
Proceedings Paper:
Zeiler, S., Nicheli, R., Ma, N. orcid.org/0000-0002-4112-3109 et al. (2 more authors) (2016) Robust audiovisual speech recognition using noise-adaptive linear discriminant analysis. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) , 20-25 March 2016, Shanghai. IEEE , pp. 2797-2801. ISBN 9781479999880

https://doi.org/10.1109/ICASSP.2016.7472187

ROBUST AUDIOVISUAL SPEECH RECOGNITION USING NOISE-ADAPTIVE LINEAR DISCRIMINANT ANALYSIS

Steffen Zeiler, Robert Nickel†, Ning Ma◇, Guy J. Brown◇, Dorothea Kolossa**

*Insititute of Communication Acoustics, Ruhr-Universität Bochum, Germany
{steffen.zeiler, dorothea.kolossa}@rub.de
†Bucknell University, Lewisburg, PA, USA, {robert.nickel}@bucknell.edu
◇Department of Computer Science, University of Sheffield, Sheffield S1 4DP, UK
{n.ma,g.j.brown}@sheffield.ac.uk

ABSTRACT

Automatic speech recognition (ASR) has become a widespread and convenient mode of human-machine interaction, but it is still not sufficiently reliable when used under highly noisy or reverberant conditions. One option for achieving far greater robustness is to include another modality that is unaffected by acoustic noise, such as video information. Currently the most successful approaches for such audiovisual ASR systems, coupled hidden Markov models (HMMs) and turbo decoding, both allow for slight asynchrony between audio and video features, and significantly improve recognition rates in this way. However, both typically still neglect residual errors in the estimation of audio features, so-called observation uncertainties. This paper compares two strategies for adding these observation uncertainties into the decoder, and shows that significant recognition rate improvements are achievable for both coupled HMMs and turbo decoding.

***Index Terms*—** Audiovisual speech recognition, uncertainty-of-observation techniques, discriminative transformation.

1. INTRODUCTION

Human-machine interaction systems have achieved a considerable level of sophistication over the past decade. However, the reliable operation of such systems in practical scenarios has been a challenge. Systems routinely struggle with the variability of human expression, both visually and acoustically, and the ever-present effects of varying environments and noise. Machines that jointly consider visual and acoustic cues tend to achieve higher levels of robustness, because visual cues are complementary to acoustic ones, and features extracted from visual and acoustic information correlate with each other. It is thereby possible to recover information that is otherwise lost in each individual modality due to noise or occlusion. However, visual and acoustic features also provide information that is *unique* to each modality (through place and manner of articulation for example) so that degradations due to the variability of human expression itself can be curbed. Furthermore, it is well known that humans gain a benefit from audiovisual integration. Lip-reading is used by hearing impaired listeners, for example, and visual information may even override acoustic cues as demonstrated by the experiments by McGurk and MacDonald [2]. Joint audiovisual processing has, therefore, been successfully used in a number of applications such as automatic emotion recognition [3], speaker diarization [4], voice conversion [5], speech enhancement [6], speaker tracking [7], and the recognition of whispered speech [8].

Early successful approaches for joint audiovisual decoding in automatic speech recognition (ASR) were conceived, amongst others, by Neti *et al.* [9], Nefian *et al.* [10], Zhang *et al.* [11], and Kratt *et al.* [12]. The goal is generally to combine an acoustic recognition engine with an automatic lip-reading mechanism. A fundamental problem in the integration of corresponding visual and acoustic cues is that the two modalities are not perfectly synchronous, but exhibit time offsets caused by preparatory articulator/lip movements. These movements occur in anticipation of future phonemes, similarly to corresponding co-articulation effects in the acoustic signal. The two mechanisms that are currently considered to be most successful in integrating such visual and acoustic features are *coupled hidden Markov models* (CHMMs) [13] and *turbo decoding* (TD) [14]. CHMMs are formally different from regular HMMs in that internal states are addressed with a two-dimensional index instead of a one-dimensional index (see Fig. 1). One dimension of the index refers to the corresponding state in the audio-only stream and the other index refers to the corresponding state in the video-only stream. CHMMs are naturally able to account for asynchronicities between the two streams [10] and, thereby, enable the processing of audiovisual data.

