ISCA Archive
http://www.isca-speech.org/archive

AVSP 2009 - International Conference
on Audio-Visual Speech Processing
University of East Anglia, Norwich, UK
September 10-13, 2009

Audiovisual Speech Recognition with Missing or Unreliable Data

Dorothea Kolossa, Steffen Zeiler, Alexander Vorwerk, Reinhold Orglmeister

Electronics and Medical Signal Processing Group, TU Berlin
{dorothea.kolossa,alexander.vorwerk,reinhold.orglmeister}@tu-berlin.de,
s.zeiler@ee.tu-berlin.de

Abstract

In order to robustly recognize distorted speech, use of visual information has been proven valuable in many recent investigations. However, visual features may not always be available, and they can be unreliable in unfavorable recording conditions. The same is true for distorted audio information, where noise and interference can corrupt some of the acoustic speech features used for recognition. In this paper, missing feature techniques for coupled HMMs are shown to be successful in coping with both uncertain audio and video information. Since binary uncertainty information may be easily obtained at little computational effort, this results in an effective approach that can be implemented to obtain significant performance improvements for a wide range of statistical model based audiovisual recognition systems.

Index Terms: missing data techniques, audiovisual speech recognition, coupled HMM

1 Introduction

Robustness of speech recognition can be significantly improved by multi-modal and notably audio-visual speech recognition. For this purpose, both HMMs and graphical models have been successfully utilized [1, 2, 3, 4]. In order to further improve robustness, missing feature recognition offers additional performance gains.

Uncertainty compensation has already been employed for this purpose in [5], using uncertainty decoding to deal with unreliable features. However, the presented method is shown to work well in the mel-spectrum domain, whereas here, a simple and computationally efficient approach for RASTA-PLP-cepstra is shown. These features have been shown to be more robust both with respect to noisy and reverberant conditions [6].

This paper is organized as follows. At first, Section 2 will introduce the audiovisual speech recognizer JASPER, which will be used for all subsequent experiments. This system is based on coupled HMMs and allows for asynchronous streams as long as synchrony is again achieved at word boundaries. Next, in Section 3, the feature extraction and uncertainty estimation are presented. In Section 4, the utitlized strategy for multi-stream missing feature recognition is described. Results for this system on the GRID database, a connected word small-vocabulary audiovisual database, are given in Section 5. These results and further implications are discussed in Section 6.

2 Audiovisual Recognition System

Audiovisual speech recognition and lipreading can both be carried out using the Java Audiovisual SPEech Recognizer JASPER, developed for the purpose of robust single- or multistream speech recognition. The system allows for a tight integration of the MATLAB and JAVA environments and capabilities, with an interface that lets preprocessing and feature extraction be carried out in MATLAB, whereas model training and recognition take place in JAVA. It is based on a flexible token passing architecture applicable for a wide range of statistical speech models, which is described in more detail below.

2.1 System Architecture

JASPER is based on an abstract model in which connected word recognition is viewed as a process of passing tokens around a transition network [7]. Within this network, each vocabulary element is represented by a word model. These word models are statistical descriptions of the evolution of the feature stream within the associated words. Since the token passing architecture only requires a narrow interface of the word models, these may be realized e.g. as conventional HMMs, coupled or product HMMs or even templates or a range of graphical models.

Fig. 1 shows an example of a possible word net structure.

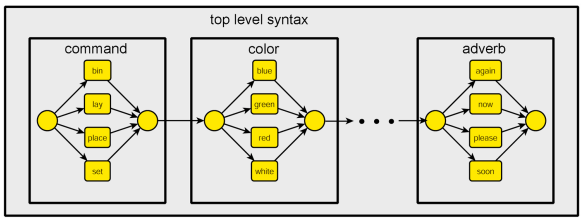

Figure 1: An example of a word network for recognition of the GRID grammar described in Section 5.1. Link nodes are depicted by circles, non-terminal nodes by bold black rectangles and word models are shown as yellow rounded rectangles.

In addition to the word models, further network elements are link nodes and non-terminal nodes. Link nodes serve as connections between non-terminal nodes and word models and associate a linked list of possible word alternatives to all tokens passing through them. Non-terminal nodes allow grouping of nodes into different hierachical levels. The highest level in the network is the non-terminal node that represents the entire language model. The elements on the lowest level are the word models.

