Effects of Image Distortions on Audio-Visual Speech Recognition

Martin Heckmann ◇, *Frédéric Berthommier* •, *Christophe Savariaux* •, *Kristian Kroschel* ◇

◇ Institut für Nachrichtentechnik
Universität Karlsruhe
Kaiserstraße 12, 76128 Karlsruhe, Germany
{heckmann, kroschel}@int.uni-karlsruhe.de

• Institut de la Communication Parlée (ICP)
Institut National Polytechnique de Grenoble
46, Av. Félix Viallet, 38031 Grenoble, France
{bertho, savario}@icp.inpg.fr

Abstract

Audio-visual speech recognition leads to significant improvements compared to pure audio recognition especially when the audio signal is corrupted by noise. This has been reproduced by many researchers. Little research has been done on the behavior of audio-visual recognition with additional degradations of the video signal, however. In this article we investigate the consequences of different types of image degradations, namely white noise, a JPEG compression, and errors in the localization of the mouth region, on the audio-visual recognition process. The first question we address is how the noise in the video stream influences the recognition scores. Therefore we added noise to both, the audio and video signal at different SNR levels. The second question is how the adaptation of the fusion parameter, controlling the contribution of the audio and video stream to the recognition, is affected by the additional noise in the video stream. We compare the results we obtain when we adapt the fusion parameter to the noise in the audio and video stream to those we get when it is only adapted to the noise in the audio stream and hence a clean video stream is assumed. For the second type of tests we use an automatic adaptation of the fusion parameter based on the entropy of the a-posteriori probabilities from the audio stream.

1. Introduction

The importance of the lips movement in human speech perception, especially in noisy conditions, is well known [1]. This motivated the inclusion of the visual information in *Automatic Speech Recognition (ASR)* systems. In most of these systems the impact of additional noise in the audio stream is considered but the video stream is assumed to be of constant quality [2, 3, 4]. In [5] and [6] the effects of a degradation of the video stream on a video only recognition are investigated but the consequences it has on the audio-visual recognition are not taken into account. In this article we want to evaluate both, the influence of degradations in the video stream on a video only and on an audio-visual recognition process. Due to the fact that for clean audio the contribution of the audio stream dominates in audio-visual recognition and only small improvements compared to an audio only recognition can be observed we combine the video data with audio data corrupted by noise.

A key aspect of the fusion of audio and video data is the control of the fusion parameter which determines the contribution of either of the two streams to the recognition. Therefore it is an important question how the additional noise in the video stream affects the correct setting of the fusion parameter. To assess this influence, in a first experiment the fusion parameter was adapted to the noise in the audio and the video stream. In a second experiment we adapted the fusion parameter only to the noise in the audio stream and assumed that the video stream was undistorted. By comparing the results we are able to judge if it is necessary to adapt the fusion parameter to the degradations in both streams or if an adaptation only to the audio stream is sufficient.

2. The recognition system

2.1. System structure

The recognition tests aiming to assess the impact of degradations in the audio and video stream on the audio-visual speech recognition are carried out with an ANN/HMM hybrid model for continuous numbers recognition. For the implementation of the system the tool STRUT from TCTS lab Mons, Belgium, was used [7]. The identification of the phonemes is performed independently for the audio and the video path (compare Fig. 1) and thus follows a *Separate Identification (SI)* or multi-stream approach [8]. The ANNs are trained to produce estimates of

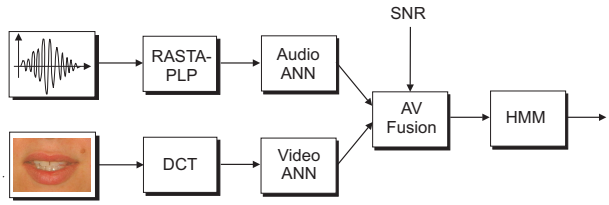


Figure 1: Separate Identification (SI) audio-visual speech recognition system

the a-posteriori probabilities $P(H_i|\mathbf{x}_A)$ and $P(H_i|\mathbf{x}_V)$ for the

occurrence of the phoneme H_i when the acoustic feature vector $\mathbf{x}_A$ and the visual feature vector $\mathbf{x}_V$ are observed, respectively.