An alternative to CHMMs was recently published by Receveur, Scheler, and Fingscheidt [14]. They recognized that the so-called *turbo* techniques developed in the context of error correcting channel codes [15] can also be applied to the information fusion problem in multimodal recognition tasks. They reported that their generalized turbo ASR approach outperformed conventional CHMMs with a significant reductions in word error rate [14].

Lastly, all types of ASR systems suffer from degradations in noisy and/or reverberant environments. The situation can be significantly improved, however, if not only a generic noise-reduction/dereverberation algorithm is applied, but also information about the *reliability* of the resulting feature vector is incorporated into the recognition process. Conventional techniques include uncertainty decoding (UD) and modified imputation (MI). Successful implementations of such *uncertain data techniques* for audiovisual speech recognition (AVSR) were proposed in [13] and [16].

In this paper we assess the use of a new *noise-adaptive linear discriminant analysis* method (NALDA) to fuse reliability information into the recognition process. NALDA was introduced in [17] in the context of audio-only ASR. Classical *linear discriminant analysis* (LDA) projects multidimensional data onto its most discriminative direction. In NALDA this direction is adaptively optimized

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

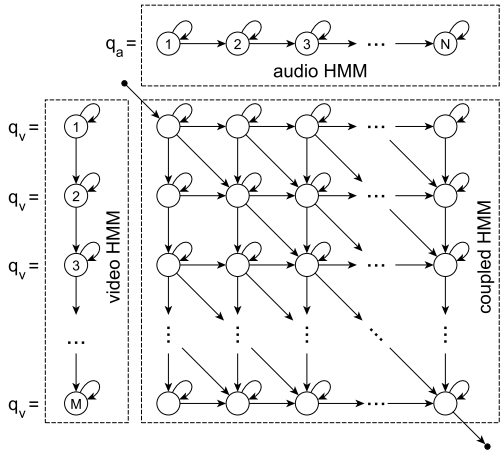

Fig. 1: Illustration of a CHMM with $N \times M$ coupled states. The marginal audio model has N audio states q_a and the marginal video model has M video states q_v. Entry and exit of a CHMM is limited to corner states (indicated by tiny black circles).

with respect to the estimated feature uncertainties. We extend the methodology to AVSR systems and show that NALDA delivers word recognition rates that are superior to conventional techniques.

2. AUDIOVISUAL SPEECH RECOGNITION

Audiovisual data differs from standard single-modality data with respect to its inherent asynchronicities: since speakers tend to bring articulators into position before phonation occurs, e.g. at the beginning of an utterance, the visual modality information can precede that of the acoustic modality by up to 120 ms [18]. Different model topologies have been proposed to deal with this issue. Ideas range from simply applying a standard HMM to concatenated features, the so-called *feature fusion* approach [9], to a wide range of so-called *decision fusion* approaches. Decision fusion can occur at different stages of the recognition process. *Early integration* fuses the information already at the state level. *Late integration* may go as far as recognizing audio and video data separately and then fusing decisions at the sentence-level [10].

In this paper, we consider two methods of classifier integration. Both allow for a certain degree of natural, asynchronous behavior[1], while at the same time also providing means to explicitly enforce some constraints on synchronicity. The first of these, the CHMM, is known to be superior to feature fusion as well as a wide range of other conventional strategies [10]. The other, more recently developed approach uses turbo decoding to integrate classifier information with provisions for modeling asynchronicities. Turbo decoding is shown to deliver a performance superior to CHMMs in [19]. We extended both models to incorporate the handling of observation uncertainties, as described in detail below.

2.1. Coupled HMM decoding

In a coupled HMM (CHMM) the joint state transition probability is modeled as a linear combination of marginal transition probabilities as illustrated in Fig. 1. An audio stream weight λ_C is used to capture the interactions among audio o_a and video observations o_v and their

[1]Given that the model topology is chosen appropriately.