The recognition process starts with a single token entering the top level syntax. Every timestep is split into two half steps. At first, all link nodes propagate their tokens down the hierarchy, until the lowest level, the word models, are reached. The actual calculations of word model log-likelihoods given the observed fea-

tures happen in the second phase. At the end of the second phase all link nodes collect tokens from their incoming connections and build a connected list of the n-best tokens. Tokens are ranked by a score, corresponding to their accumulated log-likelihood, and a global, adaptive threshold is used for efficient pruning. This process is iterated until a complete observation sequence has been processed and the outgoing link node of the top level syntax contains the token with the highest score given the model and the observed data.

2.2 Audiovisual Recognition using Coupled HMMs

For audiovisual speech recognition, two streams of feature vectors are available. $o_a(t)$ denotes the acoustical and $o_v(t)$ the visual feature vector at time t. These are not necessarily synchronized. The cause of asynchronicities lies in part in recording conditions, since sampling rates may differ. Such technical influences may be compensated by synchronous sampling and interpolation. However, other causes of asynchronicities are rooted in the speech production process itself, in which variable delays between articulator movements and voice production lead to time-varying lags between video and audio. Such lags may have a duration of up to 120ms, which corresponds to the duration of up to an entire phoneme [8]. Therefore, it is of great importance to account for variable delays between modalities, when recognition is to perform optimally.

In order to allow for such variable delays, a number of alternatives exist [2], like multi-stream HMMs, coupled HMMs, product HMMs or independent HMMs. These differ especially in the degree of required synchrony between modalities, from the one extreme of independent HMMs, where both feature streams can evolve with no coupling whatsoever, to the other extreme of multi-stream HMMs, in which a state-wise alignment is necessary or at least implicitly assumed.

As a reasonable compromise, coupled HMMs allow for both streams to have lags or evolve at different speeds, as long as they are again synchroneous at all word boundaries. Since this introduces some contraints but does not force unachievable frame-by-frame alignment, the following work is based on a realization of coupled HMMs in the above token passing framework.

2.3 Two-Stream Realization of Word Models

In coupled hidden markov models (CHMMs), both feature vector sequences are retained as separate streams. As generative models, the CHMM can describe the probability of both feature streams jointly as a function of a set of two discrete, hidden state variables, which evolve analogously to the single state variable of a conventional HMM.

Thus, the CHMMs have a two-dimensional state $\mathbf{q}$ which is composed of an audio and a video state, q_a and q_v, respectively, which can be seen in Fig. 2.

Each possible sequence of states through the model represents one possible alignment with the sequence of observation vectors. To evaluate the likelihood of such an alignment, each state pairing is connected by a transition probability, and each state is associated with an observation probability distribution.

The transition probability and the observation probability can both be composed from the two marginal HMMs. Then, the coupled transition probability becomes

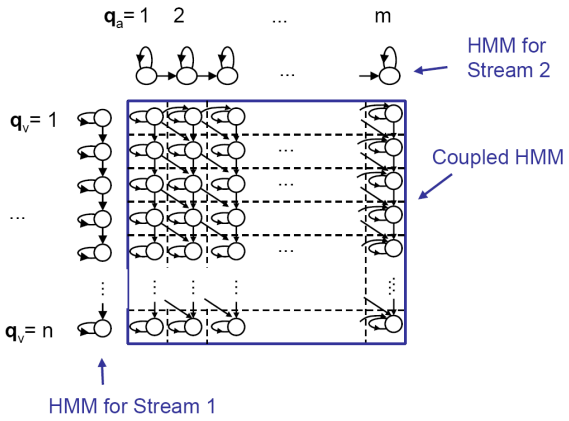

Figure 2: A coupled HMM consists of a matrix of interconnected states, which each correspond to the pairing of one audio- and one video-HMM-state, q_a and q_v, respectively.

$$\begin{aligned} &\mathrm{p}\Big(q_a(t+1)=j_a, q_v(t+1)=j_v | q_a(t)=i_a, q_v(t)=i_v\Big) \\ &\qquad = a_a(i_a,j_a)\cdot a_v(i_v,j_v) \end{aligned}$$

where $a_a(i_a,j_a)$ and $a_v(i_v,j_v)$ correspond to the transition probabilities of the two marginal HMMs, the audio-only and the video-only single-stream HMMs, respectively.

