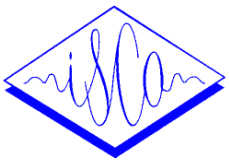

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

AUDIO-VISUAL SPEECH RECOGNITION COMPARED ACROSS TWO ARCHITECTURES

A. Adjoudani & C. Benoît
e-mail: adjouani@icp.grenet.fr

Institut de la Communication Parlée, Unité de Recherche Associée au CNRS N° 368
INPG/ENSERG - Université STENDHAL, BP 25X - F38040 Grenoble.

ABSTRACT

In this paper, we describe two architectures for combining automatic lip-reading and acoustic speech recognition. We propose a model which can improve the performances of an audio-visual speech recognizer in an isolated word and speaker dependent situation. This is achieved by using a hybrid system based on two HMMs trained respectively with auditory and visual data. Both architectures have been tested on degraded audio over a wide range of S/N ratios. The results of these experiments are presented and discussed.

10.21437/Eurospeech.1995-382

1. INTRODUCTION

Although acoustically-based automatic speech recognition systems have witnessed enormous developments over the past ten years, they still perform poorly in certain conditions, such as operation in noisy environments. However, using visual cues can significantly enhance the recognition rate [10, 12, 3].

Research with human subjects has shown that vision of the talker's face provides extensive benefit to speech recognition in difficult listening conditions [13, 4, 5, 2]. All those studies have shown that the audio-visual recognition scores are always higher than both the audio and the visual scores in all conditions, i.e., AV>A and AV>V. This is the basic challenge of bimodal integration, and thus the first goal any audio-visual ASR should reach.

2. AUDIO-VISUAL SPEECH PERCEPTION

In the area of speech perception, several models have been proposed to account for the human process of auditory and visual integration of speech [14, 11]. There have been proposed four or five different architectural structures to model the AV fusion in speech perception. Only two of these strategies are easy to implement. We have thus focused our studies on the comparison from results of these two kinds of architectures. They are hereby briefly presented:

1 - *Direct Identification Model*: it is based upon the "Lexical Access From Spectra" by Klatt [7]. The fusion process takes place before any classification. Thus, the input of such system is composed of a combined auditory and visual data. There is no common representation level over two modalities between the signal and the percept. Figure 1 shows the principle of this architecture.

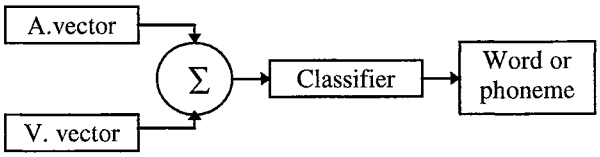

Figure 1: schematic of Direct Identification Model.

2 - *Separated Identification Model*: the A and the V inputs are directly identified through two parallel identification processes. Each input is matched against unimodal prototypes so that an A and a V score is computed from those A and V scores based on an automatic decision. Figure 2 summarizes this architecture.

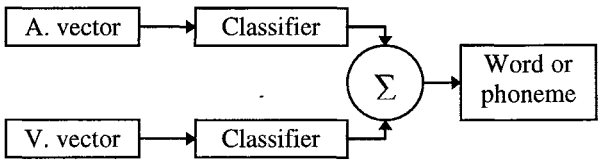

Figure 2: schematic of Separated Identification Model.

Fusion can be processed with logical data, such as in the VPAM model ("Vision Place Auditory Mode") where each modality deals with a set of phonetic features [14]. Thus, the place of articulation of the global output is that of the visual input and the mode of articulation (voiced, nasal, etc.) is that of the auditory input.

Outputs of both modalities can also be fused using probabilistic data (fuzzy logic). In this case, each pair of data corresponding to the same category is fused through probabilistic computation.

3. SYSTEM AND DATA BASE DEFINITION

The Hidden Markov Model (HMM) has been chosen as the main classifier in all our experiments. We have based our studies on isolated word recognition using a small vocabulary, i.e. 54 non-sense words uttered by a single speaker.

The V input consists of parameters carefully extracted from front and profile view of the speaker lips which have been made-up in blue [8]. A chroma-key system converts the lip region into the saturated black level in order to ease the edge detecting process. Figure 3 shows an example of the input image (in black & white).

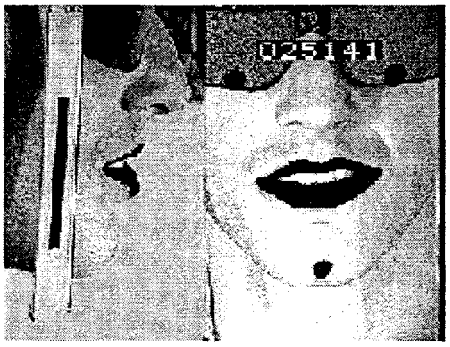

Figure 3: example of an input image for parameter extraction.

