EURASIP Journal on Applied Signal Processing 2002:11, 1260–1273

Noise Adaptive Stream Weighting in Audio-Visual Speech Recognition

Martin Heckmann
Institut für Nachrichtentechnik, Universität Karlsruhe, Kaiserstraße 12, 76128 Karlsruhe, Germany
Email: heckmann@int.uni-karlsruhe.de

Frédéric Berthommier
Institut de la Communication Parlée (ICP), Institut National Polytechnique de Grenoble, 46 Av. Félix Viallet, 38031 Grenoble, France
Email: bertho@icp.inpg.fr

Kristian Kroschel
Institut für Nachrichtentechnik, Universität Karlsruhe, Kaiserstraße 12, 76128 Karlsruhe, Germany
Email: kroschel@int.uni-karlsruhe.de

Received 31 November 2001 and in revised form 26 July 2002

It has been shown that integration of acoustic and visual information especially in noisy conditions yields improved speech recognition results. This raises the question of how to weight the two modalities in different noise conditions. Throughout this paper we develop a weighting process adaptive to various background noise situations. In the presented recognition system, audio and video data are combined following a Separate Integration (SI) architecture. A hybrid Artificial Neural Network/Hidden Markov Model (ANN/HMM) system is used for the experiments. The neural networks were in all cases trained on clean data. Firstly, we evaluate the performance of different weighting schemes in a manually controlled recognition task with different types of noise. Next, we compare different criteria to estimate the reliability of the audio stream. Based on this, a mapping between the measurements and the free parameter of the fusion process is derived and its applicability is demonstrated. Finally, the possibilities and limitations of adaptive weighting are compared and discussed.

Keywords and phrases: audio-visual speech recognition, adaptive weighting, robust recognition, multistream recognition, ANN/HMM.

1. INTRODUCTION

The limited performance of *Automatic Speech Recognition (ASR)* systems in the presence of background noise still restricts their usability in many scenarios. Different attempts have been made to increase the robustness of ASR systems but all fall short in comparison to human performance.

It is well known that the movement of the lips plays an important role in speech perception [1, 2]. The contribution of the lips is especially high in noisy speech [3, 4]. This is due to the fact that visual speech mainly conveys information about the place of articulation, which is most easily confused in the audio modality when noise is present [5]. Motivated by these findings many researchers have tried to integrate the information transmitted by lip movement into ASR systems (see [6, 7, 8, 9, 10, 11, 12] for a review). The first systems, already, showed noticeable improvements of the recognition scores in noise when the audio and video signals are jointly evaluated. Since then significant progress was made, and currently a recognition system using both, audio and video data, can outperform humans having only access to the audio signal at low *Signal to Noise Ratio (SNR)* [13]. Despite this high performance of audio-visual ASR systems, there is still a long way to go before these systems will have performance comparable to humans in an identical task.

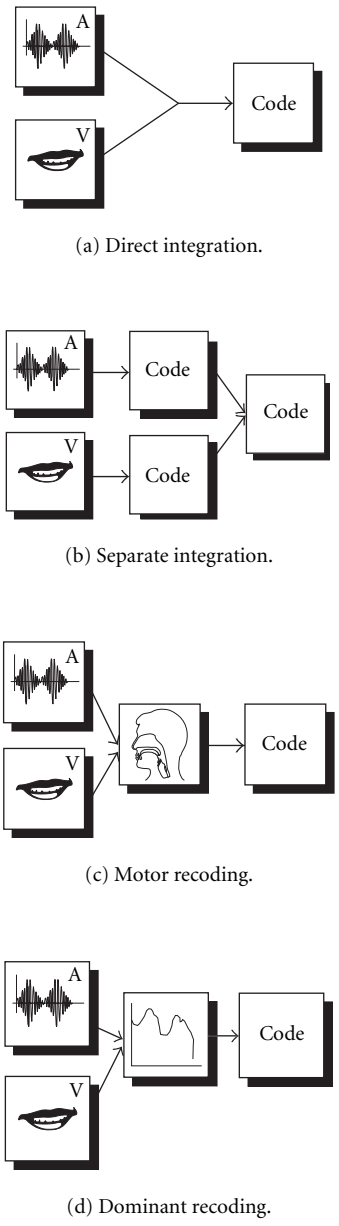
Throughout this paper, we mainly want to focus on the adaptive fusion of audio and video data under different noise conditions. We start with a quick look at different possible fusion architectures and point out why we have chosen a *Separate Integration* architecture, where fusion takes place on a decision level. Next, we present four different fusion schemes of audio and video decisions. A comparison of these fusion schemes in a wide range of noise conditions allows to identify the best scheme. In order to be adaptive to changing noise conditions, there is need for a criterion to evaluate the reliability of the audio channel. We present three different

reliability criteria and compare them in different noise conditions. We conclude this paper with a discussion of the results of our comparisons. Throughout this discussion, special attention is paid to the question of whether adaptive weights on the audio and video stream are necessary, or if it is sufficient to simply use one fixed weight for all situations.

2. FUSION OF AUDIO AND VIDEO DATA

When looking at the fusion of audio and video data for audio-visual speech recognition, the first question to be addressed is where the fusion of the data takes place. Several different architectures for the fusion process have been proposed [5, 14]. The first is integration on the feature level. In this case, audio and video features are directly combined to a larger feature vector, which is then used to identify the corresponding phoneme. This is also referred to as *Direct Integration (DI)* (see also Figure 1). In contrast to this, fusion can also take place after independent identification of each stream. Hence the fusion is rather a fusion of identification results. This is called *Separate Integration (SI)*. Between these two extremes lies the so-called *Motor Recoding (MR)* in which the input features are first transformed into a common representation, and the classification then is based upon the combined features in this representation. The articulatory gesture parameters are chosen as common representation, to which both audio and video features are mapped. A problematic point when using Motor Recoding is the choice of the representation of the articulatory gestures. In the fourth fusion architecture one stream is dominant. In this case, the decision is based on the dominant stream, and the second stream is only used to rescore the identification results of the dominant stream. This is called *Dominant Recoding (DR)*. Due to the fact that it conveys much more information than the video stream, naturally the audio stream is chosen as the dominant stream.

When comparing the different fusion architectures, Separate Integration exhibits some characteristics that make it the best choice for our task. An important property is that the fusion of the two input streams can be controlled by weighting the streams. The code elements in Figure 1 are the phonemes H_i to which we can assign a posteriori probabilities $P(H_i|\mathbf{x}_A, \mathbf{x}_V)$ for their occurrence given the acoustic feature vector $\mathbf{x}_A$ and the video feature vector $\mathbf{x}_V$ (see Figure 2). These a posteriori probabilities, or to be more precise their estimates $\hat{P}$, are generated by an *Artificial Neural Network (ANN)* [15] in each time frame. Therefore the SI, in combination with an ANN, allows an adaptive weighting of the input streams depending on their reliability. Adaptation of the weights can be done once per scenario as well as for each single frame. Furthermore comparisons of SI with other architectures showed superior performance of SI [16, 17, 18].[1] For these reasons, we decided to use an SI architecture for our recognition experiments. Once we have chosen the SI



(a) Direct integration.

(b) Separate integration.

(c) Motor recoding.

(d) Dominant recoding.

FIGURE 1: Four different fusion architectures for audio-visual recognition.

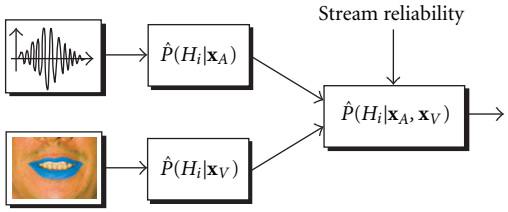


FIGURE 2: Weighting of audio and video a posteriori probabilities in a Separate Integration architecture to take into account the changing reliability of the input streams.

[1] Regarding the comparison of DI and SI, these results were confirmed by our own experiments but not reported here.

architecture, the next question to tackle is how the fusion of the identification results takes place.

The quality of the estimate of the a posteriori probabilities is related to the match of the training and test conditions. As training was in all cases performed on clean data, the reliability of these estimates, particularly in the case of the audio path, strongly depends on the noise present in the test condition. In order to cope with the changing reliability, a weighting of the audio and video probabilities is desirable. Even though the quality of the video stream was kept constant in all following tests, an adaptive weighting of the video stream depending on the quality of the audio stream does improve the performance.