For the observation probability, both marginal HMMs could equally be composed to form a joint output probability by

$$\mathrm{p}(\mathbf{o}|\mathbf{i}) = b_a(o_a|i_a)\cdot b_v(o_v|i_v). \tag{1}$$

Here $b_a(o_a|i_a)$ and $b_v(o_v|i_v)$ denote the output probability distributions for both single streams.

However, such a formulation does not allow to take into account the different reliabilities of audio and video stream. Therefore, Eq. (1) is commonly modified by an additional stream weight γ as follows

$$\mathrm{p}(\mathbf{o}|\mathbf{i}) = b_a(o_a|i_a)^{\gamma}\cdot b_v(o_v|i_v)^{1-\gamma}. \tag{2}$$

This approach, described in more detail e.g. in [1], is also adopted in JASPER and forms the basis of all presented experiments.

3 Feature Extration

3.1 Audio Feature Extraction

RASTA-PLP-coefficients are designed to concentrate on features with a certain rate of change, namely that rate of change which is typical of speech. For that purpose, features are band-pass filtered. This makes them more robust to variations in room transfer function and in speakers, and to changes caused by background noise varying more quickly or slowly than speech signals [6].

In the presented experiments, 12 RASTA-PLP cepstrum coefficients and their first and second derivatives were used, which were obtained from a power spectrum of the speech signal using a window size of 25ms with 15ms overlap. RASTA-filtering takes

place in the log-bark-spectrum, using the transfer function

$$H(z) = \frac{0.2z^4 + 0.1z^3 - 0.1z^{-3} - 0.2z^{-4}}{z^3(z - 0.94)} \quad (3)$$

given in [6]. Subsequently, the LPC-cepstrum is obtained from the RASTA-filtered log spectrum, using an LPC model order of 12. For these computations, the rastamat toolbox was used, which is available from `http://labrosa.ee.columbia.edu/matlab/rastamat/`. Finally, first and second order time derivatives were appended to the 12 RASTA-PLP-cepstral coefficients, which further improved robustness.

3.2 Visual Feature Extraction

3.2.1 Face Detection

The visual feature extraction consists of two main steps, first finding the face and mouth region and subsequently extracting the relevant visual features from those regions. For the very first frame of every video sequence, a detailed search for the face and mouth regions is conducted. Following frames use the coherence of image content between frames to find the mouth region with a less complex search strategy. If an uncertainty-threshold for the mouth-ROI position is reached, a full search is required. This happens on average only once every second.

For the face segmentation, the YCbCr colorspace is used. Based on probability density functions (pdfs) learned from 187 example images, image pixels are classified as skin and non-skin pixels. For a potential face candidate, two constraints must be met. A predefined number of pixels inside a connected region is required, i.e. the face region needs a sufficient size, and topologically, the region should have at least one hole. Holes in the homogeneous skin colored region are usually caused by the eyes or the mouth, especially in an opened position.

After a suitable image region has been found for which all plausibility tests are positive, an ellipse with the same center of gravity is fitted to that region. The height to width ratio is adjusted to a typical value of $\frac{3}{2}$ to give ample space for all facial features.

The orientation of the fitted ellipse is used to compensate for a possible lateral head inclination.

The rotation angle is obtained from the main axis of the ellipse, and an alternative method to estimate the rotation is used for verification. This second estimate of the rotation angle is calculated from the center-positions of the three largest holes in the face region. If the difference between both angles is smaller than a predefined value, the rotation angle is accepted. With a reliable estimate of the angle, the image is rotated around the center of gravity of the ellipse so that the face is vertically oriented in an upright position, otherwise no rotation takes place. An example for an extracted face region can be seen in Fig. 3.

3.2.2 Mouth Detection

Typically, a face has a number of prominent features in the horizontal direction (eyelids, eyebrows, nostrils, lips). These properties of human faces are used to guide the search into image regions with a high probability for the desired facial features. A Sobel edge filter is used to extract horizontal edges. Rows in this edge image are summed and the resulting column vector is low-pass filtered to give a smooth approximation of the vertical image profile. Local maxima are recorded in a list sorted by the strength of the maximum.

