Comparing Audio- and A-Posteriori-Probability-Based Stream Confidence Measures for Audio-Visual Speech Recognition

Martin Heckmann[◇], *Thorsten Wild*[◇], *Frédéric Berthommier*[•], *Kristian Kroschel*[◇]

[◇]Institut für Nachrichtentechnik
Universität Karlsruhe, Germany
{heckmann,wild,kroschel}@int.uni-karlsruhe.de

[•]Institut de la Communication Parlée
Institut National Polytechnique de Grenoble, France
bertho@icp.inpg.fr

Abstract

During the fusion of audio and video information for speech recognition, the estimation of the reliability of the noise affected audio channel is crucial to get meaningful recognition results. In this paper we compare two types of reliability measures. One is the use of the statistics of the phoneme a-posteriori probabilities and the other is the analysis of the audio signal itself. We implemented the entropy and the dispersion of the probabilities and, from the audio-based criteria, the so called Voicing Index. To test the criteria a hybrid ANN/HMM audio-visual recognition system was used and 5 different types of noise at 12 SNR levels each were added to the audio signal. The best sigmoidal fit for each criterion between the fusion parameter and the value of the criterion over all noise types and SNR values was performed. The resulting individual errors and the corresponding averaged relative errors are given.

1. Introduction

To improve the recognition results of automatic speech recognition systems in adverse environments, recently in particular the integration of multiple input streams is considered. One application of this multi-stream approach is audio-visual speech recognition. Under the assumption of constant recording conditions of the lips movement, only the audio stream shows a varying degree of reliability depending on the additional background noise present. As a consequence of severe audio recognition performance degradations due to additional noise, the determination of the reliability of the audio stream and the corresponding setting of the parameters in the fusion process is of great importance.

In this paper we want to investigate two different approaches to the estimation of the reliability. The first is to exploit the information which is present in the distribution of the a-posteriori probabilities of the phonemes. In the second approach the level of the background noise is directly determined from the audio signal. We investigated the entropy and the dispersion of the a-posteriori probabilities which belong to the first category and the Voicing Index which is evaluated on the audio signal. For our tests we added 5 different types of noise at 12 *Signal to Noise Ratio (SNR)* values each. Via an optimization in respect to the minimal resulting *Word Error Rate (WER)* we determined the best fit between the parameter of the fusion and the values of the different criteria. From this data we calculated the absolute error at each tested case and the averaged relative error over all cases for the three criteria in order to compare them under the same conditions.

2. The Recognition System

To compare the different criteria 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 [1].

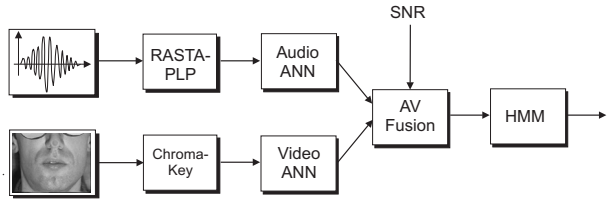

Figure 1: Separate Identification (SI) audio-visual speech recognition system

The ANNs are trained to produce estimates of 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.

Fusion of the estimated audio and video a-posteriori probabilities follows a Geometric Weighting [2]:

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

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

$$\alpha = \begin{cases} 0 & c < -1 \\ 1+c & -1 \le c \le 0 \\ 1 & c > 0 \end{cases}$$

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

The parameter c controls the weighting of audio and video data. When $c \simeq -1$ only the video signal contributes to the recognition, whereas for $c \simeq 1$ the recognition relies completely on the audio signal. $\epsilon(\alpha, \beta)$ is a normalization parameter independent on the actual phoneme. The weighting parameter is fixed during one test and applied at each individual frame.