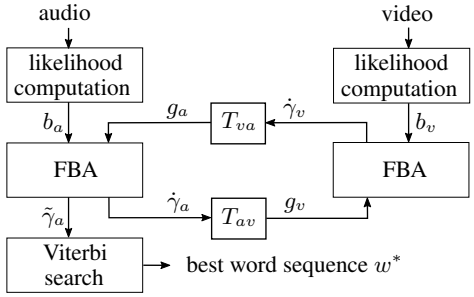

Fig. 2: Turbo decoding (TD) for AVSR. The left column comprises an FBA-based audio-only ASR system. For TD a second modality (video) is added and extrinsic probabilities $\dot{\gamma}_a$ and $\dot{\gamma}_v$ are exchanged between decoders. In the first TD iteration a flat prior $g_a(q_a) = 1,\ \forall q_a$ is used for the audio state posterior calculation. After a predefined number of iterations a best path search through the audio posteriors $\tilde{\gamma}_a$ reveals the final best word sequence w^*.

respective individual observation likelihoods b_a and b_v. For the *joint* audiovisual state likelihood we obtain

$$\mathrm{p}(o_a, o_v|q_a, q_v) = b_a(o_a|q_a)^{\lambda_C} \cdot b_v(o_v|q_v)^{1-\lambda_C}. \qquad (1)$$

For our experiments we used a token passing decoder to find the best word sequence w^* as the Viterbi path through a network of CHMM word models.

2.2. Turbo decoding

Turbo decoding [15] is an information fusion technique, which originated from a breakthrough in digital communication applications. More recently the turbo principle emerged as an alternative decoding scheme in multimodal speech recognition [20, 14] and proved to be useful for other applications such as blind speech separation [21] and speech enhancement [22].

TD is based on the iterative exchange of soft information, deduced from state posteriors, between different decoders. This extra information, g_a and g_v in Fig. 2, is used like a prior to modify the observation likelihoods b_a and b_v in the forward-backward algorithm (FBA). The modified audio and video likelihoods become

$$\tilde{b}_a(o_a|q_a) = b_a(o_a|q_a) \cdot g_a(q_a)^{\lambda_T \lambda_P}, \qquad (2)$$

$$\tilde{b}_v(o_v|q_v) = b_v(o_v|q_v) \cdot g_v(q_v)^{(1-\lambda_T)\lambda_P}, \qquad (3)$$

in which λ_T acts like an audio stream weight and the constant λ_P balances the likelihood and prior probability. From the FBA, we obtain new state posteriors $\tilde{\gamma}$, which subsume the likelihood, the prior probability and the extrinsic probability [14]. To find the extrinsic probability $\dot{\gamma}(q_t)$ for state q and frame t, we have to remove all excess information via

$$\dot{\gamma}(q_t) \propto \frac{\tilde{\gamma}(q_t)}{b(o_t|q_t) \cdot g(q_t)}. \qquad (4)$$

The final step of each such half-iteration is to map the extrinsic probabilities to the state space of the respective other decoder. This is done by a linear transformation.

$$g_a = T_{va}\, \dot{\gamma}_v \qquad \text{audio} \leftarrow \text{video} \qquad (5)$$

$$g_v = T_{av}\, \dot{\gamma}_a \qquad \text{video} \leftarrow \text{audio} \qquad (6)$$

The process of modified FBA followed by the deduction of extrinsic probabilities and their transfer to the corresponding state space is iterated for the audio and the video model a few, e.g. 4, times.

Despite objections against the applicability of plain forward-backward inference in loopy graphical models [23], we have experienced no convergence problems in our experiments.