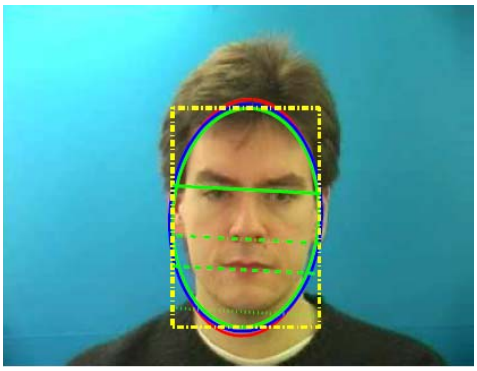

Figure 3: An example of an extracted face region. The green ellipse represents the accepted face hypothesis. The yellow square is the selected face region. The green lines correspond to the image rows with the highest probability for horizontally oriented facial features.

The list is used to find the most probable positions of 6 different facial features (hair, hairline, eyebrows, eyes, nose and mouth). A training set of 80 handlabeled images was used to calculate the optimal parameters for Gaussian approximations of the probability density functions for these facial features from line profiles. An image region with a suitable size around the most probable positions for eyes and mouth is used for block matching. Templates for eyes and mouth in opened as well as closed configuration are used for a normalized cross correlation with the image regions. If the resulting maxima in the correlation image reach a predefined threshold, the center positions for eyes and mouth are accepted.

It is asumed that the distance between the centers of the eyes is almost the same as the width of the mouth. With a ratio of $\frac{1}{2}$ between width and height of the mouth, an approximate region of interest (ROI) for the mouth is found.

Inside this region both corners of the mouth are detected. For this purpose, the green channel is thresholded to produce an approximation of the lip region. Starting from the left- and rightmost columns of the mouth-ROI going towards the middle, the first pixel exeeding the threshold is searched for. These pixels are very close to the true corners of the mouth. Both corner points are used to calculate a normalised mouth region, depicted in Fig. 4.

Figure 4: Example for a sequence of extracted mouth regions.

3.2.3 Feature Extraction

The properly selected and normalized mouth region is DCT transformed. At the moment, the first 64 coefficients of the two dimensional DCT-II are used as observations for the video model of the coupled HMM speech recognizer. Future versions will use parametric features such as snakes and optical flow based features in combination with linear discriminant analysis for dimension reduction.

4 Robust Recognition of Uncertain Features

Recognition of speech in noisy or otherwise difficult conditions can greatly profit from so-called *missing feature approaches*. In these methods, those features within a signal, which are dominated by noise or distorted by other detrimental effects, are considered "'missing"', and subsequently disregarded in speech recognition. [9]. This paradigm is also implemented in JASPER, which carries out missing feature recognition on both audio- and video features. In the following section, a short overview of missing feature recognition is given, followed by an explanation of how uncertainties are derived for both the audio and the video stream. Finally, the integration of both uncertain feature streams in a *missing-feature coupled HMM* is described.

4.1 Missing Feature Theory

When some parts of a speech signal are occluded by noise or interference, missing feature theory allows the recognizer to concentrate only on the reliable features. Both continuous-valued and binary methods exist to consider such uncertainties in the recognition process. In the current application, binary uncertainties have been considered, due to their advantages regarding computational complexity.

Where binary uncertainties are concerned, two main approaches can be distinguished, marginalization and imputation [10]. In both cases, a binary mask is necessary, which labels the unreliable regions in the feature domain. These uncertainty values are given in the following by $u_a(k,t)$, which denotes a feature-wise and frame-wise uncertainty for the audio feature stream, and by $u_v(t)$, to denote a frame-wise uncertainty for the video stream.