Noise present in the audio or video stream affects the reliability of the estimated a-posteriori probabilities. This is taken into account by putting different weights on the audio and video stream during the fusion process. The fusion process follows a Geometric Weighting [9]:

$$\hat{P}_{GW}(H_i|\mathbf{x}_A,\mathbf{x}_V) = \frac{\hat{P}^{\alpha}(H_i|\mathbf{x}_A)\hat{P}^{\beta}(H_i|\mathbf{x}_V)}{\hat{P}^{\alpha+\beta-1}(H_i)} \cdot \varepsilon(\alpha,\beta) \quad (1)$$

The weighting parameters α and β both depend on a third parameter c according to:

$$\alpha = \begin{cases} 0 & c < -1 \\ 1+c & -1 \le c \le 0 \\ 1 & c > 0 \end{cases}, \quad \beta = \begin{cases} 1 & c < 0 \\ 1-c & 0 \le c \le 1 \\ 0 & c > 1 \end{cases} \quad (2)$$

The contribution of the audio and video data is controlled by the parameter c. When $c \simeq -1$, only the video signal contributes to the recognition, whereas for $c \simeq 1$ the recognition relies completely on the audio signal. $\epsilon(\alpha,\beta)$ is a normalization parameter independent on the actual phoneme determined by evaluating the condition $\sum_{i=1}^{N} \hat{P}_{GW}(H_i|\mathbf{x}_A,\mathbf{x}_V) = 1$.

2.2. Adaptive fusion

For a realistic scenario, the setting of the weights has to be performed automatically depending on the noise level in the input streams. A prerequisite to this is the estimation of the reliability of these streams during the fusion. The distributions of the a-posteriori probabilities at the output of each of the two ANNs carries information on the reliability of the corresponding input streams. If one distinct phoneme class shows a very high probability and all other classes have a low probability, this signifies a reliable input. Whereas, when all classes have almost equal probability, the input is very unreliable. This information is captured in the entropy of the estimated a-posteriori probabilities $\hat{P}(H_{i,k}|\mathbf{x}_{A,k})$ for the occurrence of the phoneme H_i, given the acoustic feature vector $\mathbf{x}_{A,k}$ at time frame k [10, 11, 2]. The entropy can be calculated for both, the audio and the video stream, but for now we will only take the audio stream into account. The entropy of the a-posteriori probabilities in frame k is

$$H(k) = -\sum_{i=1}^{N} \hat{P}(H_{i,k}|\mathbf{x}_{A,k}) \log_2 \hat{P}(H_{i,k}|\mathbf{x}_{A,k}) \quad (3)$$

where N is the number of phonemes. In order to reduce the effects of large variations of the entropy value from one frame to the other a smoothing with a first order recursive filter was introduced

$$\tilde{H}(k+1) = (1-a) \cdot \tilde{H}(k) + a \cdot H(k)\,. \quad (4)$$

For a the value 0.0025 was chosen corresponding to a time constant of 0.6 Hz.

We want to control the fusion process based on the entropy. Therefore a mapping between the value of the entropy and the fusion parameter c has to be established. In order not to make the reliability measure depend only on one noise type we used five different noise types to establish the mapping. The noise types were white noise, noise recorded in a car at 120 km/h and, from the NOISEX database, babble noise and two types of factory noise [12]. These five different types of noise were mixed to the audio signal at 12 SNR levels each ranging from -15 dB to clean speech. Next the best mapping, in a minimum error sense, between the entropy value for each of these 60 noise scenarios and the corresponding setting of the fusion parameter c with minimal *Word Error Rate (WER)* was determined [9]. The underlying settings of the parameter c with minimum WER were determined manually. To approximate the mapping, a second order polynomial function was used. The free parameters of this polynomial function were determined with a LMS algorithm. Based on this mapping, the fusion process can be controlled adaptively in a wide range of noise scenarios. Hence for a given value of the entropy of the a-posteriori probabilities the setting of the fusion parameter close to the optimum value can be determined.