2.1. Unweighted Bayesian Product

The simplest way to combine audio and video data is to follow Bayes' rule and multiply the audio and video a posteriori probabilities to derive the combined probabilities. This approach is valid in a probabilistic sense if the audio and video data are independent. Perceptive studies showed that in human speech perception audio and video data are treated as class conditional independent [19, 20]. Under this hypothesis,

$$P(\mathbf{x}_A, \mathbf{x}_V|H_i) = P(\mathbf{x}_A|H_i)P(\mathbf{x}_V|H_i). \tag{1}$$

When applying Bayes' rule, we can write the desired a posteriori probability of the phoneme H_i as

$$P(H_i|\mathbf{x}_A, \mathbf{x}_V) = \frac{P(H_i|\mathbf{x}_A)P(H_i|\mathbf{x}_V)}{P(H_i)} \cdot \frac{P(\mathbf{x}_A)P(\mathbf{x}_V)}{P(\mathbf{x}_A, \mathbf{x}_V)}. \tag{2}$$

Replacement of the probabilities P by estimates $\hat{P}$ leads to the representation of the, as we want to call it, *Unweighted Bayesian Product (UBP)*

$$\hat{P}_{\mathrm{UBP}}(H_i|\mathbf{x}_A, \mathbf{x}_V) = \frac{\hat{P}(H_i|\mathbf{x}_A)\hat{P}(H_i|\mathbf{x}_V)}{\hat{P}(H_i)} \cdot \eta, \tag{3}$$

where the terms independent of the actual phoneme are replaced by the normalization factor

$$\eta = \frac{1}{\sum_{j=1}^{N} \hat{P}(H_j|\mathbf{x}_A)\hat{P}(H_j|\mathbf{x}_V)/\hat{P}(H_j)}, \tag{4}$$

with N being the number of phonemes. This fusion scheme is also the core of the *Fuzzy Logical Model of Perception (FLMP)* [21], which is used to model human perception.

2.2. Standard Weighted Product

In order to deal with varying reliability levels of the input streams, different authors introduced a weighted fusion, where different weights are applied to the audio and video channels. The weighting of the a posteriori probabilities proposed in [17, 18] follows (we want to refer to this as *Standard Weighted Product*)

$$\hat{P}_{\mathrm{SWP}_\lambda}(H_i|\mathbf{x}_A, \mathbf{x}_V) = \frac{\hat{P}^{\lambda}(H_i|\mathbf{x}_A)\hat{P}^{(1-\lambda)}(H_i|\mathbf{x}_V)}{\sum_{j=1}^{N} \hat{P}^{\lambda}(H_j|\mathbf{x}_A)\hat{P}^{(1-\lambda)}(H_j|\mathbf{x}_V)}. \tag{5}$$

The assumption of conditional independence is approached for equal a-priori probabilities of the phonemes or words, respectively, depending on the place of fusion. It is not actually fulfilled since equal weights on both streams correspond to weights of 0.5 instead of 1.

In addition to the intermediate setting, when the audio and video stream contribute equally to the recognition, two more distinct settings of the weights exist. When the SNR is very low, the estimation in the audio path completely fails. Therefore, the final a posteriori probability should only depend on the video features, which is achieved

$$\hat{P}_{\mathrm{SWP}_\lambda}(H_i|\mathbf{x}_A, \mathbf{x}_V) = \hat{P}(H_i|\mathbf{x}_V), \tag{6}$$

with $\lambda = 0$.

Similarly, for very high SNR, the estimation in the audio path is in general much better than the one in the video path and consequently

$$\hat{P}_{\mathrm{SWP}_\lambda}(H_i|\mathbf{x}_A, \mathbf{x}_V) = \hat{P}(H_i|\mathbf{x}_A), \tag{7}$$

with $\lambda = 1$.

The most common recognition systems are based on *Gaussian Mixture Hidden Markov Models (GM/HMM)*. These produce likelihoods instead of a posteriori probabilities. Weighting of these likelihoods corresponds to a weighting of (1) [9, 10, 16]. This approximates the assumption of conditional independence, independent of the a-priori probabilities.

Equal weights of 0.5 instead of 1 entails that not the product of the probabilities but the square root of the product is evaluated when both the audio and the video stream have the same weight. To resolve this problem, we modify the parameterization of the Standard Weighted Product. We introduce the parameters α and β, which depend both on a third parameter c according to

$$\alpha = \begin{cases} 0, & c \le -1, \\ 1 + c, & -1 < c < 0, \\ 1, & c \ge 0, \end{cases} \qquad \beta = \begin{cases} 1, & c \le 0, \\ 1 - c, & 0 < c < 1, \\ 0, & c \ge 1, \end{cases} \tag{8}$$

yielding

$$\hat{P}_{\mathrm{SWP}_{\alpha\beta}}(H_i|\mathbf{x}_A, \mathbf{x}_V) = \frac{\hat{P}^{\alpha}(H_i|\mathbf{x}_A)\hat{P}^{\beta}(H_i|\mathbf{x}_V)}{\sum_{j=1}^{N} \hat{P}^{\alpha}(H_j|\mathbf{x}_A)\hat{P}^{\beta}(H_j|\mathbf{x}_V)}. \tag{9}$$

Similarly to λ in the previous parameterization, the parame-

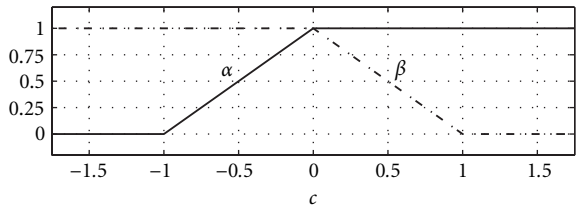


Figure 3: Dependence of the parameters α and β on the fusion parameter c.

ter c varies with the SNR, and it determines the contribution of the audio and video streams to the final probability.

When $c = 0$ the a posteriori probabilities from the audio and video path, both have the same weight as $\alpha = 1$ and $\beta = 1$ (see also Figure 3). For $c \leq -1$ (at very low SNR), $\alpha = 0, \beta = 1$ and for $c \geq 1$ (when only the audio signal carries information), $\alpha = 1, \beta = 0$. Hence this takes the situations into account, where we only want to rely on one of the two streams. In contrast to the original parameterization of the Standard Weighted Product to which we want to refer as SWP_λ, this implementation will be referred to as $\text{SWP}_{\alpha\beta}$.

2.3. Geometric Weighting

A concept integrating class conditional independence of audio and video data expressed in (1) and the idea of noise dependent stream weighting expressed in (5) is the *Geometric Weighting* [22]

$$\hat{P}_{\text{GW}}(H_i|\mathbf{x}_A, \mathbf{x}_V) = \frac{\hat{P}^{\alpha}(H_i|\mathbf{x}_A)\hat{P}^{\beta}(H_i|\mathbf{x}_V)}{\hat{P}^{\alpha+\beta-1}(H_i)} \cdot \varepsilon(\alpha, \beta). \quad (10)$$

The normalization factor

$$\varepsilon(\alpha, \beta) = \frac{1}{\sum_{j=1}^{N} \left(\hat{P}^{\alpha}(H_j|\mathbf{x}_A)\hat{P}^{\beta}(H_j|\mathbf{x}_V)/\hat{P}^{\alpha+\beta-1}(H_j)\right)} \quad (11)$$

is determined by evaluating the condition

$$\sum_{j=1}^{N} \hat{P}_{\text{GW}}(H_i|\mathbf{x}_A, \mathbf{x}_V) = 1. \quad (12)$$

Factors only dependent on $\mathbf{x}_A$ and $\mathbf{x}_V$ are eliminated by the normalization. The result of the sum in (11) is independent of H_i and hence ε only depends on the fusion weights α and β.

For the Geometric Weighting we solely employed the parameterization with α and β as defined in (8). Consequently, for $c = 0$ the assumption of conditional independence as stated in (1) is fulfilled when equal weight is put on the audio and video stream. Similar to the description in the previous section for $c = -1$ the final probability only depends on the a posteriori probability of the video stream and for $c = 1$ it only depends on the audio stream (see also Figure 3).

2.4. Full Combination

Findings in human speech perception showed that the error rate for phoneme recognition using the full frequency range is approximately equal to the product of the error rates using only nonoverlapping frequency sub-bands [23, 24]. This is known as the so-called *Product of Errors (POE) Rule*. Motivated by this rule, multistream recognition systems were built, which decompose the speech signal in multiple subbands, perform an identification of the phoneme for each sub-band, and then combine the results [25]. In general, the performance gain of this approach was not very high in noise and was countered by a loss of performance on clean speech. The loss on clean speech is alleviated by the so-called *Full Combination (FC)* approach [26]. Here phoneme identification is performed for all combinations of sub-bands, including also the full frequency range, and the identification results are then combined linearly.