In order to recognize these uncertain features, the approach of marginalization has been employed. In that case, the output probability of an HMM with M Gaussian mixtures, which is usually computed for a given state q by

$$b\big(o(t)\big) = \sum_{m=1}^{M} w_m \cdot \mathcal{N}\big(o(t), \mu_{q,m}, \Sigma_{q,m}\big) \tag{4}$$

with $\mu_{q,m}$ and $\Sigma_{q,m}$ as mean and covariance matrix of mixture m and w_m as the mixture weight, is modified as follows

$$b\big(o(t)\big) = \sum_{m=1}^{M} w_m \cdot \mathcal{N}\big(o^r(t), \mu^r_{q,m}, \Sigma^r_{q,m}\big). \tag{5}$$

Here, $o^r(t)$ stands for a reduced feature vector, which only contains those components k that are reliable at the given time, i.e. for which $u_a(k,t) = 0$. Similarly, $\mu^r_{q,m}$ is a reduced mean vector and $\Sigma^r_{q,m}$ is the reduced covariance matrix, from which all rows and columns k with $u_a(k,t) = 1$ have been removed.

4.2 Video Feature Uncertainties

To compute the uncertainties of video features, a simple, frame-wise method is used. For this purpose, a speaker-dependent mouth-model is trained on 20 hand labelled sequences, which were selected to contain at least 30 wrongly estimated mouthregions in total. From these sequences, intact mouth regions are learned separately from a model combining both non-mouth and cropped-mouth regions. The model consists of a single Gaussian pdf of the same first 64 DCT coefficients, which are also used for HMM training. Therefore, no additional feature extraction needs to take place.

Subsequently, a frame will be counted as reliable, if the log-likelihood p_m of the trained mouth model,

$$p_m = \log \frac{1}{\sqrt{(2\pi)^d |\Sigma_m|}} e^{-\frac{1}{2}(o_v(t)-\mu_m)' \Sigma_m^{-1} (o_v(t)-\mu_m)}$$

with mean μ_m and covariance matrix Σ_m exceeds that of a non-mouth model

$$p_n = \log \frac{1}{\sqrt{(2\pi)^d |\Sigma_n|}} e^{-\frac{1}{2}(o_v(t)-\mu_n)' \Sigma_n^{-1} (o_v(t)-\mu_n)}$$

with parameters μ_n and Σ_n. Thus, the uncertainty $u_v(t)$ of all video features in frame t is given by

$$u_v(t) = \begin{cases} 0 & \text{for } p_m \geq p_n, \\ 1 & \text{otherwise.} \end{cases} \tag{6}$$

A typical example of the resultant labelling of video frames can be seen in Fig. 5.

Figure 5: Mouth regions and their associated labels. A green square means that the frame has been counted as reliable, otherwise, a red square is shown.

4.3 Audio Feature Uncertainties

For audio feature uncertainties, numerous approaches exist to estimate either binary or continuous-valued uncertainties. However, most binary uncertainty values are used for spectra and log-mel-spectra, which are not very robust to variations in the speaker, the environment or the background noise. In contrast, a number of more elaborate mechanisms for estimating continuous uncertainty values in the feature domain have been developed over the past years, see e.g. [11, 12].

Here, in contrast to the above approaches, only binary uncertainties are considered, since audiovisual recognition poses even more requirements for restricting the computational burden of HMM output density evaluations. In contrast to continuous-valued uncertainties, binary missing feature approaches are more efficient to evaluate, and will therefore be used in the following.

To estimate the reliability of the audio feature vector $o_a(t)$ at time t, its value is compared to that of a background noise estimate n_a, which is obtained in the time domain during the first 250ms. Since the audio feature vector is given in the RASTA-PLP-cepstrum domain, the background noise estimate n_a is also transformed to the same domain and extended to the length of the audio feature vector. These two vectors are then compared, and a reliability decision for each of the $k = 1 \ldots 36$ feature components is made by

$$u_v(k,t) = \begin{cases} 0 & \text{for } n_a(k,t) < 0.9 \cdot o_a(k,t), \\ 1 & \text{otherwise} \end{cases} \quad (7)$$

i.e. a feature is deemed unreliable if the value of the background noise feature exceeds 90% of the observed signal feature.

This is a fairly simple approach which is only suitable for stationary background noise. Further work will concentrate on integrating adaptive background noise estimates such as those obtainable from IMCRA (improved minima controlled recursive averaging) and related techniques.