2.3. The audio-visual database

To train the ANNs and to perform the recognition tests we used a single-speaker audio-visual database recorded at the Institut de la Communication Parlée (ICP) in Grenoble, France. This database is a repetition of a set of utterances selected from the Numbers95 corpus [13]. The database comprises 1543 utterances spoken by a native English-speaking female subject. Each of these utterances consists of several continuously uttered numbers yielding to a total of 4651 words. The database was divided into a set of 851 utterances for training and a set of 692 utterances for testing. During the recordings of the database a lamp positioned in front of the speaker ensured constant illumination conditions and high contrast images. The mouth region was tracked by means of markers positioned in the speakers face. To further facilitate the tracking of the mouth region, the movements of the speakers head were restricted during the recordings by a helmet. The recordings were made on BETACAM video with 50 half-frames of size 768×288 pixels per second. Instead of combining two half-frames to a full-frame we preferred to do without the additional spatial resolution and keep a higher temporal resolution. Full-frames were generated via a linear interpolation of the missing lines from the neighboring lines in each half-frame. After localization of the mouth region based on the markers positioned in the speakers face, the corresponding region was extracted and the images were downsampled by a factor 4. This yields images of 78×64 pixels at 50 frames per second of the *Region Of Interest (ROI)*, which are stored in the RGB format. Tracking of the ROI was performed via a correlation. The localization error in the final image is low (approx. 1 pixel).

2.4. Audio and video features

Video features are generated using the *Discrete Cosine Transform (DCT)*. The DCT has shown superior performance compared to other pixel based and geometric lip features [5]. To reduce the number of coefficients we have selected the 20 coefficients with the highest energy. This includes also the first coefficient representing the mean energy of the image. The selection of the coefficients is based on an evaluation of the training set.

The audio feature extraction is performed with RASTA-PLP using 13 cepstral coefficients and the log energy.

In order to take the context of a frame into account a time window of 13 frames set up by the current frame and the 6 preceding and succeeding frames was presented to the ANN. Each frame is 12 ms long and consecutive frames have a 50% overlap. Additionally to the pure DCT and RASTA-PLP coefficients

also their first and second order derivatives were used.

3. Distortions in the video stream

In general noise in the audio signal plays a more important role than video degradations. Nevertheless the video signal can also be corrupted and as a consequence impair the recognition scores.

3.1. Sources of degradations

Possible sources of degradations of the video signal are additive noise in the capturing or the transmission device, a mismatch of illumination conditions between the training conditions and the application of the system, lossy compression of the images on the transmission from the capturing device to the recognition system, and the incorrect localization of the ROI in the images (compare Fig. 2).

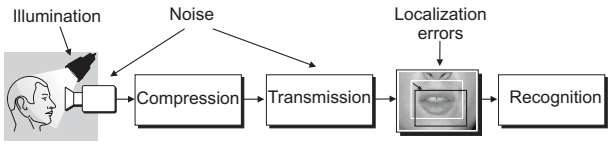
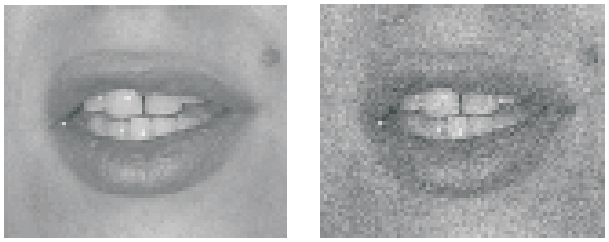


Figure 2: Possible sources of degradations to the video signal.

A change of the illumination conditions can not easily be simulated. For this purpose either the same sequences have to be recorded with different illumination conditions or 3-D recordings of the scene are necessary. Based on 3-D recordings the illumination can also be changed afterwards. Both these methods were too costly for the simulations performed here. Hence we limit our investigations in this article to the effects of additive white noise, a compression of the images with a JPEG algorithm and a translation of the mouth region in respect to its normal position.

3.1.1. White noise

In Fig. 3 an image from the database is shown a) before noise was added and b) with additional white noise at an SNR level of 0dB. To each image a different realization of a simulated noise process was added. Effects of the noise on the tracking of the mouth region were not investigated.



a) b)

Figure 3: In a) an image taken from the audio-visual database is shown. The same image is shown in b) when white noise at an SNR of 0 dB was added.

