# OPTIMAL WEIGTHING OF POSTERIORS FOR AUDIO-VISUAL SPEECH RECOGNITION

*Martin Heckmann*[◇•], *Frédéric Berthommier*[◇]

*Kristian Kroschel*[•]

[◇]Institut National Polytechnique de Grenoble
Institut de la Commuinication Parlée
46, Av. Félix Viallet, 38031 Grenoble, France
bertho@icp.inpg.fr

[•]Universität Karlsruhe
Institut für Nachrichtentechnik
Kaiserstraß e 12, 76128 Karlsruhe, Germany
{heckmann, kroschel}@int.uni-karlsruhe.de

## ABSTRACT

We investigate the fusion of audio and video a posteriori phonetic probabilities in a hybrid ANN/HMM audio-visual speech recognition system. Three basic conditions to the fusion process are stated and implemented in a linear and a geometric weighting scheme. These conditions are the assumption of conditional independence of the audio and video data and the contribution of only one of the two paths when the SNR is very high or very low, respectively. In the case of the geometric weighting a new weighting scheme is developed whereas the linear weighting follows the Full Combination approach as employed in multi-stream recognition. We compare these two new concepts in audio-visual recognition to a rather standard approach known from the literature. Recognition tests were performed in a continuous number recognition task on a single speaker database containing 1712 utterances with two different types of noise added.

## 1. INTRODUCTION

In recent years much attention was put on the integration of multiple streams, so called multi-stream, for noise robust speech recognition. In most cases these streams are the result of different preprocessing of the audio signal. One application of this multi-stream approach is audio-visual speech recognition. Here the two streams, namely the acoustic signal and the lips movement, carry complementary information. The visual modality additionally contributes information on the place of articulation. E.g. /t/ and /p/ are well discriminated via the lips movement but can easily be confused when looking on a noisy acoustical signal only.

In this paper we state three basic conditions which, in our opinion, should be fulfilled in a weighting scheme for audio-visual fusion. We implement these conditions in a newly developed geometric weighting scheme in the framework of a so called *Separate Identification (SI)* architecture [1], which can be seen as a special case of multi-stream recognition. Furthermore, as an example of a linear weighting scheme, we apply the so called *Full-Combination (FC)* approach, known from multi-stream recognition [2], to our knowledge for the first time to audio-visual recognition. In the implementation of the FC we also take these basic conditions into account. We then compare the new geometric and linear fusion scheme to one common for audio-visual speech recognition.

## 2. RECOGNITION SYSTEM

To compare the different fusion methods we use an ANN/-HMM hybrid model for continuous audio-visual number recognition. Identification of the phonemes is performed independently for the audio and the video path (compare Fig. 1) and thus follows a SI or multi-stream approach. The ANNs are trained to produce 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. The goal of the fusion process is to obtain a good estimation of the a posteriori probability $P(H_i|\mathbf{x}_A, \mathbf{x}_V)$ when both $\mathbf{x}_A$ and $\mathbf{x}_V$ are present.

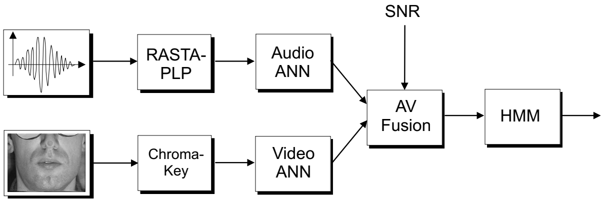

**Fig. 1**. Separate Identification (SI) audio-visual speech recognition system

Implementation of the system was carried out with the tool STRUT from TCTS lab Mons, Belgium [3]. To train

This work was partly funded by the EC program SPHEAR and is a part of the project RESPITE

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 [4]. The database consists of 1712 utterances each comprising several numbers representing a street address, zipcode or similar, as in NUMBERS95 from the Oregon Graduate Institute (OGI). For training 884 and for testing 828 utterances were used. The audio feature extraction is performed with RASTA-PLP using 12 cepstral coefficients and the log energy. Video features are extracted via a *chroma key* process, which requires coloring of the speakers lips with blue ink (see also Fig. 1)[5]. Due to the coloring, the lips can then be located easily and their movement parameters can be extracted in real time. As lips parameters

- outer lip width
- inner lip width
- outer lip height
- inner lip height
- lip surface and
- mouth surface surrounded by lips

were chosen.

