Adaptive Audio-Visual Speech Recognition with Distorted Audio and Video Data

Martin Heckmann [1], Frédéric Berthommier [2], Christophe Savariaux [2], Kristian Kroschel [3]

[1] *Honda Research Institute Europe, 63073 Offenbach, Germany, Email martin.heckmann@honda-ri.de*

[2] *Institut de la Communication Parlée (ICP), 38031 Grenoble, France, Email: {bertho, savario}@icp.inpg.fr*

[3] *Institut für Nachrichtentechnik, Universität Karlsruhe, 76128 Karlsruhe, Germany, Email: kroschel@int.uni-karlsruhe.de*

Introduction

The importance of the lips movement in human speech perception, especially in noisy conditions, is well known and motivated the inclusion of the visual information in *Automatic Speech Recognition (ASR)* systems [1, 2, 3]. In this article we want to evaluate 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.

The recognition system

System structure

The recognition tests are carried out with an *Artifical Neural Network/Hidden Markov Model (ANN/HMM)* hybrid model for continuous numbers recognition. The identification of the phonemes is performed independently for the audio and the video path and thus follows a *Separate Identification (SI)* or multi-stream approach [4]. The ANNs are trained to produce estimates of the a-posteriori probabilities for the occurrence of the phonemes when the acoustic and visual feature vectors are observed.

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. To find the adequate setting of the parameter controlling the fusion process in changing noise scenarios an adaptive algorithm was developed [5]. The basis of this algorithm is the evaluation of the entropy of the a-posteriori probabilities at the output of the audio ANN. Hence the adaptation algorithm is only based on information from the audio stream and does not take additional degradations in the video stream into account.

The audio-visual database

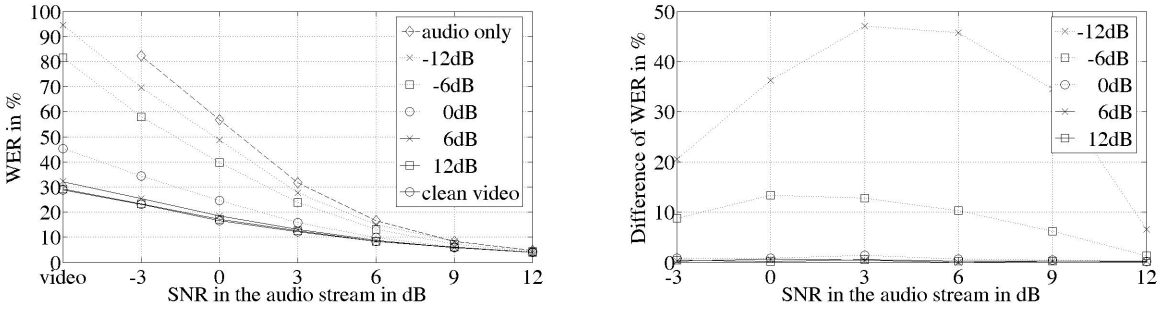
To train the ANNs and to perform the recognition tests we used a single-speaker audio-visual database. The database comprises 1543 utterances spoken by a native English-speaking female subject (851 used for training and 692 for testing). Each of these utterances consists of several continuously uttered numbers. 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. The recordings were made with 50 half-frames 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, a *Region Of Interest (ROI)* of 78×64 pixels was extracted.

Audio and video features

Video features are generated using the *Discrete Cosine Transform (DCT)*. To reduce the number of coefficients we have selected the 20 coefficients with the highest energy. 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.

Distortions in the video stream

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. Here we only present results for additive white Gaussian noise.



a) b)

Figure 1: In a) the results when white noise was added to the video stream and babble noise to the audio stream can be seen. In the most left column the results for a video only recognition are displayed. In b) the 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, is given.

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.

Recognition scores

The tests were carried out in two steps: First we investigated the impact of the image degradation 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. Secondly we determined the loss of performance when using the adaptive weighting algorithm due to the additional image distortions.

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.

In Fig. 1 the recognition results in *Word Error Rates (WERs)* are displayed. 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 1.a) also shows the audio-visual results when the fusion parameter was adapted manually so as to give minimum WERs. On the x-axis the SNR levels in the audio stream are indicated and the different curves represent the noise level in the video stream. 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. 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. 1.b) 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. The effects of neglecting 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. Such high noise levels in the video stream are quite unrealistic though.

Conclusion

Our results show that distortions in the video stream have only a small impact on the overall system performance. In particular we could show that it is sufficient to adjust the weighting of the audio and video stream during recognition only to the noise in the audio stream. The negligence of the additional video distortions has only a very small impact on the recognition results when SNR levels lie in a realistic range (in general the SNR in the video stream can be expected to be significantly better than 20 dB). These results also hold when using a lossy compression of the images [6]. Whereas a shift of the mouth inside the image leads to severe performance degradations making a precise tracking of the mouth region indispensable when using DCT features.

References

[1] 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.

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

[3] 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.

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

[5] 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.

[6] M. Heckmann, F. Berthommier, Cristophe Savariaux, and K. Kroschel. Effects of image distortions on audio-visual speech recognition. In *6th Audio-Visual Speech Processing Conference (AVSP)*.