When applying this concept to audio-visual recognition we have to consider two input streams. Taking all combinations of the input streams plus the empty stream containing only the a-priori probabilities into account we have a total of four streams: the audio, the video, the combined audio-visual and the empty stream. Hence three ANNs have to be trained to generate the corresponding probabilities. The weighting of the streams is performed by a linear combination of the a-priori and a posteriori probabilities according to

$$\begin{aligned}\hat{P}_{\text{FC}}(H_i|\mathbf{x}_A, \mathbf{x}_V) = {} & a_1\hat{P}(H_i|\mathbf{x}_A, \mathbf{x}_V) + a_2\hat{P}(H_i|\mathbf{x}_A) \\ & + a_3\hat{P}(H_i|\mathbf{x}_V) + a_4\hat{P}(H_i).\end{aligned} \quad (13)$$

In order to reduce the number of neural networks to be trained on each independent stream (which grows exponentially with the number of streams), the so-called *Full Combination Approximation (FCA)* was introduced [26]. Here class conditional independence is assumed between the streams and hence the identification result for a combination of streams can be derived from the identification results of the individual streams (compare to (2)). Then the a posteriori probability of the combined audio-visual stream is evaluated according to

$$\begin{aligned}\hat{P}_{\text{FCA}}(H_i|\mathbf{x}_A, \mathbf{x}_V) = {} & a_1\frac{\hat{P}(H_i|\mathbf{x}_A)\hat{P}(H_i|\mathbf{x}_V)}{\hat{P}(H_i)} \cdot \eta + a_2\hat{P}(H_i|\mathbf{x}_A) \\ & + a_3\hat{P}(H_i|\mathbf{x}_V) + a_4\hat{P}(H_i),\end{aligned} \quad (14)$$

with η as defined in (4). The first term in (14) results from the postulation of class conditional independence, and the other terms ensure the same behavior as Geometric Weighting when only one of the streams is reliable. The a_k are the weights with which the individual streams contribute to the final probability. They are set to $a_1 = \alpha \cdot \beta$, $a_2 = \alpha(1 - \beta)$, $a_3 = (1 - \alpha)\beta$, and $a_4 = (1 - \alpha) \cdot (1 - \beta)$, with α and β as given in (8). When the estimation process for the different probabilities is not consistent, and hence the sum over all probabilities does not equal one, an independent normalization for each stream is necessary. At $c = 0$ the assumption of conditional independence is fulfilled. Similarly for $c = 1$ and $c = -1$ all the weight is assigned to the audio or video

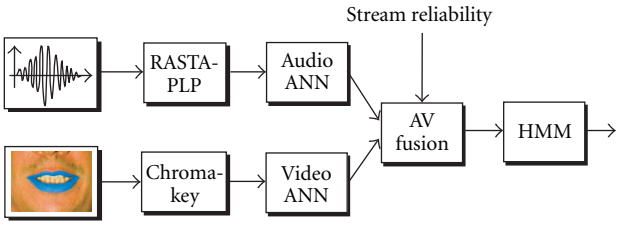


FIGURE 4: Implementation of the SI audio-visual speech recognition system.

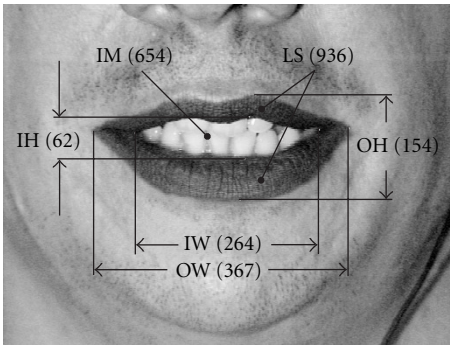


FIGURE 5: One exemplary image from the database with the extracted lip parameters visualized. The numerical values are given in pixels.

stream, respectively. In our implementation, the degrees of freedom of the FCA and the Geometric Fusion are limited to one. This might not be optimal but a multidimensional optimization with multiple degrees of freedom would be much more costly to perform.

3. THE RECOGNITION TASK

As a common task to evaluate the presented fusion schemes we have chosen the recognition of continuously uttered English numbers. This task comprises many of the problems of continuous speech recognition, whilst still being not too costly to implement. One of the distinct features of a continuous recognition task is the necessity to discriminate between speech segments and silence passages, which is especially problematic in noisy speech. Due to the very limited availability of audio-visual speech data, we had to record a new database to train our system.

3.1. The audio-visual database

For the recording of the database, selected utterances from NUMBERS95 [27] were chosen and repeated by a single native English-speaking male subject. The database contains 1712 sentences or 6432 words. It was subdivided into two subsets of similar size for training and final recognition. Synchronous recordings of the speech signal and video images of the head and mouth region at 50 frames per second were taken. Recordings were made on BETACAM video and standard audio tapes and A/D converted with 8 kHz off-line.

3.2. The recognition system

Our audio-visual speech recognition system is based on a hybrid *Artificial Neural Network/Hidden Markov Model (ANN/HMM)* structure. ANN/HMM hybrid systems represent an alternative concept for continuous speech recognition to pure HMM systems giving competitive recognition results [28]. As already mentioned in the previous section, our system follows an SI architecture (see Figure 4). The implementation of our system was carried out using the tool STRUT from TCTS lab Mons, Belgium [29].

The emphasis of our research lies on the *fusion* of the audio and video data during the recognition process which requires large amounts of data to obtain meaningful results. Therefore, following [16], we rely on geometric lip features and simplify the extraction of the features significantly by a *chroma key* process. The chroma key process requires coloring the speakers lips with blue lipstick. Due to the coloring, the lips can then be located easily, and their movement parameters can be extracted in real time. As lips parameters, the following were chosen:

(i) outer lip width (OW);
(ii) inner lip width (IW);
(iii) outer lip height (OH);
(iv) inner lip height (IH);
(v) lip surface area (LS);
(vi) inner mouth area surrounded by lips (IM).

In Figure 5 the results of the feature extraction are visualized. The detected lip boundaries and the corresponding numerical values in pixels are given. The extracted video parameters were linearly interpolated from the original 50 Hz to 8 kHz, in order to be synchronous with the audio data. Following the interpolation, each lip parameter was low-pass filtered to remove high frequency noise introduced by the parameter extraction and to further smooth the results of the interpolation. Audio feature extraction was performed using RASTA-PLP [30].

To take temporal information into account, several successive time frames of the audio and video feature vectors are presented simultaneously to the input of the corresponding ANNs. The concept of visemes was not used. Each acoustical articulation is assumed to have a synchronously generated corresponding visual articulation. Hence the recognition process is based on phonemes. Individual phonemes are modeled via left-to-right HMM models. The number of states of the HMMs used to represent the different phonemes was adapted to the mean length of the corresponding phoneme. Word models were generated by the concatenation of the corresponding phoneme models. Recognition is based on a dictionary with the phonetic transcription of 30 English numbers. Complete sentences containing a sequence of numbers were presented to the system during the recognition process. The sentences consist of free format numbers making a grammar model unnecessary.

Training of the ANNs was in all cases performed on clean data. During our recognition tests we added 5 different types

Table 1: Average of the relative error in percent for audio alone recognition and the fusion schemes *Standard Weighted Product (SWP)* parameterized with λ (SWP_λ) and α and β ($\text{SWP}_{\alpha\beta}$), *Geometric Weighting (GW)*, *Full Combination (FC)*, *Full Combination Approximation (FCA)*, and *Unweighted Bayesian Product (UBP)* over all noise types and SNR levels. Additionally the 95% confidence interval for the relative error is given.

Method	Audio	SWP_λ	$\text{SWP}_{\alpha\beta}$	UBP	FC	FCA	GW
Error e	137.6 ± 2.6	$0.0 \pm -$	-0.9 ± 2.4	9.2 ± 2.1	-2.2 ± 2.3	-28.9 ± 2.0	-30.2 ± 2.0

of environmental noise at 12 different SNR levels to the audio signal, resulting in 60 different test conditions. Adding noise to the recorded signal instead of adding it during the recordings does not take into account the changes in articulation speakers produced when background noise is presented [31] and therefore generates somehow nonrealistic scenarios. On the other hand it opens the possibility to test exactly the same utterances in different noise conditions and tremendously facilitates the recordings of the data. As additive noise we have chosen white noise, noise recorded in a car at 120 km/h and babble noise and two types of factory noise taken from the NOISEX database [32]. Noise was only added to the audio signal. We considered the video stream to be of constant quality and did not alter the video signal throughout the tests.

4. EVALUATION OF THE FUSION SCHEMES

The first step in the evaluation is to compare the fusion schemes under identical conditions using a manual setting of the optimal weights.

4.1. Manual weight adaptation

Throughout this first stage of evaluation, the fusion parameter c in the Standard Weighted Product with α and β parameterization ($\text{SWP}_{\alpha\beta}$), the FCA and the Geometric Fusion was adapted manually at each SNR level in order to get the best possible recognition score. During a test in a particular noise condition, the fusion parameter was held constant over all frames. Tests in that particular noise condition with different settings of the fusion parameter were repeated until the minimum *Word Error Rate (WER)* was reached. For the Standard Weighted Product with its original parameterization, the parameter λ instead of c was adapted to each noise scenario.