To take temporal information into account, several time frames of the audio and video feature vectors are presented simultaneously at the input of the corresponding ANNs. The generated posteriors are directly forwarded to the HMM without a conversion to so called scaled likelihoods. Left-right HMM models were employed to model the individual phonemes. The order of the HMMs used to represent the different phonemes was adapted to the mean length of the corresponding phoneme. Word models are 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. No distinction between phonemes and visemes was made in the fusion process. Each acoustical articulation is assumed to have a synchronously generated corresponding visual articulation.

## 3. FUSION OF AUDIO AND VIDEO

The goal of the fusion process is to calculate the a posteriori probability, when both the audio and the video features are present, via the posteriors of each individual stream. In this chapter we state the 3 conditions to the fusion process and introduce a new geometric fusion method, the Full Combination approach and the standard fusion method known from the literature, which will serve as a benchmark.

### 3.1. Boundary Conditions to the Fusion Process

Massaro&Stork [6] and Movellan&Challderon [7] gave evidence that the acoustical and the visual features can be considered as conditionally independent:

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

Considering this assumption and applying Bayes' rule the desired a posteriori probability of the phoneme $H_i$ can be formulated as:

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

Unfortunately the ANNs only give us an estimate $\hat{P}$ of the true posteriors. Particularly in the case of the audio path, this estimate strongly depends on the additional noise present. It is therefore desirable to control the influence of the audio and video features depending on the noise level. 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:

$$\hat{P}(H_i | \mathbf{x}_A, \mathbf{x}_V) \rightarrow \hat{P}(H_i | \mathbf{x}_V) \tag{3}$$

We want to refer to this in the following as *Condition I* to the fusion process. Similar for very high SNR the estimation in the audio path is in general much better than the one in the video path and consequently

$$\hat{P}(H_i | \mathbf{x}_A, \mathbf{x}_V) \rightarrow \hat{P}(H_i | \mathbf{x}_A) \tag{4}$$

This will be referred to as *Condition II*. Finally the evaluation of the a posteriori probability according to Eq. 2 in cases where both, the audio and the video path, contribute equally will be referred to as *Condition III*. In the next two sections we develop a geometric and a linear weighting method which takes *Conditions I-III* into account.

### 3.2. Geometric Weighting

To develop the Geometric Weighting scheme we replace the terms independent of the actual phoneme $H_i$ by a normalization factor:

$$\varepsilon(\alpha, \beta) = \frac{1}{\sum_{j=0}^{N-1} \frac{\hat{P}^{\alpha}(H_j|\mathbf{x}_A)\hat{P}^{\beta}(H_j|\mathbf{x}_V)}{\hat{P}^{\alpha+\beta-1}(H_j)}} \tag{5}$$

Then the final estimate of the a posteriori probability of the phoneme $H_i$ can be written as:

$$\hat{P}_{GW}(H_i | \mathbf{x}_A, \mathbf{x}_V) = \frac{\hat{P}^{\alpha}(H_i|\mathbf{x}_A)\hat{P}^{\beta}(H_i|\mathbf{x}_V)}{\hat{P}^{\alpha+\beta-1}(H_i)} \cdot \varepsilon(\alpha, \beta) \tag{6}$$

The weighting parameters $\alpha$ and $\beta$ both depend on a third parameter $c$ according to:

$$\begin{aligned} \alpha &= \frac{1}{1 + \exp(-c - 5)} \\ \beta &= \frac{1}{1 + \exp(c - 5)} \\ & -\infty \le c \le \infty \end{aligned} \tag{7}$$

The parameter $c$ varies with the SNR and is so far adjusted manually so as to give best recognition results. When $c = 0$ the posteriors from the audio and video path both have the same weight as $\alpha \simeq 1$ and $\beta \simeq 1$. Hence *Condition III* is fulfilled. For very low $c$ and thus very low SNR $\alpha \simeq 0, \beta \simeq 1$ and for very high $c$ corresponding to very high SNR $\alpha \simeq 1, \beta \simeq 0$. Consequently these two cases fulfill *Conditions I&II*.

In the following sections we will refer to this fusion method as *Geometric Weighting*.