4.4 The missing feature coupled HMM

The probability evaluation is carried out by means of marginalization, as described above in Section 4.1. In the case of coupled HMMs, the output probability computation is factorized into two streams, as given by Eq. (2). This means that marginalization can also be carried out independently for each stream. As a final optimization, since the video feature uncertainty is only computed once for each frame and extends to the entire video feature vector at that time, only the following expression needs to be evaluated at each frame and for each HMM state $\mathbf{i} = (i_a i_v)$.

$$\mathrm{p}(\mathbf{o}|\mathbf{i}) = \begin{cases} b_a(o_a|i_a)^\gamma \cdot b_v(o_v|i_v)^{1-\gamma} & \text{for } u_v(t) = 0, \\ b_a(o_a|i_a)^\gamma & \text{for } u_v(t) = 1. \end{cases} \quad (8)$$

This simplified version is due to that fact, that $\mathrm{p}(o_v|q_v) = 1$, when the entire feature vector is unreliable.

5 Experiments and Results

5.1 Database

The GRID database is a corpus of high-quality audio and video recordings of 1000 sentences spoken by each of 34 talkers [13]. Sentences are simple, syntactically identical phrases of the form `<command:4><color:4><preposition:4><letter:25><digit:10><adverb:4>`, where the number of choices for each component is indicated. The corpus is available on the web for scientific use at `http://www.dcs.shef.ac.uk/spandh/gridcorpus/`.

5.2 Experimental Setup

To evaluate the missing feature concepts for audiovisual speech recognition, corrupted video sequences are required. However, the mouth region is almost always clearly visible and it is found with a large accuracy, normally well above 99%. In order to test the algorithm under difficult conditions, two especially problematic speakers, which caused failures in the ROI selection at least occasionally, were selected for testing. For these two speakers, numbers 16 and 34, a mouth region is still correctly extracted in 99.1% of the cases. Therefore the tests were limited to only that subset for which at least one mouth region is either invisible or wrongly selected. The resulting test set consists of 318 sentences, in which an accurate mouth region is available in 94.6% of the frames. On this test set, the audio information is additionally distorted by adding white noise at signal to noise ratios (SNRs) ranging from 0dB to 30dB.

5.3 Audiovisual Recognition Results

Fig. 6 shows the overall recognition results for those two speakers of the GRID database, for whom accurate mouth detection was most problematic with the chosen approach. The recognition result is given as the accuracy in percent, PA, defined by $PA = \frac{N-D-I-S}{N}$, with N as the number of reference labels, D the deletions, S the substitutions and I the insertions.

As can be seen, both video-only and audio-only recognition profit notably from the use of missing feature techniques. Additionally, the performance improvement from audiovisual

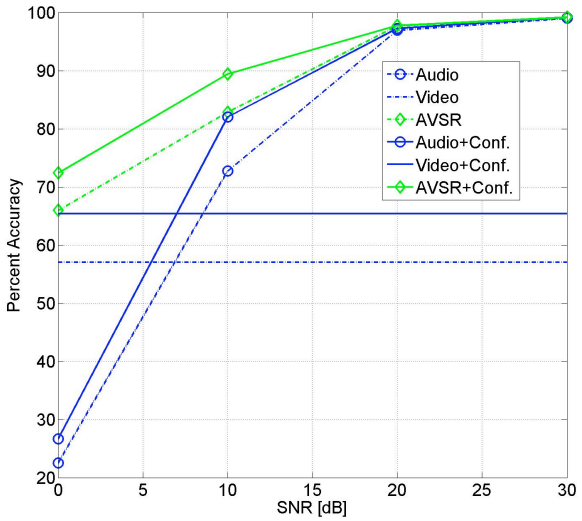

Figure 6: Recognition results for audio-only, video-only and audiovisual recognition. Dash-dotted lines correspond to conventional recognition, whereas bold lines indicate results using estimated audio and video confidences, i.e. using missing feature techniques.

CHMM recognition is significant, especially in the low-SNR range. Here, the joint recognition always improves the accuracy when compared to the better of the two single-stream recognizers, and often comes close to halving the error rate. This is also shown in Fig. 7 which displays the relative error rate reductions achieved by audiovisual missing feature recognition, when compared to the better one of the two single-stream recognizers.