In the following evaluation of the different fusion schemes, we will use the *Relative Word Error Rate (RWER)* instead of the WER. The reference point of the RWER is the WER resulting from a fusion according to the *Standard Weighted Product* with the original λ parameterization (SWP_λ) for the corresponding noise scenario. The $\text{RWER}(\text{SNR}, n)$ at a given noise type n and SNR level is defined as

$$\text{RWER}(\text{SNR}, n) = \frac{\text{WER}_{\text{fusion}}(\text{SNR}, n) - \text{WER}_{\text{ref}}(\text{SNR}, n)}{\text{WER}_{\text{ref}}(\text{SNR}, n)}. \tag{15}$$

To take all noise conditions into account the mean relative error for a particular fusion scheme over all noise conditions was calculated

$$e = \frac{1}{60} \sum_{(\text{SNR},n)} \text{RWER}(\text{SNR}, n). \tag{16}$$

An improvement compared to the Standard Weighted Product results in a negative RWER.

Table 1 compares the different mean relative errors. Both the FC and the FCA were implemented but due to the very poor performance of the identification network trained on the combined clean audio and video features in noise, resulting from a training on clean data, the performance of the FC was significantly worse than that of the FCA. For the Standard Weighted Product a parameterization with α and β is compared to the original parameterization with λ, which serves as the reference point for the evaluation of the relative error. Parameterization with α and β, which results in equal weights of 1 at $c = 0$ instead of 0.5 at $\lambda = 0.5$, leads to a small but consistent improvement over all noise types.

The results are given in detail in Figure 6 and Table 2, which show the graphical and numerical results, when car noise was added to the audio signal. For comparison, also the scores for the audio and video stream alone are given. Due to its poor performance, the FC is not included in this comparison. The SWP_λ is included to serve as a reference point. From Figure 6 and Tables 1 and 2, it follows that all weighted fusion schemes are able to fulfill the basic postulation of audio-visual recognition. This postulation states that the audio-visual score should always be better or equal to the audio or video score alone [18]. From a useful fusion scheme we further expect that it is able to generate synergy effects from the joint use of audio and video data in a way that the resulting error rates are significantly lower than the error rates from either stream alone. The Standard Weighted Product rather yields poor performance and shows only little gain from the joint use of audio and video data. Geometric Weighting and FCA give very similar results, which are much better for audio-visual recognition at medium SNR than audio or video recognition alone. For low SNRs the Geometric Weighting performs slightly, though not significantly better, than the FCA, but gives identical results for medium and high SNR. The Unweighted Bayesian Product is the only fusion scheme which does not fulfill the basic postulation. At very low SNR values, the recognition scores drop below those of the video channel alone, whereas at medium and high SNR values the scores are very similar, or identical, to those of the Geometric Weighting or the FCA.

Due to its superior performance, we only employed the Geometric Weighting in the following tests.

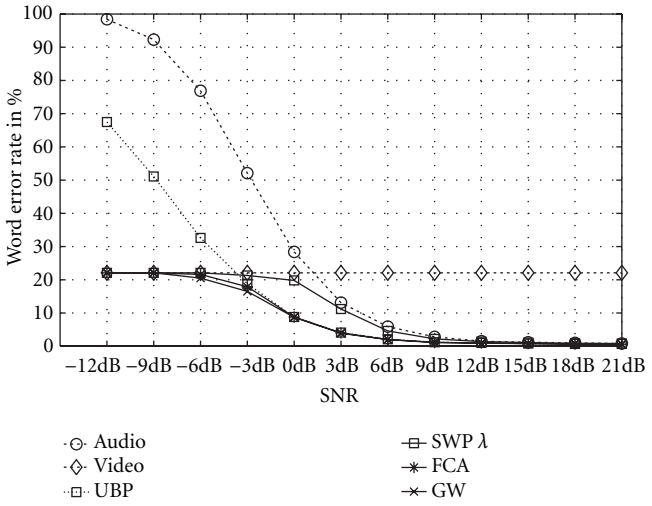


Figure 6: Word error rates for each individual stream and for audio-visual recognition with different fusion schemes. The fusion parameter was set by hand. Car noise was added to the audio channel.

Table 2: Comparison of the word error rates in percent for recognition with different fusion schemes, when car noise is added (WER on video alone is 22.1%).

	−12 dB	− 6 dB	0 dB	6 dB	12 dB	Clean
Audio	98.4	76.9	28.4	5.9	1.5	0.8
SWP_λ	22.1	22.1	19.8	4.6	1.4	0.8
UBP	67.5	32.6	8.8	2.1	1.0	0.7
FCA	22.1	21.6	8.8	2.0	0.8	0.6
Geometric	22.1	20.5	8.7	2.0	0.8	0.6

4.2. Automatic weight adaptation

For a real-time scenario, the setting of the weights has to be performed automatically depending on the noise level. A prerequisite to this is the estimation of the reliability of the audio stream during the fusion. The reliability estimation can follow two different approaches, either relying on the statistics of the a posteriori probabilities or directly on the speech signal. We will first present two measures based on the distribution of the a posteriori probabilities and will then also present a measure based on the speech signal.

4.2.1 Audio stream reliability estimation methods

Entropy of a posteriori probabilities

The distribution of the a posteriori probabilities at the output of the ANN carries information on the reliability of the input stream to the ANN. 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 quasi 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 [16, 33, 34]. The average entropy of the a posteriori probabilities over all frames is

$$H = -\frac{1}{K}\sum_{k=1}^{K}\sum_{n=1}^{N}\hat{P}(H_{n,k}|\mathbf{x}_{A,k})\log_2\hat{P}(H_{n,k}|\mathbf{x}_{A,k}), \tag{17}$$

where N is the number of phonemes and K the number of frames. 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. Experiments showed that for this mapping it is necessary to exclude segments where the pause is the most likely state, due to many false identifications of pauses at low SNR levels. Therefore only those frames, where the silence state is not amongst the 4 most probable phonemes, are taken into account for the calculation of the entropy.

Dispersion of a posteriori probabilities

A measure similar to the entropy is the dispersion of the a posteriori probabilities [16, 34]

$$D = \frac{1}{K}\sum_{k=1}^{K}\frac{2}{M(M-1)} \cdot \sum_{m=1}^{M}\sum_{l=m+1}^{M}\left(\log\left(\hat{P}(H_{m,k}|\mathbf{x}_{A,k})\right) - \log\left(\hat{P}(H_{l,k}|\mathbf{x}_{A,k})\right)\right), \tag{18}$$

where the probabilities $\hat{P}(H_{m,k}|\mathbf{x}_{A,k})$ are sorted in descending order, beginning with the highest one. Hence the difference between the M most likely phonemes is calculated and summed up. In our setup the best results were obtained for $M = 3$. As for the entropy, only frames where a silence is not among the 4 most likely phonemes are taken into account.

Voicing index as audio reliability measure

It is known that speech contains many harmonic components, whereas in many everyday life situations background noise is nonharmonic. Thus the lower the ratio of the energy of the harmonic to the nonharmonic components is, the more noise is present in the signal. A measure to asses this relation is the so-called *voicing index* [35]. The voicing index gives the conditional probability of a speech segment to be clean enough to be recognized when the harmonicity index R of this speech segment is known.

For the calculation of the harmonicity index, the speech signal is segmented into overlapping frames of 1024 sample values length and each speech segment is pre-emphasized and demodulated. The demodulation is performed by a rectification followed by a filtering with a trapezoidal band-pass filter. The cut-off frequencies of the band-pass filter are [0, 90, 350, 1000] Hz and hence cover the range of possible pitch values. After demodulation, the autocorrelation function of the speech segment is calculated. Inside a time window of the possible pitch values ([1/350, 1/90] s) the maximum value of the autocorrelation function is picked. We

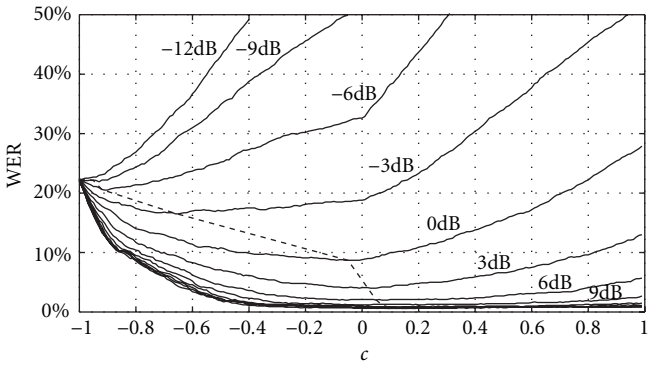


FIGURE 7: Relation between the fusion parameter c and the WER, when adding car noise at 12 SNR levels ranging from −12 dB to clean speech. The dashed line connects the points of minimum WER at a given SNR.

derive the value of the harmonicity index R by the normalization of this maximum value by the zero time-lag of the autocorrelation function, representing the mean energy of the demodulated speech frame. When setting a threshold for the signal to be "clean enough to be recognized" at an SNR level of 0 dB, the corresponding conditional probability, and hence the voicing index, can be formulated as $P(\text{SNR} > 0\,\text{dB}|R)$. For the evaluation of this conditional probability we use an estimate of the conditional probability density function. We added white noise at 0 dB SNR to 288 sentences of the database, and we compiled a bi-dimensional histogram of the relationship between the local SNR value (in each 1024 bins time frame) and the harmonicity index. A sigmoidal mapping function between the harmonicity R and the voicing index is derived from this histogram and used to estimate the conditional probability.