Implementation of the system was carried out with the tool STRUT from TCTS lab Mons, Belgium [3]. To train the ANNs and to perform the recognition tests we used a single-speaker audio-visual database recorded at the Institut de la Communication Parlée (ICP) in Grenoble, France [4]. The database contains 1712 utterances each comprising several continuously uttered numbers. The input vectors to the MLPs consist of RASTA-PLP coefficients and 6 geometric lip features.

3. Automatic Fusion of Audio and Video Data

At high SNR the audio signal contributes much more information to the audio-visual recognition process than the video signal (compare $< 1\%$ of WER on the audio signal on clean speech to 22.1% WER for the video signal in our setup). This situation changes with increasing noise level in the audio signal until the audio signal almost contributes no information, whereas the reliability of the video signal is not affected by the additional acoustic noise[1]. For a given SNR, variations of the audio and video weights severely affect the final recognition score. As a consequence, an estimation of the reliability of the audio signal and the corresponding setting of the weights is of great importance for the fusion process.

The reliability estimation can follow two different approaches, namely relying on the statistics of the a-posteriori probabilities or directly on the speech signal. We will first present two statistics-based measures and will then also present a measure based on the speech signal.

3.1. Statistics-Based Reliability Measures

The distribution of the a-posteriori probabilities at the output of the MLP carries information on the reliability of the input stream to the MLP. 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. Two methods proposed in the literature to exploit these distributions for the fusion in audio-visual speech recognition are the use of the entropy and the dispersion of the probabilities [5][6].

Entropy of A-Posteriori Probabilities

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 $\mathbf{x}_{A,k}$ in time frame k can be determined via:

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

where N is the number of phonemes and K the number of frames. Experiments showed that it is necessary to exclude segments where a pause is the most likely phoneme to get estimates of the entropy suitable for the control of the fusion process due to many false identifications of pauses at low SNR levels. Therefore only those frames, where the pause 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:

$$D = \frac{1}{K}\sum_{k=1}^{K}\frac{2}{N(N-1)}\cdot \quad (4)$$
$$\sum_{n=1}^{N}\sum_{l=n+1}^{N}\Big(log(\hat{P}(H_{n,k}|\mathbf{x}_{A,k})) - log(\hat{P}(H_{l,k}|\mathbf{x}_{A,k}))\Big)$$

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

3.2. Voicing as Audio Reliability Measure

The Voicing Index is established from the acoustic signal, independently of the recognition process. It determines the degree of harmonicity of the acoustic signal, considering that speech contains many voiced segments and background noise in general is non-harmonic. In each time frame, we calculate a value which depends on the local SNR and which provides information about the reliability of the speech, as proposed in [7]. This information can be interpreted as a probability for an audio-frame to be clean enough to be recognized, knowing the value of the harmonicity index R, i.e. $P(SNR > 0dB|R)$. This harmonicity index is calculated after pre-emphasis and demodulation of the signal, using the autocorrelation function. In each time frame of 1024 bins, the waveform is rectified and filtered by a trapezoidal band-pass filter (with the cut-off frequencies: $[0, 90, 350, 1000]Hz$). Then, the maximal value is picked from the autocorrelogram in the pitch domain within a window of $[1/350, 1/90]s$. To obtain R, the amplitude is normalized by the zero time-lag of the autocorrelation function. The probability $P(SNR > 0dB|R)$ is a function of the observable R which is established using a large statistic.

We added white noise at $0dB$ SNR to 288 sentences of the database, and we compiled a bi-dimensional histogram of the relationship existing between the local SNR value (in each 1024 bins time frame) and the harmonicity index. Giving a threshold at $0dB$, the mapping function having a sigmoidal shape is derived from this histogram, and it allows an estimate of the probability for the signal to be "clean enough". We name this probability *Voicing Index* which is derived from the harmonicity index R. First tests of the use of the Voicing Index for the fusion in audio-visual speech recognition are reported in [8].