3. USING MISSING AND UNCERTAIN DATA IN AVSR

When attempting ASR within natural settings, such as a home or office environment, the system is frequently confronted with significant levels of additive noise. The detrimental effect of noise on the recognition process can be substantially reduced with proper preprocessing of the recorded input signal [24]. The implicit noise power estimation of any type of noise reduction mechanism serves two purposes: (1) with knowledge of the noise power it becomes possible to (optimally) filter the incoming distorted speech to enhance the signal components of interest [1], and (2) the estimated noise power may serve as a gauge for the reliability level of each component [25]. High levels of noise would render a particular component less reliable. Low levels of noise would sway the recognition mechanism to place more confidence into any estimate derived from the associated speech component. Fusing information about missing and uncertain data into an ASR process typically requires three technical steps:

1. The execution of a speech enhancement algorithm that delivers: (a) an estimate of the underlying noise power at each point of a time-frequency decomposition of the incoming signal, and (b) an estimate of the underlying clean speech power at each time-frequency point.
2. A transformation of the estimated clean spectra and their associated noise powers into the domain of the recognition features, which includes the estimated feature vectors and their associated uncertainty measures.
3. Using the feature vectors and their uncertainties in a statistical recognition engine to decode the word sequence of the targeted underlying speech signal.

The implementation details of each of the three steps within our proposed method are described in the following three subsections.

3.1. Signal enhancement and uncertainty estimation

For the preprocessing of the signal we generally followed the recommendation ETSI ES 202 050 for an *advanced front-end feature extraction algorithm* after the European Telecommunications Standards Institute [26]. Our experimental data consisted of two-channel signals sampled at 16 kHz. We applied a simple delay-and-sum and a simple null-steering beamformer to derive an initial estimate of the targeted speech signal $\hat{x}_{\text{in}}[n]$ and an initial estimate of the noise signal $\hat{v}_{\text{in}}[n]$ [16]. Both estimates were converted into the STFT domain with a 400-sample Hamming window, a frame overlap of 240 samples, and an FFT length of 512 samples [1]. We use $\hat{X}_{\text{in}}(k,t)$ to denote the STFT of the initial signal estimate and $\hat{V}_{\text{in}}(k,t)$ for the STFT of the initial noise estimate. Parameter k represents the frequency index and t denotes the time frame index.

The STFT $\hat{X}_{\text{in}}(k,t)$ was subjected to a speech enhancement algorithm with Wiener gain, the *decision directed approach* for the estimation of the a-priori SNR $\xi(k,t)$ [1], and an *improved minima controlled recursive averaging* (IMCRA) for estimating the noise power $N(k,t)$ [27]. The estimated noise power $N(k,t)$ was weighted with gain-factor $\xi(k,t)/(1+\xi(k,t))$ to arrive at an estimate of the spectral uncertainties $\hat{\Sigma}_N(k,t)$ (see Nesta *et al.* [28]).

3.2. Uncertainty propagation

The enhanced signal spectral estimates $\hat{X}(k,t)$ and the associated spectral uncertainties $\hat{\Sigma}_N(k,t)$ were converted into the 13-dimensional MFCC domain after Astudillo *et al.* [29] via uncertainty propagation. Cepstral mean subtraction is applied [24]. We augmented our MFCC vector with the usual Δ and $\Delta\Delta$ coefficients (see [24] for example). The associated Δ and $\Delta\Delta$ uncertainties are augmented in the uncertainty vector accordingly. As a result we obtain a 39-dimensional audio recognition feature estimate $\boldsymbol{o_a}(t)$ and an associated 39-dimensional feature uncertainties vector $\hat{\Sigma}_{\boldsymbol{o_a}}(t)$ for each signal frame.