Similar to the previous criteria, the voicing index was evaluated only in those segments where the pause was not amongst the 4 most probable phonemes. First tests of the use of the voicing index for the fusion in audio-visual speech recognition are reported in [36].

4.2.2 Evaluation of global audio stream weights

After the definition of the various measures to be used in the estimation of the reliability of the audio stream, the questions at hand are: how sensitive are the recognition results to variations of the fusion parameter c, and how consistent are the reliability measures over different noise types and SNR levels?

To answer the first question we can have a look at Figure 7. Here the recognition results are plotted for c varying between $-1 \leq c \leq 1$. As additive noise, car noise at 12 SNR levels was used. The points of minimum WER used for the manual weight adaptation in Section 4.1 are connected by a dotted line in Figure 7. The goal of the automatic adaptation is now to find the mapping between the reliability estimation measure and the fusion parameter c, which results in the same minimum WERs in all noise conditions. As can be seen in the figure, there are large regions where the WER does not increase significantly over a wide range of values of the fusion parameter c. On the other hand, there are also regions at low SNR where small variations of the fusion parameter have a strong impact on the WER. In general, Figure 7 demonstrates that the fusion is not very sensitive to the setting of c for SNR> 0 dB, and hence an automatic choice of c should at least give reasonable results for these SNR values.

The next question is the sensitivity of the audio reliability estimation measures to different noise types. To test this sensitivity, we used all 5 noise types at 12 SNR levels each and calculated the average value of the corresponding reliability measure (entropy, dispersion, voicing index) over the whole test set for a given noise scenario. In Figure 8 we plotted the value of the reliability measure over the different optimal settings (i.e., the minimum WER $= f(c)$ points) of the fusion parameter c. Each point of the curves corresponds to one of the 12 SNR values and each of the first five curves corresponds to one noise type. If the criteria were independent of the noise type, all points of the curve would lie on one continuously decreasing (for the entropy) or increasing (for the dispersion and the voicing index) curve. This is obviously not the case. Nevertheless, the curves lie more or less close together, which indicates that the variation of the criteria with the noise type is rather small. Exceptions are the babble noise in the case of the dispersion and white noise for the voicing index.

If we want to have a reliability measure that does not depend on the noise type, we have to search for a mapping between the reliability measure ϱ and the fusion parameter c which is optimal in a minimum error sense. Our optimization criterion for the mapping $c(\varrho)$ is the minimization of the squared relative word error over all noise types n and all SNR levels [37]

$$f = \frac{1}{60} \sum_{(\text{SNR},n)} \text{RWER}(\text{SNR}, n)^2, \tag{19}$$

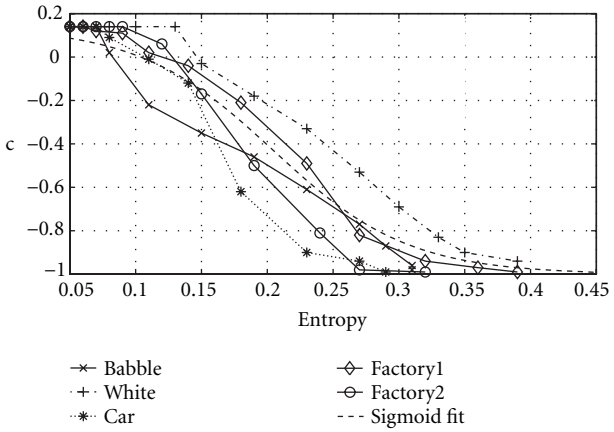
with the relative word error

$$\text{RWER}(\text{SNR}, n) = \frac{\text{WER}_{\min}(\text{SNR}, n) - \text{WER}_{\text{measure}}(\text{SNR}, n)}{\text{WER}_{\min}(\text{SNR}, n)}. \tag{20}$$

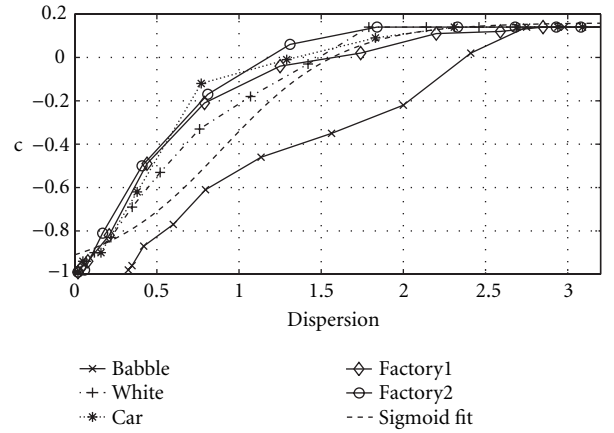
$\text{WER}_{\text{measure}}$ is the error rate obtained when using one of the reliability measures to control the fusion process, and $\text{WER}_{\min}$ is the minimum error rate when setting the fusion parameter manually (as defined in Section 4.1). The mapping between the reliability measure ϱ and the fusion parameter c is approximated by a sigmoidal function

$$c(\varrho) = \frac{h}{1 + g \cdot \exp(\varrho + d)} - 1, \tag{21}$$

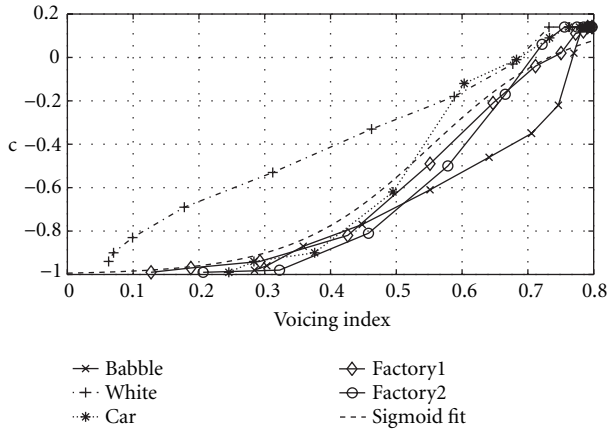
where h, g, and d are the parameters which define the shape of the sigmoidal function being subject to the optimization. The results of the optimization for each criterion can be seen in Figure 8. The sigmoidal mapping function is visualized as a dashed line. During the optimization the parameters g and d are evaluated following a gradient descent algorithm where



(a)



(b)



(c)

Figure 8: Relation between the three criteria and the fusion parameter c for five different noise types at varying SNR.

the unknown derivative is approximated by the difference quotient. The optimization procedure was repeated with different settings of the parameter h which gives the maximal attainable value of the function. The results of the setting of h yielding the minimum error in these optimizations are reported here. We want to point out, that the distance of the sigmoidal curve to the other curves in Figure 8 is not a direct measure for the quality of the fit. As a consequence of the minimization of the word errors, the optimal sigmoidal fit is the one which causes variations of the fusion parameter c from the optimal value which induces the smallest increase in word error. Hence, in regions where variations of c cause only a small increase of the word error, the distance of the sigmoidal curve and the curves resulting from the reliability criteria can be significant, whereas the resulting word error rates are still very close to optimal.

Table 3: Average RWER in percent for each criterion over all noise types and SNRs when fusion is done according to Geometric Weighting either with global or Frame Dependent evaluation of the criteria. Additionally the 95% confidence interval for the relative error is given.

	Entropy	Dispersion	Voicing index
Global	−27.5 ± 2.0	−26.4 ± 2.0	−27.0 ± 2.0
Frame dependent	−26.3 ± 2.0	−22.6 ± 2.0	−26.7 ± 2.0

4.2.3 Evaluation of adaptive audio stream weights

So far word error rates were calculated for a setting of the fusion parameter c being constant in one noise condition. The average value of the reliability measure was calculated in this noise condition and a global value of c for this noise condition was selected accordingly. This assumes that the whole test set is known at recognition time, which of course is unrealistic in a real life recognition system. Rather it is necessary to calculate the correct setting of the fusion parameter instantaneously for each frame. This also opens the possibility to cope with nonstationary noise and variations of the SNR of the speech signal. We therefore repeated the tests in the previous section with audio stream weights adapted on a frame by frame basis. To reduce the influence of estimation errors, the values of the fusion parameter were smoothed over time with a first order recursive filter with a cut off frequency of 0.6 Hz. Table 3 compares the results of the optimization for the different criteria, when the value of the fusion parameter is fixed over the whole test set (*Global*) and when it is varied (*Frame Dependent*). As for the previous recognition results, the average RWER is based on the results obtained with SWP_λ and hence evaluated according to (15) and (16). In Figure 9 the results of the automatic fusion, the manual setting of the fusion parameter, and the fusion using the Unweighted Bayesian Product are compared. For the automatic fusion the voicing index was chosen as the reliability measure and its evaluation was performed on a frame by frame basis. The curve corresponding to the global evaluation of the voicing index is almost identical to the frame-wise evaluation and therefore not included in the plot.