4. Comparing the Criteria

For the comparison we used white noise, noise recorded in a car at 120 km/h and, from the NOISEX database, babble noise and two types of factory noise. We mixed these different types of noise to the audio signal at 12 SNR levels ranging from $-12dB$ to clean speech. In order to have a reference for the automatic fusion results we determined the minimum WER and the corresponding setting of the fusion parameter c for all noise types at all 12 SNR values. These optimum values were determined by adapting the fusion parameter manually (one example for the

[1] Varying noise levels in the video signal are not considered here

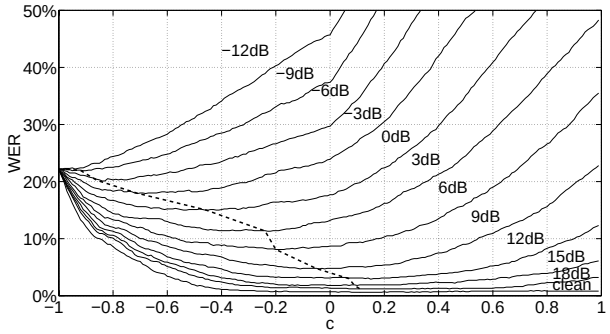

Figure 2: Relation between the fusion parameter c and the WER when adding white noise at 12 SNR levels ranging from -12dB to clean speech. The dashed line connects the points of minimum WER at a given SNR.

relation between the fusion parameter and the SNR is given in Fig.2). Next we evaluated the relation between the values of each of the three criteria and the optimal fusion parameter for a given noise type and SNR (the results can be seen in Fig. 3). In an optimization, the sigmoidal function which does *best* represent this relation between the criterion and the fusion parameter was searched (compare the dashed curve in Fig. 3). The word *best* stands for the function which minimizes the relative error

$$\mathrm{WER}_{rel}(\mathrm{SNR}, n) = \frac{\mathrm{WER}(\mathrm{SNR}, n) - \mathrm{WER}_{min}(\mathrm{SNR}, n)}{\mathrm{WER}_{min}(\mathrm{SNR}, n)} \tag{5}$$

over all types of noise n and SNR values. As optimization criterion the mean squared relative error over all noise types and all SNR values (altogether 60 values) is used:

$$e = \frac{1}{60} \sum_{(\mathrm{SNR}, n)} \mathrm{WER}_{rel}(\mathrm{SNR}, n)^2 \tag{6}$$

5. Results

First the relation between the three criteria and the fusion parameter c for a given noise type at varying SNR is visualized in Fig. 3. From these figures it can be seen, that the curves are the closest together when using entropy as a criterion and they are the furthest apart when using the Voicing Index. These observations suggest that the determination of the fusion parameter c depending on the entropy will give the best results.

In figures 4 the results of the fusion using the three criteria are visualized. The graphs show the minimum WER and the WER obtained using each of the criteria. When interpreting these graphs it has to be taken into account, that the optimization minimized the relative error at each SNR value and hence results in rather high absolute errors at high WERs. The Voicing Index leads to larger errors for white noise, and in addition to this, for babble noise at SNR values around $3-6$dB. The latter results are clearly due to the harmonic components present in the babble noise. The Voicing Index makes the assumption that the harmonic components stem mainly from the desired speech signal and the non-harmonic components from the noise. In the case of babble noise this is not fulfilled. Problems also occur when the background noise contains no harmonic components at all, as in the case of white noise.

For the comparison of the three criteria the relative error based on the minimum error (as defined in Eq. 5) also is helpful. Its numerical value averaged over all noise types and SNRs is given in Table 1. The averaged relative error shows that the