### 3.3. Full Combination Approximation

As a method of linear weighting we consider the so called *Full Combination (FC)* approach developed in the framework of multi-stream speech recognition [2]. In our case the two streams are the audio and video signal. In the FC approach the final a posteriori probability is derived via the weighted linear sum over all possible combinations of the streams. In the case of audio-visual recognition we have to consider during combination the "empty" stream, containing only the priors, the audio, the video and the combined audio/video stream:

$$\begin{aligned} \hat{P}_{FC}(H_i|\mathbf{x}_A, \mathbf{x}_V) \quad = \quad & a_1\hat{P}(H_i|\mathbf{x}_A, \mathbf{x}_V) + \\ & a_2\hat{P}(H_i|\mathbf{x}_A) + \\ & a_3\hat{P}(H_i|\mathbf{x}_V) + \\ & a_4\hat{P}(H_i) \end{aligned} \tag{8}$$

In Eq.8

$$\hat{P}(H_i|\mathbf{x}_A, \mathbf{x}_V) = \gamma \frac{\hat{P}(H_i|\mathbf{x}_A)\hat{P}(H_i|\mathbf{x}_A)}{\hat{P}(H_i)} \tag{9}$$

is the a posteriori probability of the combined audio and video stream. To evaluate this probability again conditional independence of the audio and video data is assumed. This assumption leads to the so called *Full Combination Approximation (FCA)*[2]. The parameter $\gamma$ replaces the terms independent of the phoneme $H_i$ and is evaluated correspondingly to Eq. 6.

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 $\alpha$ and $\beta$ as given in Eq. 7. 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. *Condition III* is fulfilled via $c = 0$. Similar for $c \ll 0$ and $c \gg 0$, respectively *Conditions I&II* are fulfilled.

In the common approach for FCA the weights of each stream are adapted according to the reliability of the particular stream. The reliability of the audio signal strongly depends on the noise level, whereas, as a consequence of the recording conditions, the reliability of the visual signal is constant. To determine the weights we applied a manual optimization approach similar to the one used for the Geometric Fusion. In our implementation the degrees of freedom of the FCA and the Geometric Fusion are limited to one, which might not be optimal but a multidimensional optimization is not practicable.

### 3.4. Standard Fusion

In the literature conditional independence of the audio and video features is assumed in systems which rely on likelihoods via direct weighting of Eq. 1 [5][8]. Whereas in systems based on a posteriori probabilities, as our implementation, Eq. 1 is modified according to (compare [1][9]):

$$\hat{P}_{Std}(H_i|\mathbf{x}_A, \mathbf{x}_V) = \hat{P}^{\alpha}(H_i|\mathbf{x}_A)\hat{P}^{\beta}(H_i|\mathbf{x}_V) \cdot \delta(\alpha, \beta) \tag{10}$$

with the weighting factors given in Eq. 7 [1] and the normalization factor $\delta$, which is straightforward. In this case only *Condition I&II* are fulfilled for $\alpha = 1, \beta = 0$ and $\alpha = 0, \beta = 1$.

This fusion method will be referred to as *Standard Fusion*.

## 4. PERFORMANCE COMPARISON

To compare the different fusion modalities we added noise at different SNR levels to the audio signal and then performed recognition tests implementing the different fusion strategies. The recognition systems were identical except for the fusion process. The free parameter $c$ in the fusion process was tuned by hand so to obtain the best possible result for each fusion modality. The a priori probabilities involved in the FCA and the Geometric Fusion were estimated from the training set of the database. Two sets of tests with two different types of noise were carried out. In the first test white Gaussian noise was added. In the second test noise, which we recorded in a car under real driving conditions on a motor-way, was used. This noise is almost stationary and has a lowpass characteristic.

The results of the comparison with white noise are visualized in Fig.2. It can be seen that for all SNR values the audio-visual recognition performs better or at least as good as video or audio alone. The FCA and the Geometric Fusion perform very similar in all cases and they give significantly better results than the standard fusion. Furthermore, starting at SNR values around $-6dB$ up to clean speech, synergy effects arising from the joint use of audio and video data are clearly visible for the FCA and the Geometric Fusion. At low SNR values the Geometric Fusion performs slightly better than the FCA. The numerical values at some exemplary points are given in Tab. 1.