A dozen of parameters are stored every 20 ms along the speech waveform but only six of them have been retained as the input of the optical recognizer. This choice has two reasons:

i. Training the visual alone HMM with all front and profile parameters led to the high score of 87%. However, in order to get a tougher evaluation of the AV recognition system and better evaluate the benefit of the visual cues, we preferred to train the video HMM with a smaller number of parameters to obtain the more realistic video score of 78%.

ii. Profile parameters are usually difficult to extract and need additional equipments (extra camera, etc.). Generally, only the front view of the speaker is used as the main visual source of information. The six front parameters seemed thus more appropriate to an AV speech recognition.

The parameters used are as follow:

internal lip width (A)
external lip width (A')
internal lip height (B)
external lip height (B')
intero-labial lip area (S)
lip total area (S').

They are shown in figure 4.

For each visual vector, the first derivatives have been appended which makes a total of 12 coefficients for each set of visual data.

The audio input consists of 12 cepstral coefficients computed on a 40 ms Hamming window. First derivatives and energy are also appended. In order to have a complete synchronization between audio and visual information, the window displacement is set to 20 ms. Therefore, visual and audio vectors remain in synchrony.

The corpus is made of nine repetitions of 54 non-sense words of the form $V_1CV_2CV_1$ (V_1, $V_2 \in$ [a, i, y]; $C \in$ [b, v, z, ʒ, ʁ, l]) uttered by a French male speaker.

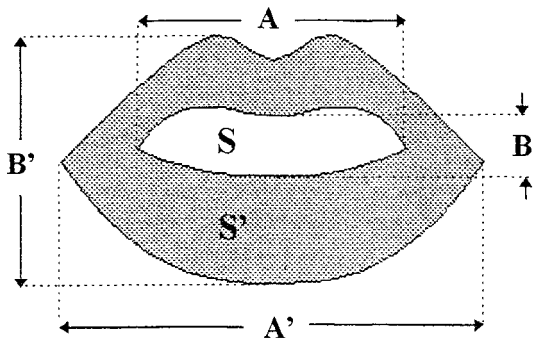

Figure 4: visual parameters used as the input of the optical recognizer.

This somewhat limited corpus is chosen because it has been extensively studied at the ICP in the visual modality, both through geometrical analyzes [1, 9] and perceptual tests [2, 6]. In all experiments, seven tokens of each word have been selected as the training set while the remaining two tokens were used as the test set.

In order to evaluate the performance of the system and the benefits of the visual information, the audio recognizer has been trained with clear acoustic data while tests have been run on noisy data at different S/N ratios of additive noise.

4. EXPERIMENTAL RESULTS

As a first step towards building an AV speech recognizer, we followed the architecture of the *Direct Identification Model*. In this experiment, visual parameters have been appended to audio data in order to build a single audio-visual vector. The results have been reported on figure 5. The V line shows the percentage of correct responses of the video recognizer alone. The A curve shows the scores of the audio alone recognizer. In order to have results independent of the test set and to take into account the variability among stimuli, audio and video HMM training and test set have been built up with four different combinations of test and training tokens. The dashed curves represent the average of the test scores while the line segments show the standard deviation for each test conditions. Since the V score is independent of the noise intensity, we have presented the V standard deviation as a shaded area. It is important to note that we have arbitrarily selected, as the reference for the V alone score, the test-train combination with the poorer V scores (plain horizontal line). The A and AV scores (plain curves) represent the scores obtained with the same test-train combination, i.e. first seven repetitions are used as training set and the last two repetitions as test set. In all other experiments, this test-train configuration has been used as the main reference.

The number of correct responses dramatically decreases as the intensity of the noise increases. The AV curve shows the result of the audio-visual recognizer with the mentioned fusion model.

As a second step, we have implemented the second fusion strategy, i.e. *Separated Identification Model.*

Different techniques may be applied to integrate the outputs of both modalities. Three of them have been tested:

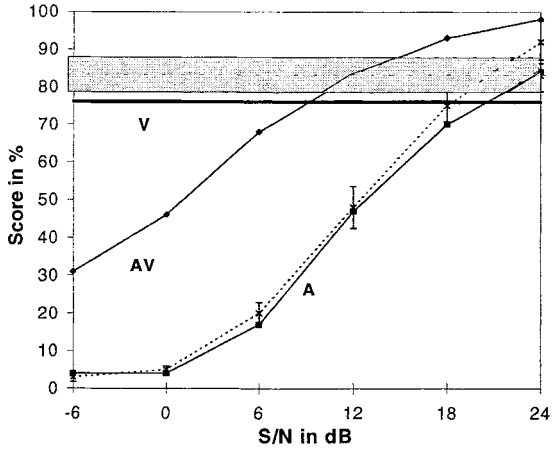

Figure 5: percentage of correct responses at the output of audio, video and audio-visual recognizer using the Direct Identification Model.