In all experiments, the stream weight γ for the audiovisual recognizer and that for the audiovisual recognizer with confidences, γ_c, were separately adjusted to their optimal values shown in Table 1. This adjustment could be carried out automatically based on an SNR estimate. Ideally, however, no adjustment should be necessary at all. As Table 2 indicates, the use of confidences appears to be a step in the right direction. Here, the influence of the stream weight on the accuracy PA is shown. It is given in terms of the mean absolute variation of the accuracy relative to

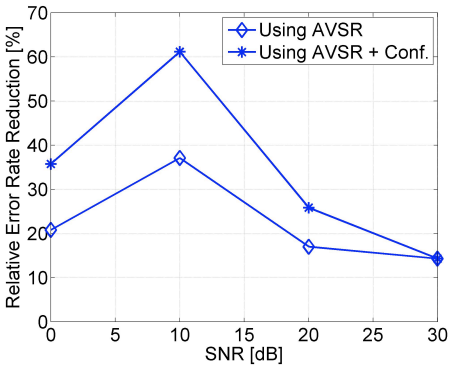

Figure 7: Relative error rate reductions achieved by audiovisual recognition, with and without the use of binary confidences.

Table 1: Optimal Settings of Stream Weights.

SNR	0	10	20	30
γ	0.5	0.89	0.95	0.98
γ_c	0.7	0.89	0.93	0.97

the stream weight, i.e. the average value of $\left|\frac{\Delta PA}{\Delta \gamma}\right|$. The numbers were obtained by averaging over all results from the parameter tuning phase, during which 5 different stream weight settings were tested per SNR.

Table 2: Mean variability of accuracy $\left|\frac{\Delta PA}{\Delta \gamma}\right|$ relative to changes in stream weight γ.

SNR	0dB	10dB	20dB	30dB
Conventional AVSR	37.6	88.8	17.5	0
Missing features	18.6	48.8	5.0	0

6 Conclusions

As shown in previous publications, audiovisual integration gives significant error rate reductions when compared to the best of two single-stream recognizers. In the presented method, where coupled HMMs allow for loosely coupled streams with variable lags, these improvements can halve the error rate compared to the better single stream recognizer, especially at low SNRs.

For additional gains in performance, missing feature recognition has been applied successfully. This approach is combined with a fairly simple and easily computed binary uncertainty estimate, which results in a system significantly less computationally demanding than e.g. uncertainty decoding.

Already for the considered binary uncertainties, both audio-only and video-only recognition gain in performance, and audio-visual recognition is also notably improved by this approach.

As an additional advantage, the performance becomes less dependent on stream weighting, much reducing the need for robust SNR estimation and stream weight adaptation, since the system automatically focuses on the most reliable of both feature streams.

References

[1] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio-visual speech recognition," Johns Hopkins University, CLSP, Tech. Rep. WS00AVSR, 2000. [Online]. Available: citeseer.ist.psu.edu/neti00audiovisual.html

[2] 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.

[3] 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, pp. 488–495.

[4] J. Gowdy, A. Subramanya, C. Bartels, and J. Bilmes, "Dbn based multi-stream models for audio-visual speech recognition," in *Proc. ICASSP*, vol. 1, May 2004, pp. I–993–6 vol.1.

[5] G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos, "Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 3, pp. 423–435, March 2009.

[6] H. Hermansky and N. Morgan, "Rasta processing of speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 4, pp. 578–589, Oct 1994.

[7] S. Young, N. Russell, and J. Thornton, "Token passing: a simple conceptual model for connected speech recognition systems," Cambridge University Engineering Department, Tech. Rep. CUED/FINFENG /TR.38, 1989.

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

[9] J. Barker, P. Green, and M. Cooke, "Linking auditory scene analysis and robust ASR by missing data techniques," in *Proceedings WISP 2001*, 2001.

[10] B. Raj and R. Stern, "Missing-feature approaches in speech recognition," *Signal Processing Magazine, IEEE*, vol. 22, no. 5, pp. 101–116, 2005.

[11] 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.

[12] R. F. Astudillo, D. Kolossa, and R. Orglmeister, "Uncertainty propagation for speech recognition using rasta features in highly nonstationary noisy environments," in *Proc. ITG*, 2008.

[13] 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.