[1] Weighting with $\alpha = (1 - \lambda)$ and $\beta = \lambda$, $0 \leq \lambda \leq 1$, as originally implemented in the literature leads to identical results

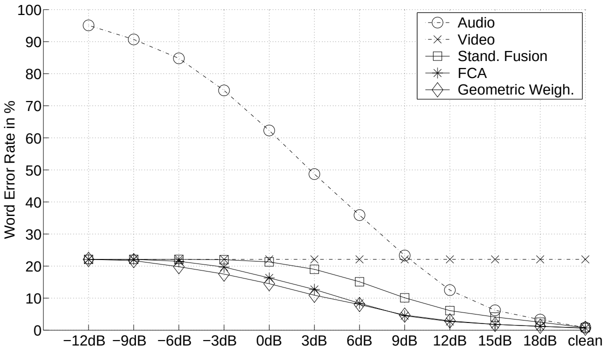

**Fig. 2**. Comparison of the word error rates for recognition with different fusion methods when white noise is added

| | -6dB | -3dB | 3dB | 6dB | 9dB | clean |
|---|---|---|---|---|---|---|
| Audio | 84.8% | 74.8% | 48.7% | 35.9% | 23.3% | 0.8% |
| Standard | 22.1% | 22.1% | 19.0% | 15.2% | 10.1% | 0.8% |
| FC | 21.8% | 20.0% | 12.6% | 8.8% | 4.8% | 0.6% |
| Geometric | 20.1% | 18.0% | 11.4% | 8.1% | 4.8% | 0.6% |

**Table 1**. Comparison of the word error rates for recognition with different fusion methods when white noise is added (WER on video alone is 22.1%)

Fig.3 shows the results when using noise recorded in a car. The results are very similar to those obtained with white noise. Besides the rather poor performance of the Standard Fusion also the small difference between the Geometric Fusion and the FCA for low SNR values is visible.

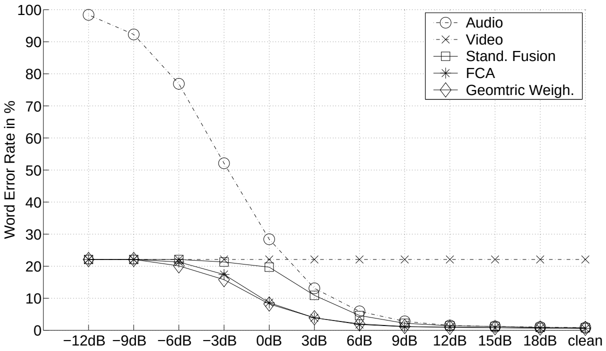

**Fig. 3**. Comparison of the word error rates for recognition with different fusion methods when noise recorded in a car is added

## 5. CONCLUSION

We formulated three conditions to the fusion of audio and video data in a hybrid ANN/HMM audio-visual recognition system, which allows direct weighting of the posteriors. We applied these conditions to a geometric and a linear weighting scheme. The conditions comprise the assumption of conditional independence of the audio and video data but also take the situations, where only the data of one channel is reliable, into account. The linear weighting scheme introduced follows the Full Combination Approximation approach, known from multi-stream recognition. We then compared these two new concepts to one already known from the literature. To test the different fusion modalities we used an audio-visual database containing 1712 sentences each representing a sequence of numbers. A continuous recognition was performed where two different types of noise at different SNRs were added to the audio channel. The results for both noise conditions are identical. The FCA and the Geometric Fusion clearly outperform the Standard Fusion in all situations. At high noise levels the newly introduced Geometric Fusion gives slightly better results than the FCA, whereas for medium and good SNR their performance is almost identical.

## 6. ACKNOWLEDGMENTS

We want to thank Hervé Glotin for his technical aid to setup the audio recognition system and Vincent Charbonneau for carrying out parts of the simulations.

## 7. REFERENCES

[1] A. Rogozan and P. Deléglise, “Adaptive fusion of acoustic and visual sources for automatic spech recognition,” *Speech Communication*, vol. 26, pp. 149–161, 1998.

[2] A. Morris, A. Hagen, H. Glotin, and H. Bourlard, “Multi-stream adaptive evidence combination for noise robust asr,” *to appear in Speech Communication*, 2000.

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

[4] M. Heckmann, F. Berthommier, C. Savariaux, and K. Kroschel, “Labeling audio-visual speech corpora and training an ann/hmm audio-visual speech recognition system,” in *Proc. ICSLP 2000 Bejing*, 2000.

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

[6] D. W. Massaro and D. G. Stork, “Speech recognition and sensory integration,” *American Scientist*, vol. 86, no. 3, 1998.

[7] 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., Berlin, 1996, NATO ASI Series, pp. 473–487, Springer.

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

[9] 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 Audio Processing*, vol. 7, pp. 629–642, 1999.