3.3. Uncertainty-based decoding

In conventional ASR systems, the observation likelihoods $b_a(o_a|q_a)$ and $b_v(o_v|q_v)$ are typically computed as Gaussian mixture models

$$b_{q_s}(\mathbf{o_s}) = p(\mathbf{o_s}|q_s) \sum_{m=1}^{M} W_{q_s,m} \cdot \mathcal{N}\left(\mathbf{o_s}; \boldsymbol{\mu}_{q_s,m}, \boldsymbol{\Sigma}_{q_s,m}\right), \quad (7)$$

where $W_{q_s,m}$, $\boldsymbol{\mu}_{q_s,m}$ and $\boldsymbol{\Sigma}_{q_s,m}$ are the parameters of the m^{th} Gaussian mixture component of state q in stream s ($s \in \{a, v\}$). Each Gaussian component density would be evaluated via

$$\mathcal{N}\left(\mathbf{o_s}; \boldsymbol{\mu}_{q_s,m}, \boldsymbol{\Sigma}_{q_s,m}\right) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}_{q_s,m}|}} \cdot \ldots$$
$$\exp(-\tfrac{1}{2} \left(\mathbf{o_a} - \boldsymbol{\mu}_{q_s,m}\right)^T \boldsymbol{\Sigma}_{q_s,m}^{-1} \left(\mathbf{x} - \boldsymbol{\mu}_{q_s,m}\right)). \quad (8)$$

We refer to models in which off-diagonal elements of covariance matrix $\boldsymbol{\Sigma}_{q_s,m}$ are forced to zero as *Gaussian-Density/Diagonal* (**GDD**) models. Models with fully populated covariance matrices are referred to as *Gaussian-Density/Full* (**GDF**) models.

In order to account for the possibly time-varying *reliability* of the audio stream, one may replace observation likelihoods of audio-streams in audio-only, coupled-HMM, or turbo decoders with likelihoods derived from uncertainty-of-observation techniques. These utilize estimates of the observation uncertainty $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{o_a}}(t)$ at each time frame. In uncertainty decoding (UD) [30], for example, the Gaussian component densities are updated with a time-dependent "correction" in covariance, i.e.

$$\boldsymbol{\Sigma}'_{q_a,m} = \boldsymbol{\Sigma}_{q_a,m} + \hat{\boldsymbol{\Sigma}}_{\boldsymbol{o_a}}(t). \quad (9)$$

Thus, the observation uncertainty is added to the covariance of each state output probability distribution. Uncertainty Decoding (denoted by **GDU** in the following tables) was used successfully for audio-visual speech recognition in [31]. In conjunction with uncertainty propagation techniques and stream weight optimization, however, the respective performance gains of UD become small.

In contrast to uncertainty decoding, noise-adaptive LDA transforms not only the covariance matrices of the Gaussian component models, but also the mean vectors. For this purpose, it uses the observation uncertainties $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{o_a}}(t)$ to determine the most discriminative feature transform matrix $\boldsymbol{W}_{\text{NALDA}}$ at each time frame t, as described in [17]. Once the input feature vectors have been mapped to the most discriminative D'-dimensional subspace by $\tilde{\boldsymbol{o}}_{\boldsymbol{a}}(t) = \boldsymbol{W}_{\text{NALDA}}(t)\boldsymbol{o_a}(t)$, the HMM output distributions are transformed for each state q_a by

$$\tilde{\boldsymbol{\mu}}_{q_a}(t) = \boldsymbol{W}_{\text{NALDA}}(t)\boldsymbol{\mu}_{q_a}. \quad (10)$$

Likewise, each covariance matrix is updated by

$$\tilde{\boldsymbol{\Sigma}}_{q_a}(t) = \boldsymbol{W}_{\text{NALDA}}(t)\boldsymbol{\Sigma}_{q_a}\boldsymbol{W}^{\top}_{\text{NALDA}}(t). \quad (11)$$