3.1.2. JPEG compression

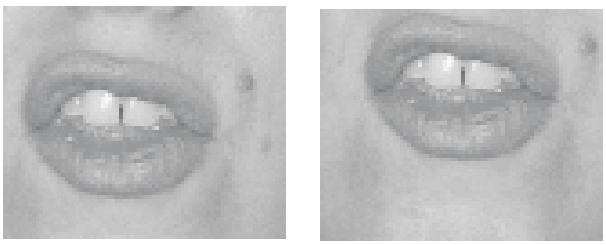
The most widespread used algorithms for moving image compression are the MPEG algorithms [14]. They are based on the still image compression standard JPEG [15]. For this reason the artifacts generated by both algorithms are very similar in nature. As the JPEG algorithm is much easier to implement we used this algorithm only to simulate compression artifacts in the video signal. Figure 4 shows the same image as in Fig. 3.a) after JPEG compression with a quality factor of 40 (a) and 10 (b). At a quality factor of 40 the degradations are still small but at a quality factor of 10 the impairments are clearly visible.

a) b)

Figure 4: The same image as in Fig. 3 a) is shown here after JPEG compression with a quality factor of 40 a) and 10 b).

3.1.3. Localization errors

The DCT used for the extraction of the video features is not shift-invariant. Hence a change of the position of the ROI in the image affects the values of the extracted features. Due to the fact that the database used for the experiments consists only of images of the mouth region, such localization errors of the ROI are difficult to simulate. For this purpose it was necessary to perform the extraction of the mouth region from the original analog video images with a different position of the ROI. This process is very time consuming and hence was only performed for one deviating position and only on a subset of the database. This subset contains 73 sequences with a total of 279 numbers. An image from this subset in its original position and after a shift of the ROI of 6 pixels to the left and 11 pixels to the bottom is shown in Fig. 5. This shift was kept constant for all images of the sequence. In the resulting image the upper bound of the upper lip is close to the border of the image but is in all images completely inside the image.



a) b)

Figure 5: An image from the reduced subset with the ROI in its original position (a) and after a shift (b).

3.2. Recognition scores

For each of the presented image distortions we performed video only and audio visual recognition tests. The tests were carried out in two steps: First we investigated the impact of the image degradations on the recognition itself. Therefore we performed video only and audio-visual tests for which we adapted the fusion parameter manually to give the best possible results in each noise scenario. Second we determined the loss of performance of the presented adaptive weighting algorithm due to the additional image distortions. This algorithm to control the weights on the audio and video stream assumes that the video stream is of constant quality and only adapts to noise in the audio stream. With our tests we wanted to answer the question if the control of the weighting based on the audio stream is sufficient in a realistic scenario where degradations occur in the audio and video stream or if it is rather necessary to develop a weighting algorithm which takes both streams into account.

For the audio-visual tests we added babble noise at SNR levels ranging from -3 to 12 dB to the audio stream. The recognition system was in all cases trained on clean audio and video.

3.2.1. White noise

In Fig. 6 the recognition results in *Word Error Rates (WERs)* are displayed when white noise was added to the video stream. In the very left column the results for the pure video recognition are plotted. As can be seen from the plot, small amounts of additive noise have almost no effect on the recognition performance, whereas when the image is severely degraded performance decreases significantly.

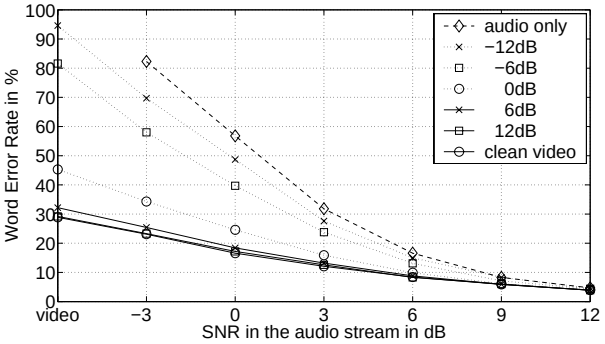


Figure 6: Fusion results when white noise was added to the video stream and babble noise to the audio stream. In the most left column the results for a video only recognition are displayed.

Figure 6 also shows the audio-visual results when the fusion parameter c was adapted manually so as to give minimum WERs. The numbers on the x-axis correspond to the SNR levels in the audio stream and the different curves represent the noise level in the video stream. Similar to the video only results almost no impairments of the recognition for SNR levels above 0 dB can be seen. Lower SNR levels in the video stream result in severe degradations of the recognition scores though. With an increase of the SNR in the audio stream the effects of the noise in the video stream on the audio-visual recognition decrease. This is due to the fact that in these situations the audio stream dominates the recognition. The results also show that even for very high distortions in the video stream the joint audio-visual recognition is still better than the audio only recognition.