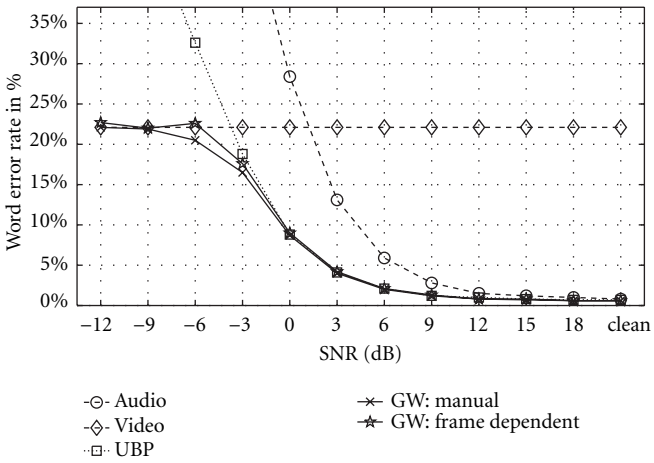


Figure 9: Word error rates for audio and video alone, fusion with Unweighted Bayesian Product, and the Geometric Weighting. For Geometric Weighting the fusion parameter was determined manually to give the best possible scores and adaptively according to the voicing index on a frame by frame basis.

5. DISCUSSION

In the previous sections, we presented different weight combination and estimation schemes of audio and video a posteriori probabilities in an audio-visual recognition task. Different tests were carried out to assess the performance of the different weighting schemes. In all tests, we used 5 different types of noise at 12 SNR levels each to obtain results not limited to one special scenario.

5.1. Performance of weight combination schemes

In the first test, the free parameters of the weighting schemes were adapted manually to each noise condition. Three of the presented weighting schemes, namely the Unweighted Bayesian Product, the FCA, and the Geometric Weighting, are based on the assumption of class conditional independence of audio and video features. The fourth one, Standard Weighted Product, only approximates this assumption for equal a-priori probabilities of the phonemes, which was not the case in our tests. Furthermore, the parameterization of the Standard Weighted Product is characterized by having a sum of weights equal to one. So, when both streams have equal weights of 0.5, the square root of the two a posteriori probabilities is taken instead of the product as for the other methods. In order to have weights equal to 1 on both streams in the equal weight condition (as for Geometric Weighting), we changed the parameterization of the Standard Weighted Product from λ and $(1-\lambda)$ to α and β, respectively. This led to a small, but consistent, improvement in comparison to the original form.

Yet the main result of this first comparison is the clear superior performance of the weighting schemes following the assumption of class conditional independence over the Standard Weighted Product, thanks to the introduction of the a-priori probabilities. Especially the FCA and the Geometric Weighting showed very similar results, where one reason is the similarity of the two algorithms (FCA is based on arithmetic weighting). Both attain the pure a posteriori probabilities when all weight is put on either channel and produce the a posteriori probability following class conditional independence for equal weights. They differ only in the way the probabilities are weighted, apart from these three special cases. The results indicate a small but not very significant advantage of the Geometric Weighting for low SNR values and equal performance for the other values. Therefore we only took the Geometric Weighting into consideration in the succeeding experiments.

5.2. Performance of audio stream reliability measures

The next test was designed to reveal the performance of the weighting scheme found best in the previous test in a more realistic scenario, where the adaptation of the weights is done automatically and not by hand. In the first step of the comparison we investigated a static case, where we first evaluated the reliability measure over the whole dataset and then performed the fusion with the setting of the fusion parameter corresponding to the measure. The mapping of the reliability measure to the fusion parameter took a wide range of noise conditions into account. For the mapping a fit in the minimum error sense between the value of the measure in a particular noise condition and the corresponding optimal fusion parameter was established. The results showed that large improvements compared to the audio-only recognition can be achieved under all noise conditions investigated, however for low SNR values the WERs are still too high to achieve useful recognition. An open question is how the optimized mapping generalizes to new, previously unseen, noise conditions. The consistency of the results (see Figure 8) proposes a possibility for generalization, even though final answers can only be found by tests in noise conditions not present during the design of the mapping.

In the last step of the comparison, we made the transition from the unrealistic static case, where the whole test set has to be known before determination of the fusion parameter, to an evaluation of the measure on a frame by frame basis. In general, we expected an increase in performance from the fact that a frame-level fusion is able to take variations of the SNR during one utterance and from one utterance to the other into account and it is capable to cope with nonstationary noise (like babble or factory noise). On the other hand, the limitation of the estimation interval of the reliability criteria to one frame has a high impact on the quality of the estimation. This effect was alleviated via smoothing the values with a first order recursive filter, although this reduces the ability to quickly adapt to intensity variations. The results of the frame-wise adaptation showed that both effects, the larger flexibility and the lesser precision seem to trade off one another. The results of the frame dependent evaluation are very similar to those evaluated on the whole test-set (see Table 3). Even though there was no performance gain from the frame-wise evaluation, the results show that the reliability estimation criteria are applicable to a realistic system. Both, the entropy and the voicing index, showed only small

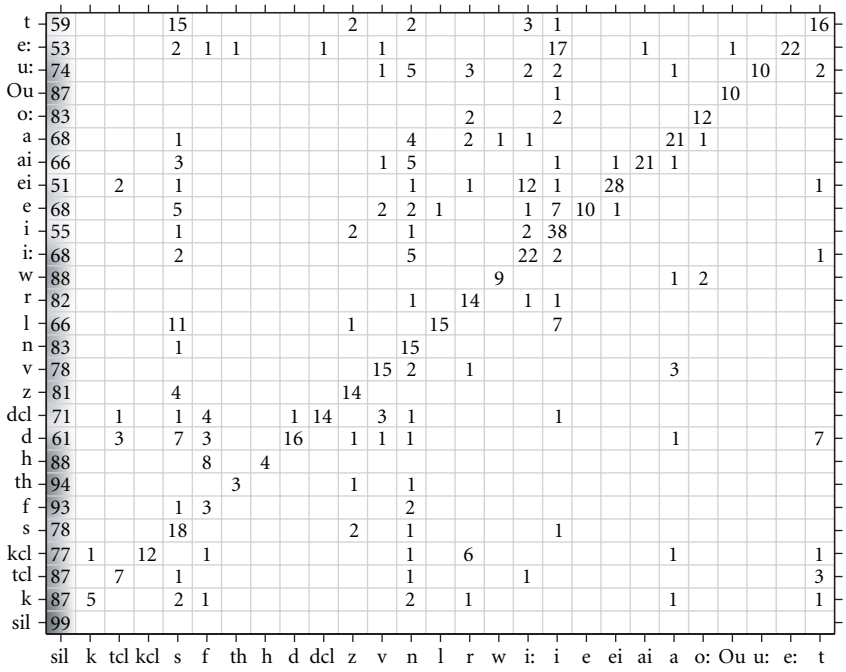


FIGURE 10: Confusion matrix of the phoneme identification from the audio stream at −6 dB when car noise was added to the signal, showing percentage of each phoneme on the y-axis identified as phonemes on the x-axis.

deviations from the optimum values. In the global evaluation the results of the entropy criterion were better than those of the voicing index, but in the more realistic frame-wise adaptation the entropy criterion deteriorated more than the voicing index. The voicing index, however, gave less consistent results for babble noise (which contains many harmonic components) and white noise (which has no harmonic components) than the entropy. The dispersion, especially in the frame-wise adaptation, was not competitive to the other criteria. To summarize, the entropy and the voicing index criterion can be used efficiently to control the adaptive fusion process.

5.3. Unweighted Bayesian Product versus adaptive weights

One interesting result of our comparison is the good performance at medium to high SNR of the Unweighted Bayesian Product, which does not require any weighting and hence no reliability estimation either. As can be seen in Figure 6, the performance of the Unweighted Bayesian Product is almost identical to that of the Geometric Weighting for medium and high SNR values (e.g., SNR $\geq$ 0 dB), whereas for low SNR values (e.g., SNR $<$ 0 dB), the performance sharply decreases.