	Keyword Accuracies (%) with Oracle Uncertainties							Keyword Accuracies (%) with Estimated Uncertainties						
SNR	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	avg.	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	avg.
Video	72.20	72.20	72.20	72.20	72.20	72.20	72.20	72.20	72.20	72.20	72.20	72.20	72.20	72.20
Audio GDF	72.30	78.07	81.28	84.82	86.50	89.32	82.05	65.09	75.04	80.83	86.73	90.34	92.54	81.76
Audio GDD	72.11	77.00	81.91	84.58	87.21	89.59	82.06	71.90	79.05	82.56	87.75	91.64	91.60	84.08
Audio GDU	76.30	78.83	83.79	84.47	86.88	87.70	83.00	72.99	77.57	81.60	88.73	91.49	91.74	84.02
Audio GDN	82.72	86.81	88.49	90.71	92.17	92.44	88.89	74.00	78.94	85.19	90.93	92.40	93.30	85.79
CHMM GDF	82.52	85.14	86.58	88.32	89.88	90.33	87.13	76.67	82.56	87.30	89.74	92.37	93.88	87.09
CHMM GDD	82.78	85.93	87.95	89.86	90.96	92.32	88.30	84.72	85.81	88.68	90.47	91.22	92.09	88.83
CHMM GDU	82.39	84.40	85.29	85.85	87.85	88.08	85.64	83.63	84.59	87.77	88.97	91.18	90.64	87.80
CHMM GDN	87.32	89.23	91.28	92.56	93.71	93.86	91.33	84.13	87.59	90.28	92.40	93.36	93.43	90.20
Turbo GDF	84.49	86.92	88.17	89.31	91.08	92.04	88.67	81.79	86.41	90.11	91.53	93.98	**95.32**	89.86
Turbo GDD	84.46	88.06	89.42	90.92	92.58	93.50	89.82	85.75	88.58	90.45	92.16	93.68	93.52	90.69
Turbo GDU	88.45	89.08	90.54	91.33	92.79	92.13	90.72	84.34	87.57	89.67	91.48	93.60	92.71	89.89
Turbo GDN	**90.03**	**92.34**	**93.77**	**94.76**	**94.97**	**95.66**	**93.59**	**87.21**	**89.48**	**92.08**	**93.09**	**95.26**	95.12	**92.04**

Table 2: Keyword accuracies from our experiments with oracle and estimated uncertainties. Best results are marked in bold.

With these updated parameters and features, and with the reduced dimension D', the audio observation densities are evaluated according to (7). We use **GDN** to refer to a system with this type of NALDA-based uncertainty evaluation.

From our video data we extracted 31-dimensional LDA-transformed DCT coefficients of the mouth region as described in [6]. The likelihood computation of the video features involved standard diagonal Gaussian mixture models without the inclusion of any observation uncertainties.

4. EXPERIMENTAL SETUP AND RESULTS

For our experimental evaluation we used audio data from the first CHiME challenge [32] in combination with matching video data from the GRiD corpus [33]. The recordings consist of 1000 sentences spoken by 33 talkers each. All utterances include the annunciation of a letter (A...Z, excluding W) and a digit (0...9). Audio and video files are not start/endpoint-aligned between the CHiME and the GRiD corpus. We therefore performed a start/end-point matching via the word alignment files provided on `http://spandh.dcs.shef.ac.uk/gridcorpus/`.

The entire set of training data was used in the initial model training. Subsequently, development data of the first five speakers was used to adjust all free parameters, i.e. the audio stream weights for the CHMM and the TD decoder. Table 1 shows the corresponding values that we obtained for the four different types of decoders. We set $\lambda_P = 0.1$ for all TD experiments and $D' = 37$ for GDN.

The experimental results are shown in Table 2. We measured the success of each of the considered recognition schemes via the *keyword accuracy*, i.e. the percentage of correctly identified letters and digits. For the *oracle uncertainty* results on the left-hand side of Table 2, the uncertainties $\hat{\mathbf{\Sigma}}_{\boldsymbol{o_a}}(i)$ are given by the "true" squared error between the features of the respective clean data and the processed features. The right-hand side shows the keyword accuracies for *estimated uncertainties,* computed after Sections 3.1 and 3.2.

All four previously introduced likelihood functions were compared and there are three different decoding mechanisms at play: Standard, single-modality decoders are used to generate the audio-only and video-only results, and coupled HMM and turbo decoding were used to obtain the audiovisual results.