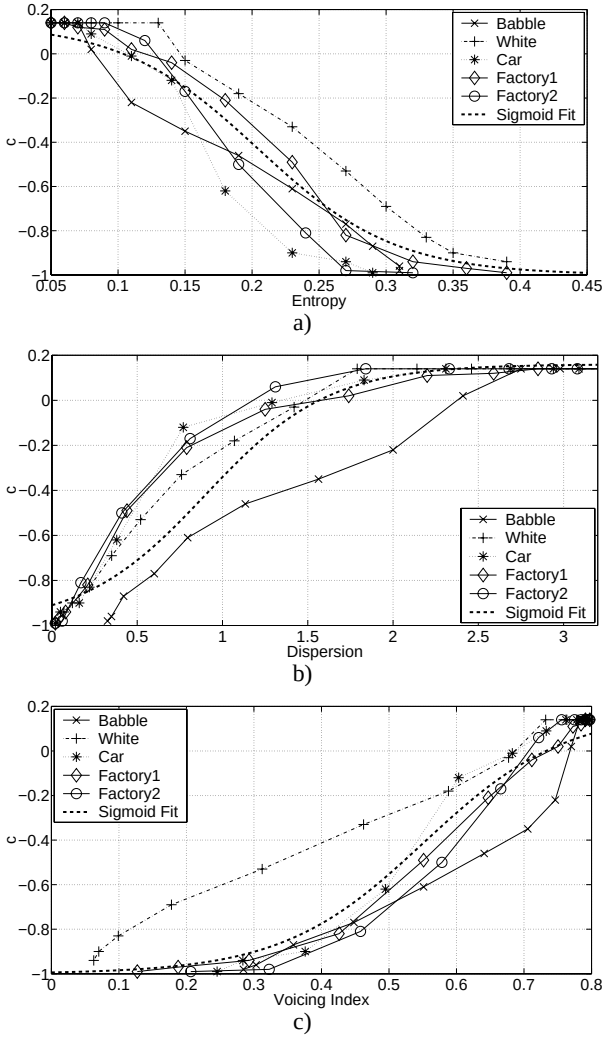

Figure 3: Relation between the three criteria and the fusion parameter c for five different noise types at varying SNR. In a) entropy in b) dispersion($N = 3$) and in c) the Voicing Index was used as criterion

entropy and the Voicing Index perform very similar, whereas the results obtained via the Dispersion are worse.

6. Discussion

From the comparison of the criteria it follows that the results given by the Voicing Index, and, to a smaller extend, those of the dispersion criterion, depend on the type of noise which is added to the speech signal. Both criteria show inconsistent results with white noise and babble noise. When looking at the results it also has to be taken into account that the calculation of the Voicing Index is only performed during speech activity and the detection of these speech segments so far is performed externally on clean speech. The two other criteria do not rely on an external speech pause detection. Hence the results of this comparison are expected to degrade in disfavor of the Voicing Index, when the speech segments are determined from the noisy signal.

In contrast to this, the entropy criterion has an almost homogeneous repartition of the errors over the different noise types and SNR values. These results differ from the findings reported