In Fig. 7 the results of the second step of the recognition tests are displayed. Here the difference between the best possible recognition scores when the fusion parameter was set manually and those resulting from the adaptive evaluation of the fusion parameter based on the entropy of the a-posteriori probabilities is visualized. As can be expected from the previous results, the effects of disregarding the additional noise in the video stream during the adaptive setting of the fusion parameter are very small for video SNR levels below 0 dB. However for lower video SNR levels the loss of performance compared to the best possible values gets significant. Only for these values an additional evaluation of the quality of the video stream for the adaptive setting of the weights is necessary. The results of the adaptive audio-visual recognition at very low SNR in the video stream are even inferior to the audio only results. Such high noise levels in the video stream are quite unrealistic though.

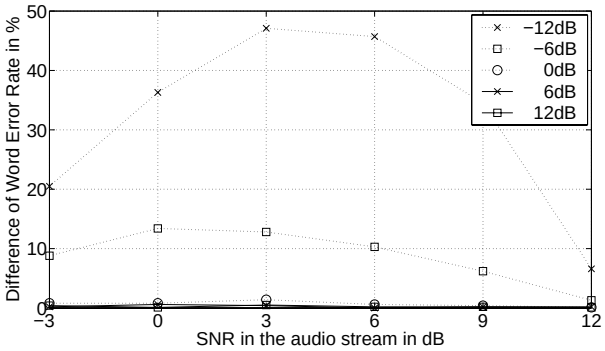


Figure 7: Difference of WERs if the fusion parameter was either adapted to both, the noise in the audio and video stream, or only to the audio stream. White noise was added to the video stream.

When looking at Fig. 7 it can be seen that the difference between adaptive and optimal setting of the weights at an SNR of -12 dB in the video stream initially increases with increasing SNR in the audio stream and then decreases again. This is at first sight a strange behavior. It originates from the changing influence of the video stream on the recognition with the changing noise in the audio stream. For very low SNR in the audio stream the influence of the video stream is high due to the adaptive weighting. The difference of the recognition results for both streams in this case is only small though. They both lead to WERs of about 90%. Hence also the degradation resulting form the adaptive weighting can only be small. With increasing SNR in the audio stream the recognition scores in the audio stream get better, but the adaptively determined influence of the video stream is still high. Due to the increasing difference between the recognition results of the audio and video stream the difference between optimal and adaptive weighting increases, too. When the SNR in the audio stream increases further, the influence of the video stream gets smaller and consequently the results for optimal and adaptive weighting approach each other again.

3.2.2. Compressed images

The tests for the JPEG compressed images were performed in a similar way as for the additive white noise. From Fig. 8 it can be seen that the effects of the compression are also rather small for low compression ratios. Correspondingly, for highly

compressed images there is still a significant gain at the audio-visual recognition compared to the audio alone recognition.

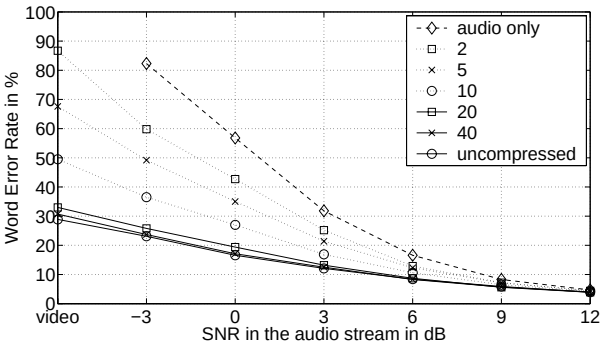


Figure 8: Fusion results when the images where compressed with a JPEG algorithm at a variable quality factor (from 40 to 2). In the most left column the results for a video only recognition are displayed.

In Fig. 9 the difference in performance between the optimal manual setting and the adaptive setting of the fusion parameter based only on the noise in the audio stream are given. In line

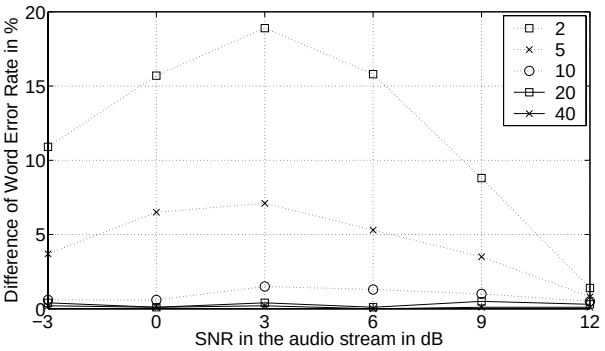


Figure 9: Difference of WERs if the fusion parameter was either adapted to both, the noise in the audio and video stream, or only to the audio stream. The images were compressed with the JPEG algorithm.

with the previous tests the loss of performance from neglecting the video noise is small and only plays a significant role when the image quality is very poor.