In general, the performance difference for estimated and oracle uncertainties is quite small for SNRs ≥ 0 dB, but the results deviate more when the SNR is negative. In nearly all SNR conditions, the best performance is obtained with turbo decoding and NALDA-based uncertainty evaluation.

Oracle uncertainties				
pdf / Recognizer	GDF	GDD	GDU	GDN
CHMM (λ_C)	0.8	0.8	0.9	0.9
TD (λ_T)	0.7	0.7	0.8	0.8
Estimated uncertainties				
pdf / Recognizer	GDF	GDD	GDU	GDN
CHMM (λ_C)	0.8	0.7	0.7	0.8
TD (λ_T)	0.8	0.6	0.6	0.7

Table 1: Stream weights λ_C and λ_T for all tested combinations of the uncertainty estimators, recognition types and pdf types.

5. CONCLUSIONS

We have considered the use of observation uncertainties in audiovisual speech recognition, using coupled HMMs and turbo decoding. As already noted by [31], in coupled HMM decoding, stream weight adaptation and uncertainty compensation by UD both provide significant advantages in isolation, but using uncertainty compensation in addition to optimized stream weighting provides only small benefits. This finding was replicated in our experiments. However, noise adaptive LDA [17], another, more recent uncertainty-of-observation technique, has proven to be of significant value in this context.

Additionally, we have incorporated uncertainty-of-observation-techniques into turbo decoding, an approach to audiovisual integration that was recently introduced in [14]. Here, again, uncertainty decoding was only of rather small benefit when optimized stream weights were used, whereas noise-adaptive LDA has shown large benefits for optimal oracle stream weights, and has been valuable with estimated uncertainties as well.

In the presented approach, fixed stream weights were used in all experiments. Optimization of stream weights on a frame-by-frame basis has proven its merit for coupled-HMM systems in [34]. It will be interesting to extend this technique to the presented turbo-decoding system, adapting the stream weight according to estimated SNR, observation uncertainty, and model-based reliability measures like dispersion and entropy, in order to also consider the time-varying utility of video information in the process.

6. REFERENCES

[1] P. C. Loizou, *Speech Enhancement, Theory and Practice*, CRC-Press, 2007.

[2] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[3] X. Zhibing, T. Yun, and G. Ling, "A new audiovisual emotion recognition system using entropy-estimation-based multimodal information fusion," in *Proc. ISCAS*, May 2015, pp. 726–729.

[4] F. Vallet, S. Essid, and J. Carrive, "A multimodal approach to speaker diarization on TV talk-shows," *IEEE Trans. on Multimedia*, vol. 15, no. 3, pp. 509–520, April 2013.

[5] K. Sawada, M. Takehara, S. Tamura, and S. Hayamizu, "Audio-visual voice conversion using noise-robust features," in *Proc. ICASSP*, May 2014, pp. 7899–7903.

[6] A. H. Abdelaziz, S. Zeiler, and D. Kolossa, "Twin-HMM-based audiovisual speech enhancement," in *ICASSP*, 2013.

[7] V. Kilic, M. Barnard, W. Wenwu, and J. Kittler, "Audio constrained particle filter based visual tracking," in *Proc. ICASSP*, May 2013, pp. 3627–3631.

[8] F. Xing, C. Busso, and J.H.L. Hansen, "Audio-visual isolated digit recognition for whispered speech," in *19th Europ. Sig. Proc. Conf.*, Aug 2011, pp. 1500–1503.

[9] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio-visual speech recognition," Tech. Rep. WS00AVSR, Johns Hopkins University, CLSP, 2000.

[10] A. V. 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.

[11] Y. Zhang, Q. Diao, S. Huang, W. Hu, C. Bartels, and J. Bilmes, "DBN based multi-stream models for speech," in *Proc. ICASSP*, April 2003, vol. 1, pp. I–836 – I–839.

[12] J. Kratt, F. Metze, R. Stiefelhagen, and A. Waibel, "Large vocabulary audio-visual speech recognition using the Janus speech recognition toolkit," in *DAGM-Symposium*, 2004.