For SNR $>$ 0 dB, audio and video channels carry complementary phonetic information which is well fused by Bayes' rule [20]. For SNR $<$ 0 dB there is a gain for the weighting principle, and Bayes' rule seems to start producing wrong results. Decreasing audio stream reliability results a-priori in an increase of the entropy of the corresponding categorization results, which is also exploited in the stream reliability criterion based on the entropy. This should result in a flattening of the distribution of the probability values and a corresponding increase in its entropy. In the extreme case, where the stream under consideration does not contribute any information, the output distribution of this stream becomes a uniform distribution. During fusion the uniform distribution does not interfere with the distribution of the reliable input stream, as the product of the uniform distribution does not alter the shape of the second distribution. Consequently, the phonetic identification is not impaired by the unreliable stream. If this is true, why can we then observe a sharp decrease of performance at low SNR?

To answer this question, we should have a look at the confusion matrix of the phoneme identification. In a confusion matrix, the elements of the matrix determine the percentage of the stimuli on the y-axis as being identified as the output class on the x-axis. For the confusion matrix in Figure 10, car noise at −6 dB was added to the audio signal.[2] Already a first quick look on the main diagonal of the confusion matrix reveals that the distribution of errors is clearly nonuniform. There are phonemes which are identified very well, and others which show only poor identification scores. The silence state "sil" obviously plays a special role. With increasing noise level more and more phonemes are confused with the silence state. This is partly taken into account in the mapping between the fusion parameter c and the stream reliability measures by the fact that only segments, where the silence is not among the 4 most likely states, are used for the evaluation of the criteria.

[2] The phonetic symbols follow the ARPABET notation.

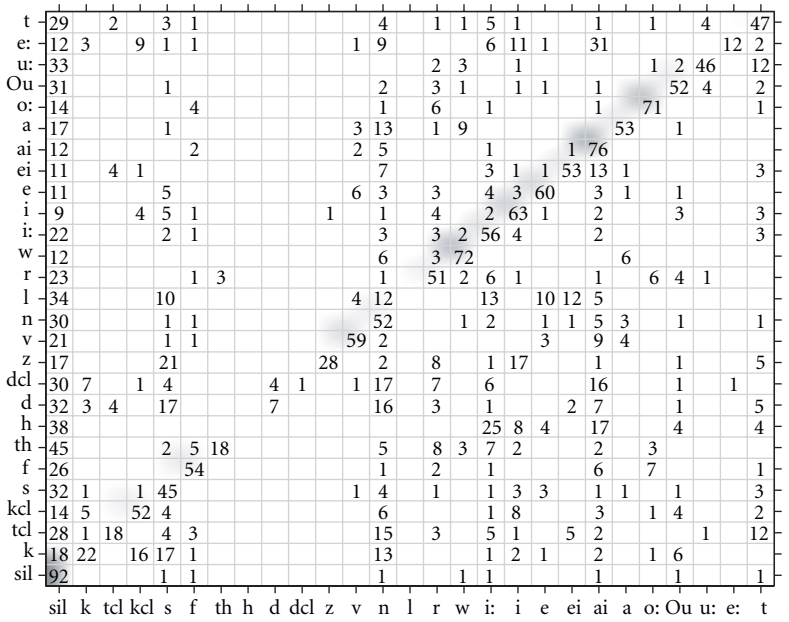


Figure 11: Confusion matrix of the phoneme identification from the video stream, showing percentage of each phoneme on the y-axis identified as phonemes on the x-axis.

Both the entropy and the dispersion criteria are improved by this modification. Furthermore also the phonemes "s," "n," and "i" attract many other phonemes. On the other hand, there are phonemes which are very poorly recognized and hardly any other phoneme is confused with them (e.g., "z" and "e:"). It follows from this analysis that the distribution of the a posteriori probabilities does not flatten but rather build certain peaks at some attractor phonemes and dips at phonemes which are hardly identified or confused.

Though, to impair the audio-visual recognition, not only an increase of errors in the audio stream has to occur, but these errors also have to be correlated with those committed in the video stream. The combination according to Bayes' rule is able to compensate for uncorrelated errors to a certain degree. Therefore, to judge the consequences of the deformation of the audio a posteriori probability distribution, it is indispensable also to look at the video stream confusion matrix visualized in Figure 11. Comparing the two confusion matrices demonstrates that the phonemes confused in the audio stream ("s," "n," and "i") also lead to confusions in the video stream. In the video stream the silence state also is the origin for many confusions. Hence the errors of phonetic identification in the audio and video stream are correlated, and in this case Bayes' rule is not able to perform a compensation.

It appears that in both confusion matrices the dominant cause for confusions is the silence state, but not equally. At −6 dB 75% of the phonemes are confused in the audio stream with the silence state. In the video stream, only 22% are misleadingly recognized as pauses.[3] When fusing audio and video following the Unweighted Bayesian Product, this strong preference for pauses at noisy audio leads to a confusion of 43% of the phonemes with pauses. Whereas when Geometric Weighting is used to weight the audio and video probabilities this confusion drops to 24%. The weighting has the tendency to select the modality having less confusion with the silence state.

Table 4: Mean relative error for GW with voicing index evaluated on a frame by frame basis and unweighted Bayesian product. The errors are calculated over all noise types and SNR levels and for SNR ≥ 0 dB and SNR < 0 dB, separately. Additionally, the 95% confidence interval for the relative error is given.

	GW/voicing index	UBP
All noise conditions	-26.7 ± 2.0	9.2 ± 2.1
SNR ≥ 0 dB	-39.4 ± 2.9	-35.0 ± 3.0
SNR < 0 dB	-1.1 ± 1.5	97.6 ± 1.7

Nevertheless, at medium SNR levels, the performance of the Unweighted Bayesian Product is very close to that of the Geometric Weighting. To further quantify this, in Table 4, in addition to the mean RWER of the Geometric Weighting and the Unweighted Bayesian Product over all noise conditions, also the mean RWER of the Unweighted Bayesian Product for SNR levels above and below 0 dB is given (all three evaluated according to (15) and (16)). From this evaluation it can be seen that the difference of performance between the UBP and the GW increases largely for SNR < 0 dB. Regardless of the remaining performance difference, there are applications where the SNR is typically higher than 0 dB, and a loss of performance is counterbalanced by a simple and intrinsically stable implementation.

[3]The distinction of speech from pauses in the video stream is not trivial as many nonarticulatory lip movements are part of continuous speech and are easily confused with a real utterance.

6. CONCLUSION

Our objective was to compare a number of schemes for an adaptive combination of audio and video a posteriori probabilities estimated by an ANN for an audio-visual recognition task under different noise conditions. In a first test we looked at the effectiveness of different weight combination schemes for audio and video data. The results demonstrated that a multiplicative combination respecting class conditional independence of the streams gives the best results. Next, we compared different criteria for an adaptive estimation of the audio stream reliability using the Geometric Weighting method. The performance of both, the criterion based on the entropy of the a posteriori probabilities and the one based on the ratio of the harmonic to the nonharmonic components in the speech signal, was very close to the best achievable performance determined by a manual adjustment. We showed that an adaptive weighting scheme based on the entropy and the voicing index can be built yielding consistent performance in various noise conditions. Finally, we investigated if a constant weight on the audio and video stream in all noise conditions would give comparable performance to the adaptive weighting. The test we made showed that when the SNR is higher than 0 dB, the Unweighted Bayesian Product performs as well as Geometric Weighting, so weighting, fixed or adaptive, is unnecessary. Whereas for SNR values below -3 dB performance losses are tremendous if no weighting is performed. An analysis of the confusion matrices showed that the confusion of all phonemes with the silence state is the main cause of the failure of the Unweighted Bayesian Product for SNR < 0 dB. We remark that this is related to the continuous speech recognition task and the problem of speech detection in noise. Therefore an algorithm (namely FCA and GW) incorporating Bayes' rule, which performs well for SNR ≥ 0 dB, and a weighting principle, being dominant for SNR < 0 dB, seems to be optimal. The weighting globally performs as a switch between the two modalities, favoring the one having less confusions with the silence state. This complements Bayes' rule, when this type of confusion occurs.

All tests are based on a database with a single male speaker whose lips were colored in blue, to facilitate the lip feature extraction. Most of the tests were repeated on a database with a single female speaker where no additional coloring of the lips was used [38]. The results of these tests are comparable to those reported here.

ACKNOWLEDGMENTS

We want to thank Christophe Savariaux for the recording of the database in April 1999, our subject John Barker for his effort and patience, Christian Sörensen, Thorsten Wild, and Vincent Charbonneau for carrying out many simulations, Gunther Sessler for his hints on statistics and Jürgen Lüttin and Gerasimos Potamianos for their thorough review of the paper. This work was partly funded by the EC program SPHEAR and is a part of the project RESPITE.

REFERENCES

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

[2] Q. Summerfield, "Lipreading and audio-visual speech perception," *Philos. Trans. Roy. Soc. London Ser. B*, vol. 335, no. 1273, pp. 71–78, 1992.

[3] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility," *Journal of the Acoustical Society of America*, vol. 26, no. 2, pp. 221–215, 1954.