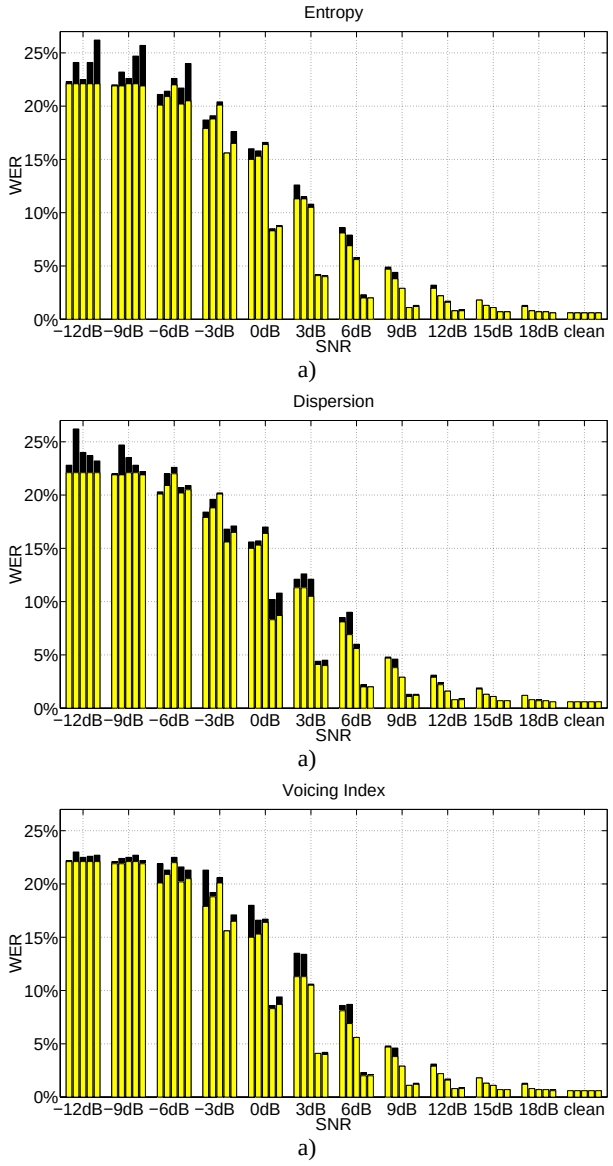

Figure 4: Minimum absolute WERs and absolute WERs resulting from the use of a) entropy b) dispersion($N = 3$) and c) Voicing Index as criterion for the fusion. The WERs are given for each noise type (from left: white, babble, factory1, factory2 and noise recorded in a car) at each SNR level. The WER when exploiting only the video channel is 22.1%.

in [6], where also entropy and dispersion are compared. They come to the conclusion that the dispersion is better suited to the automatic fusion than the entropy. Whereas it has to be mentioned that they only perform this comparison on clean speech with intent to perform a frame dependent fusion. When looking at our results it can be seen that the performance on clean speech and for good SNR is very similar for all criteria. The most significant differences occur at medium and low SNR values.

7. Conclusion

We compared two different approaches to the estimation of the reliability of the input channels to an audio-visual speech recognition system. One approach is based on the statistics of the a-posteriori probabilities as generated by the MLPs. We implemented the entropy and the dispersion of the probabilities, both following this approach. A second approach is to generate measures directly from the input data stream. We used the Voicing Index which depends on the ratio of the harmonic to the non-harmonic components in a speech signal. To get meaningful results we tested these three criteria on a variety of different noise types at a large SNR range. Our comparison showed that the entropy criterion slightly is better than the Voicing Index. The results obtained with the dispersion criterion are worse than those obtained with the two other criteria. In the case of the Voicing Index, as with other criteria based directly on the speech signal, problems arise when the type of noise does not fit into the model. Especially babble noise, which contains a lot of harmonic components from the background speech, and white noise, which has no harmonic components at all, cause problems for the use of the Voicing Index.

	Entropy	Dispersion	Voicing Index
relative error	4.45%	5.91%	4.81%
variance of rel. err.	0.0028	0.0048	0.0038

Table 1: Average and variance of the relative error for each criterion over all noise types and SNRs

8. Acknowledgments

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

9. 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] M. Heckmann, F. Berthommier, and K. Kroschel, "Optimal weighting of posteriors for audio-visual speech recognition," in *to appear in Proc. ICASSP 2001*, Salt Lake City, Utah, 2001.

[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*, Beijing, China, 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] G. Potamianos and C. Neti, "Stream confidence estimation for audio-visual speech recognition," in *Proc. ICSLP 2000*, Bejing, China, 2000, pp. 746–749.

[7] F. Berthommier and H. Glotin, "A new snr feature mapping for robust multistream speech recognition," in *Proc. ICPhS 1999*, San Fransisco, CA, 1999.

[8] H. Glotin, D. Vergyri, C. Neti, G. Potamianos, and J. Luetin, "Weighting schemes for audio-visual fusion in speech recognition," in *to appear in Proc. ICASSP 2001*, Salt Lake City, Utah, 2001.