joint uncertainty decoding for noise robust speech recognition," in *Proc. Interspeech*, Lisbon, Portugal, Sep. 2005.

[10] V. Ion and R. Haeb-Umbach, "A novel uncertainty decoding rule with applications to transmission error robust speech recognition," *IEEE Trasaction on Audio, Speech and Language Processing*, vol. 16, no. 5, pp. 1047–1060, 2008.

[11] D. Kolossa, A. Klimas, and R. Orglmeister, "Separation and robust recognition of noisy, convolutive speech mixtures using time-frequency masking and missing data techniques," *IEEE Workshop on Application of Signal Processing to Audio and Acoustics*, 2005.

[12] A. H. Abdelaziz, S. Zeiler, and D. Kolossa, "Audio-visual speech recognition for uncertain acoustical observations," in *Proc. ITG Fachtagung Sprachkommunikation*, Braunschweig, Germany, September 2012.

[13] A. H. Abdelaziz and D. Kolossa, "Decoding of uncertain features using the posterior distribution of the clean data for robust speech recognition," in *Proc. Interspeech*, Portland, Oregon, USA, 2012.

[14] A. H. Abdelaziz, S. Zeiler, D. Kolossa, V. Leutnanty, and R. Haeb-Umbachy, "GMM-based significance decoding," *Accepted for ICASSP*, 2013.

[15] M. Cooke, J. Barker, S. Cunningham, and X. 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.

[16] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[17] A. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic Bayesian networks for audio-visual speech recognition," *EURASIP Journal on Applied Signal Processing*, vol. 11, pp. 1274–1288, 2002.

[18] R. F. Astudillo, "Integration of short-time fourier domain speech enhancement and observation uncertainty techniques for robust automatic speech recognition," Ph.D. dissertation, Technische Universität Berlin, 2010.

[19] J. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," ICSI, Tech. Rep. TR-97-021, 1997.

[20] Institute for Perception-TNO and Speech Research Unit-RSRE, retrieved November 2012. [Online]. Available: http://spib.rice.edu/spib/data/signals/noise/

[21] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, pp. 357–366, 1980.

[22] *Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms*, ETSI, ES.202.050 Std., 2003.

[23] G. Bradski and A. Kaehler, *Computer Vision with the OpenCV Library*. O'Reilly Media, 2008.

[24] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, 6th ed. Pearson, 2007.

[25] R. F. Astudillo, D. Kolossa, and R. Orglmeister, "Accounting for the uncertainty of speech estimates in the complex domain for minimum mean square error speech enhancement," in *Proc. Interspeech*, Brighton, United Kingdom, 2009.

[26] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio*, vol. 11, no. 5, p. 466475, 2003.