3.2.3. Localization errors

As can be seen from Fig. 10 a translation of the mouth region out of its standard position impairs the recognition much more than the distortions considered so far. The video only recognition results for the reduced test set, but with the mouth region in its correct position, gives a WER of 35.5%. This is significantly worse than the 28.9% obtained on the whole test set. The reason for this is most likely a tilt and a rotation of the head in the underlying images (compare Fig. 3.a and Fig. 5.a). As a consequence of the additionally introduced translation of the mouth region the WER increases dramatically to 92.5%. Thus the video stream conveys hardly any useful information. This can also be seen when looking at the audio-visual results in Fig. 10. The difference between the audio only and the audio-visual recognition

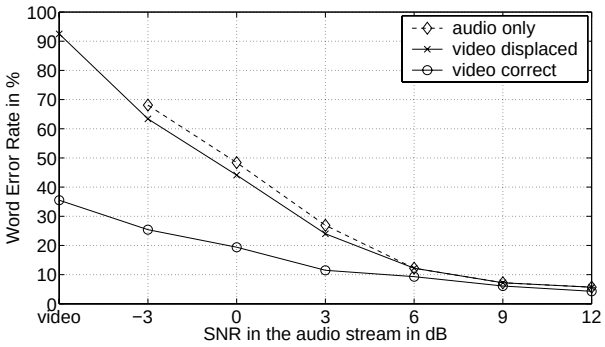


Figure 10: Fusion results on the reduced test set with the mouth region in its original position and after a translation. In the most left column the results for a video only recognition are displayed.

with the translated mouth region is very small. Due to the very poor results it is unnecessary to consider the results of the adaptive fusion.

4. Conclusion

We investigated the impact of distortions in both, the audio and the video stream, on the performance of an audio-visual recognition system. As image distortions we considered white noise, a compression of the images according to the JPEG standard at different compression ratios, and a translation of the mouth region from its original position in the image. For the audio-visual recognition tests we added babble noise at different SNR levels to the audio stream. In the first step of the recognition tests we performed video only and audio visual recognition tests for which the fusion parameter was adapted manually so as to give the best possible recognition results for each noise scenario. The results for additive white noise and degradations introduced by the compression are very similar in nature. In both cases only small effects can be observed for moderate degradations and even with severe degradations in the video stream the combined audio-visual recognition still showed significant improvements compared to the audio alone recognition. These results are in accordance with those obtained in [5]. Here the images were also degraded in a similar way but the recognition results were only based on the video stream. Despite these positive results the translation of the mouth region out of its original position severely deteriorates the recognition results. Deviations from the standard position lead to very high error rates in the video stream. Similar results are reported in [6]. Consequently the benefit from the additional use of the video stream for the recognition is negligible. To avoid these problems a precise localization of the mouth region during the feature extraction process is necessary. Algorithms capable of doing this are reported in [16], [17], and [18], for example. Small translations can be handled if they are present during the training phase of the recognition system. In this case the ANNs are able to adapt to these translations.

An important aspect of our investigations was to consider the impact of degradations in the video stream on the audio-visual recognition and especially the necessary changes of the setting of the fusion parameter. We elaborated a mapping between a measure for the noise level in the audio stream, namely the entropy of the a-posteriori probabilities, and the fusion pa-

rameter which controls the contribution of the audio and video stream during the recognition. This mapping enables an adaptive setting of the fusion parameter with optimal recognition results in a wide range of noise levels in the audio stream. In a second step of our recognition tests we investigated how the additional noise in the video stream, which is not taken into account for the evaluation of the entropy, affects the performance of the algorithm. The results showed that for small to medium degradations disregarding the noise in the video stream in the setting of the weights has only a small impact on the recognition performance. It follows from this that it is sufficient in an audio-visual recognition system to adaptively control the setting of the fusion parameter based on the audio stream. Only in situations where the image quality is extremely poor, an additional evaluation of the video stream could lead to significant improvements. However, these situations are not very realistic and also show a very low overall performance. In general the SNR in the video stream can be expected to be significantly better than 20 dB and for the quality factor in JPEG compression a common value is 70. A value of 40 already leads to clearly visible degradations.