[4] N. P. Erber, "Auditory-visual perception of speech," *Journal of Speech and Hearing Disorders*, vol. 40, pp. 481–492, 1975.

[5] Q. Summerfield, "Some preliminaries to a comprehensive account of audio-visual speech perception," in *Hearing by Eye: The Psychology of Lipreading*, B. Dodd and R. Campell, Eds., pp. 3–51, Lawrence Erlbaum Associates, Hove, UK, 1987.

[6] E. D. Petajan, *Adaptive determination of audio and visual weights for automatic speech recognition*, Ph.D. thesis, University of Illinois, Urbana, Ill, USA, 1984.

[7] D. G. Stork, G. Wolff, and E. Levine, "A neural network lipreading system for improved speech recognition," in *Proc. International Joint Conference on Neural Networks*, pp. 285–295, Baltimore, Md, USA, June 1992.

[8] P. Duchnowski, U. Meier, and A. Waibel, "See me, hear me: Integrating automatic speech recognition and lip-reading," in *Proc. International Conference on Spoken Language Processing*, pp. 547–550, Yokohama, Japan, 1994.

[9] P. L. Silsbee and A. C. Bovik, "Computer lipreading for improved accuracy in automatic speech recognition," *IEEE Trans. Speech, and Audio Processing*, vol. 4, no. 5, pp. 337–351, 1996.

[10] S. Dupont and J. Lüttin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, vol. 2, no. 3, pp. 141–151, 2000.

[11] C. Neti, P. Potamianos, J. Lüttin, et al., "Audio-visual speech recognition," Tech. Rep., Center for Language and Speech Processing, Johns Hopkins University, Baltimore, Md, USA, 2000.

[12] T. Chen, "Audiovisual speech processing: lip reading and lip synchronization," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 9–21, 2001.

[13] G. Potamianos, C. Neti, G. Iyengar, and E. Helmuth, "Large vocabulary audio-visual speech recognition by machines and humans," in *Proc. 7th European Conference on Speech Communication and Technology*, Aalborg, Denmark, September 2001.

[14] J. L. Schwartz, J. Robert-Ribes, and P. Escudier, "Ten years after summerfield: A taxonomy of models for audio-visual fusion in speech perception," in *Hearing by Eye II*, R. Campell, B. Dodd, and D. Burnham, Eds., pp. 85–108, Taylor & Francis Books, London, UK, 1998.

[15] B. Gold and N. Morgan, *Speech and Audio Signal Processing*, John Wiley & Sons, New York, NY, USA, 2000.

[16] A. Adjoudani and C. Benoit, "On the integration of auditory and visual parameters in an HMM-based ASR," in *Speechreading by Man and Machine: Models, Systems and Applications*, D. G. Stork and M. E. Hennecke, Eds., NATO ASI Series, pp. 461–472, Springer, Berlin, 1996.

[17] A. Rogozan and P. Deléglise, "Adaptive fusion of acoustic and visual sources for automatic speech recognition," *Speech Communication*, vol. 26, no. 1-2, pp. 149–161, 1998.

[18] 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*, vol. 7, no. 9, pp. 629–642, 1999.

[19] J. R. Movellan and G. Chadderdon, "Channel separability in the audio-visual integration of speech: A Bayesian approach,"

in *Speachreading by Man and Machine: Models, Systems and Applications*, D. G. Stork and M. E. Hennecke, Eds., NATO ASI Series, pp. 473–487, Springer, Berlin, 1996.
[20] D. W. Massaro and D. G. Stork, "Speech recognition and sensory integration," *American Scientist*, vol. 86, no. 3, pp. 236–244, 1998.
[21] D. W. Massaro, *Speech Perception by Eye and by Ear: A Paradigm for Physiological Inquiry*, Lawrence Erlbaum Associates, Hillsdale, NJ, USA, 1987.
[22] M. Heckmann, F. Berthommier, and K. Kroschel, "Optimal weighting of posteriors for audio-visual speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Salt Lake City, Utah, USA, May 2001.
[23] H. Fletcher, "The nature of speech and its interpretation," *J. Franklin Instit.*, vol. 193, no. 6, pp. 729–747, 1922.
[24] J. B. Allen, "How do humans process and recognize speech," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 2, no. 4, pp. 567–576, 1994.
[25] H. Bourlard and S. Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands," in *Proc. International Conf. on Spoken Language Processing*, pp. 422–425, Philadelphia, Pa, USA, October 1996.
[26] A. Morris, A. Hagen, H. Glotin, and H. Bourlard, "Multi-stream adaptive evidence combination for noise robust ASR," *Speech Communication*, vol. 34, pp. 25–40, 2001.
[27] R. A. Cole, T. Noel, L. Lander, and T. Durham, "New telephone speech corpora at CSLU," in *Proc. 4th European Conference on Speech Communication and Technology*, pp. 821–824, Madrid, Spain, September 1995.
[28] N. Morgan and H. Boulard, "Continuous speech recognition," *IEEE Signal Processing Magazine*, vol. 12, no. 3, pp. 24–42, 1995.
[29] University of Mons, *Step by Step Guide to using the Speech Training and Recognition Unified Tool (STRUT)*, Mons, France, May 1997.
[30] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "RASTA-PLP speech analysis technique," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, pp. 121–124, San Fransisco, Calif, USA, March 1992.
[31] E. Lombard, "Le signe de l'élévation de la voix," *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, vol. 37, pp. 101–119, 1911.
[32] 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," Tech. Rep., Speech Research Unit, Defence Research Agency, Malvern, UK, June 1992.
[33] C. Bregler, H. Hild, S. Manke, and A. Waibel, "Improving connected letter recognition by lipreading," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, pp. 557–560, Minneapolis, Minn, USA, 1993.
[34] G. Potamianos and C. Neti, "Stream confidence estimation for audio-visual speech recognition," in *Proc. International Conf. on Spoken Language Processing*, vol. III, pp. 746–749, Bejing, China, October 2000.
[35] F. Berthommier and H. Glotin, "A new SNR feature mapping for robust multistream speech recognition," in *Proc. Int. Congress on Phonetic Sciences (ICPhS)*, vol. 1, pp. 711–715, San Fransisco, Calif, USA, 1999.
[36] H. Glotin, D. Vergyri, C. Neti, G. Potamianos, and J. Lüttin, "Weighting schemes for audio-visual fusion in speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 173–176, Salt Lake City, Utah, USA, May 2001.
[37] M. Heckmann, T. Wild, F. Berthommier, and K. Kroschel, "Comparing audio- and a-posteriori-probability-based stream confidence measures for audio-visual speech recognition," in *Proc. 7th European Conference on Speech Communication and Technology*, pp. 1023–1026, Aalborg, Denmark, September 2001.
[38] M. Heckmann, K. Kroschel, F. Berthommier, and C. Savariaux, "DCT-based video features for audio-visual speech recognition," submitted to *International Conf. on Spoken Language Processing*, Denver, Col, USA, 2002.

Martin Heckmann studied electrical engineering at the University of Karlsruhe, Germany and the Institut National des Sciences Appliquées (INSA) in Lyon, France. He received his diploma in 1997 from the University of Karlsruhe. Currently he is pursuing a Ph.D. at the Institute for Telecommuncations Engineering at the University of Karlsruhe. His research interests include audio-visual speech recognition, face and lip tracking, and image enhancement. Parts of his research he conducted at the Institut de la Communication Parlée (ICP) in Grenoble, France. He is a student member of IEEE and ISCA.

Frédéric Berthommier received the M.D. degree from the University of Paris 7 (Lariboisière St-Louis) and his Ph.D. in biomedical engineering from the University of Grenoble I in 1992. He is a CNRS researcher in the "Institut de la Communication Parlée" in Grenoble since 1993. Frédéric's research interests include auditory scene analysis, auditory modelling, auditory perception, audio-visual speech processing and perception of odour mixtures. He is a task manager for the European projects SPHEAR and RESPITE focusing on the development of robust speech recognition systems.

Kristian Kroschel studied electrical engineering at the universities of Karlsruhe and Erlangen-Nürnberg, Germany. In 1971 he received the Ph.D. degree from the University of Karlsruhe and since 1977 he is a professor for communication engineering at the Department of Electrical Engineering and Information Technology at the University of Karlsruhe. From 1987 through 1991 he headed in parallel the group Digital Signal Processing and Diagnosis at the Fraunhofer Institut of Information and Data Processing (IITB) in Karlsruhe. In the field of signal processing and communication he has published more than 60 papers and 3 books, partly together with other authors. His special interest is on noise reduction and echo compensation applied to speech communication and recognition. He was a visiting professor, at the Technion Haifa, Israel, the University of California, Santa Barbara, California, the Bandung Institute of Technology, Indonesia, and other institutes. He is a member of the VDE, the German association of engineers in electrical engineering, electronics, and information technology.