[13] A. Vorwerk, S. Zeiler, D. Kolossa, R. Astudillo, and D. Lerch, *Robust Speech Recognition of Uncertain or Missing Data - Theory and Applications*, chapter "Use of Missing and Unreliable Data for Audiovisual Speech Recognition", Springer, 2011.

[14] S. Receveur, D. Scheler, and T. Fingscheidt, "A turbo-decoding weighted forward-backward algorithm for multimodal speech recognition," in *Proc. of 5th Int. Workshop on Spoken Dialog Syst.*, Napa, California, January 17-20, 2014.

[15] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: Turbo-codes," in *Proc. ICC*, Geneva, 1993, vol. 2, pp. 1064–1070.

[16] D. Kolossa, R. F. Astudillo, A. Abad, S. Zeiler, R. Saeidi, P. Mowlaee, J. P. da Silva Neto, and R. Martin, "CHiME Challenge: Approaches to robustness using beamforming and uncertainty-of-observation techniques," in *CHiME*, 2011.

[17] D. Kolossa, S. Zeiler, R. Saeidi, and R. F. Astudillo, "Noise-adaptive LDA: A new approach for speech recognition under observation uncertainty," *IEEE Signal Processing Letters*, vol. 20, no. 11, pp. 1018–1021, 2013.

[18] J. Luettin, G. Potamianos, and C. Neti, "Asynchronous stream modelling for large vocabulary audio-visual speech recognition," in *Proc. ICASSP*, 2001, pp. 169–172.

[19] D. Scheler, S. Walz, and T. Fingscheidt, "On iterative exchange of soft state information in two-channel automatic speech recognition," in *Proc. ITG Facht. Sprachkomm.*, 2012.

[20] S.T. Shivappa, B. D. Rao, and M. M. Trivedi, "Multimodal information fusion using the iterative decoding algorithm and its application to audio-visual speech recognition," in *Proc. ICASSP*, 2008, pp. 2241–2244.

[21] D. H. Tran Vu and R. Haeb-Umbach, "Blind speech separation exploiting temporal and spectral correlations using turbo decoding of 2D-HMMs," in *Proc. EUSIPCO*, 2013.

[22] S. Hong and H. Maitre, "Turbo iterative signal processing," in *Proc. DSP/SPE*, 2009, pp. 495–500.

[23] R. J. McEliece, D. MacKay, and J. Cheng, "Turbo decoding as an instance of Pearl's belief propagation algorithm," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 2, pp. 140–152, 1998.

[24] X. Huang, A. Acero, and H. W. Hon, *Spoken Language Processing, A Guide to Theory, Algorithm, and System Development*, Prentice Hall, Upper Saddle River, New Jersey, 2001.

[25] D. Kolossa, S. Zeiler, A. Vorwerk, and R. Orglmeister, "Audiovisual speech recognition with missing or unreliable data," in *Proc. AVSP*, 2009, pp. 117–122.

[26] European Telecommunications Standards Institute, "Advanced front-end feature extraction algorithm," *ETSI ES 202 050 V1.1.5*, January 2007.

[27] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Trans. Speech and Audio Proc.*, vol. 11, no. 5, pp. 466–475, Sep 2003.

[28] F. Nesta, M. Matassoni, and R. F. Astudillo, "A flexible spatial blind source extraction framework for robust speech recognition in noisy environments," in *Proc. CHiME*, 2013, pp. 33–40.

[29] R. Astudillo, D. Kolossa, and R. Orglmeister, "Propagation of statistical information through non-linear feature extractions for robust speech recognition," in *Proc. MaxEnt2007*, 2007.

[30] 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, May 2005.

[31] G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos, "Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition," *IEEE Trans. Aud., Sp., Lang. Proc.*, vol. 17, no. 3, pp. 423–435, 2009.

[32] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME speech separation and recognition challenge," *Computer Speech and Language*, vol. 27, no. 3, pp. 621–633, 2013.

[33] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 120, no. 5, pp. 2421–2424, November 2006.

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