5. Acknowledgments

We want to thank our subject Laurie Rebut for her effort and patience, Pandu Ranga Rao Devarakota for carrying out many simulations, and the reviewers of AVSP'03 for their helpful comments. This work was partly funded by the EC program SPHEAR and is a part of the project RESPITE.

6. References

[1] W. H. Sumby and I. Pollack. Visual contribution to speech intelligibility. *Journ. of the Acoustical Society of America*, 26:221–215, 1954.

[2] G. Potamianos and C. Neti. Stream confidence estimation for audio-visual speech recognition. In *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, pages 746–749, Bejing, China, 2000.

[3] S. Dupont and J. Luettin. Audio-visual speech modeling for continuous speech recognition. *IEEE Trans. on Multimedia*, 2:141–151, 2000.

[4] P. Teissier, J. Robert-Ribes, J.-L. Schwartz, and A. Guérin-Dugué. Comparing models for audiovisual fusion in a noisy-vowel recognition task. *IEEE Trans. Speech and Audio Processing*, 7:629–642, 1999.

[5] G. Potamianos, H. P. Graf, and E. Cosatto. An image transform approach for hmm based automatic lipreading. In *Proc. of the Int. Conf. on Image Proc. (ICIP)*, Chicago, 1998.

[6] J. Y. Kim, J. Lee, and K. Shirai. A study on various factors concerned with lipreading performance at dynamic environment. In *International Conference on Speech Processing (ICSP)*, pages 923–927, Daejon, Korea, 2001.

[7] University of Mons, Mons. *Step by Step Guide to using the Speech Training and Recognition Unified Tool (STRUT)*, May 1997.

[8] A. Rogozan and P. Deléglise. Adaptive fusion of acoustic and visual sources for automatic spech recognition. *Speech Communication*, 26:149–161, 1998.

[9] M. Heckmann, F. Berthommier, and K. Kroschel. Noise adaptive stream weighting in audio-visual speech recognition. *Journal on Applied Signal Proc.: Special Issue on Audio-Visual Proc.*, 2002:1260–1273, 2002.

[10] C. Bregler, H. Hild, S. Manke, and A. Waibel. Improving conected letter recognition by lipreading. In *Proc. IEEE Int. Conf. Acoustics Speech and Signal Processing (ICASSP)*, 1993.

[11] A. Adjoudani and C. Benoit. On the integration of auditory and visual parameters in an hmm-based asr. In D.G. Stork and M.E. Hennecke, editors, *Speachreading by Man and Machine: Models, Systems and Applications*, NATO ASI Series, pages 461–472, Berlin, 1996. Springer.

[12] A.P. Varga, H.J.M. Steeneken, M. Tomlinson, and D. Jones. The noisex-92 study on the effect of additive noise on automatic speech recognition. Technical report, Speech Research Unit, Defence Research Agency, Malvern, U.K., 1992.

[13] R. A. Cole, T. Noel, L. Lander, and T. Durham. New telephone speech corpora at cslu. In *Proc. European Conf. on Speech Communication and Technology (EUROSPEECH)*, pages 821–824, 1995.

[14] D. Le Gall. MPEG: A video compression standard for multimedia applications. *Commun. ACM*, 34(4):46–58, 1991.

[15] W. B. Pennebaker and J. L. Mitchell. *JPEG Still Image Data Compression Standard*. Van Nostrand Reinhold, New York, NY, 1993.

[16] X. Zhang, C. C. Broun, R. M. Mersereau, and M. A. Clements. Automatic speechraeding with applications to human-computer interfaces. *Journal on Applied Signal Proc.: Special Issue on Audio-Visual Proc.*, 2002:1228–1247, 2002.

[17] U. Meier, R. Stiefelhagen, J. Yang, and A. Waibel. Towards unrestricted lip reading. *Int. Journal on Pattern Rec. and Art. Intelligence*, 14(5):571–585, 2000.

[18] Gerasimos Potamianos, A. Verma, C. Neti, Giridharan Iyengar, and Sankar Basu. A cascade image transform for speaker independent automatic speech reading. In *IEEE International Conference on Multimedia and Expo (II)*, pages 1097–1100, 2000.