1- A simple way to integrate the two outputs consists of selecting one candidate among the two best candidates of each modality. It is obvious that in order to select the more reliable source, an additional information is needed to evaluate the quality of audio input. Thus, when the audio input signal is degraded, the system can switch to the video output and vice versa. This information can be extracted using the dispersion of output probabilities. Figure 6 shows the output probabilities of all models when the same word has been uttered under two different test conditions.

Many experiments have been run in order to estimate the noise level of the audio input signal using the probability dispersion. Finally, only the fourth highest probabilities have been used to evaluate the audio signal quality. In fact, using all probabilities has given poor results on the S/N estimation.

To have a dispersion reference, a similar operation has been run on the highest fourth video output probabilities.

Thus, for each word, the probability dispersion of each modality is calculated. The modality with the higher dispersion is selected and finally, the best candidate of this modality is chosen for the AV output.

The curve named S1 on figure 7 shows the results of this experiment.

2- Another approach consists of integrating the both outputs using probabilistic rules. A simple way is to select the candidate which maximizes the product of both output probabilities. The following formula can summarize this operation:

$$P_{AV}(C_{best}|W) = \max_i P_V(C_i|W) \times P_A(C_i|W) \qquad (1)$$

In which C_{best} represents the candidate which maximizes this product. $P_V(C_i|W)$ and $P_A(C_i|W)$ are respectively the output probability of the video and the audio HMM when the word W has been uttered.

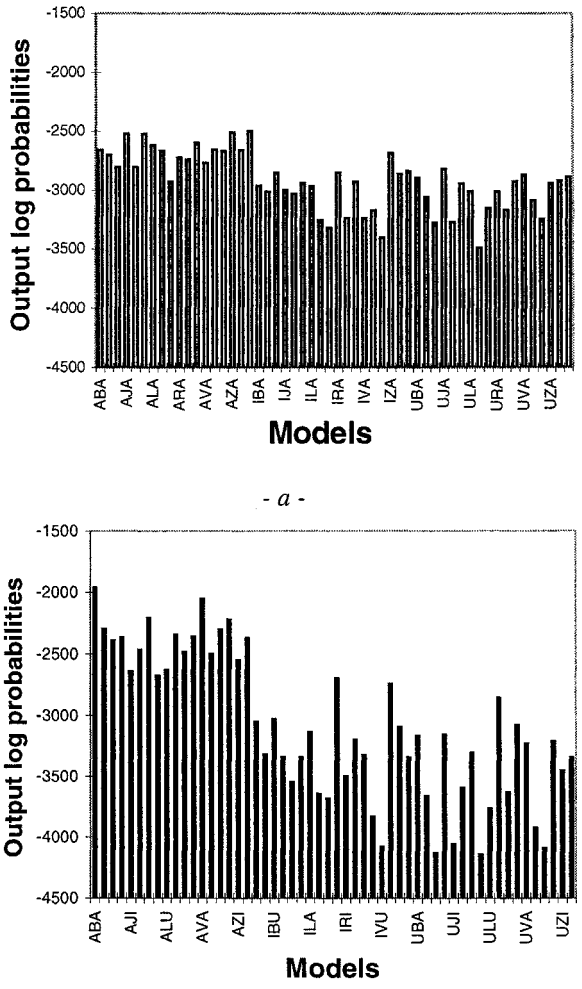

Figure 6: output probabilities of the word ABA from audio HMM in
a) 0 dB test condition,
b) 24 dB test condition.

Results are plotted as the S2 curve on figure 7.

3- In order to improve our probabilistic method especially in low S/N conditions, we have assigned a weighting factor to each modality. These coefficients are modulated by S/N, i.e., in low level of additive noise, more weight is assigned to the auditory output while in highly degraded acoustic input, video output has the major contribution on the final decision. Using equation 1 gives:

$$P_{AV}(C_{best}|W) = \max_i P_V(C_i|W)^{\lambda} \times P_A(C_i|W)^{(1-\lambda)} \qquad (2)$$

In which λ is the normalized weighting factor.

$$\lambda = \frac{\sigma_V}{\sigma_A + \sigma_V} \qquad (3)$$

In which σ_V and σ_A are respectively the dispersion of video and audio output probabilities calculated as explained before.

Results of this integration technique are plotted as the S3 curve in figure 7.

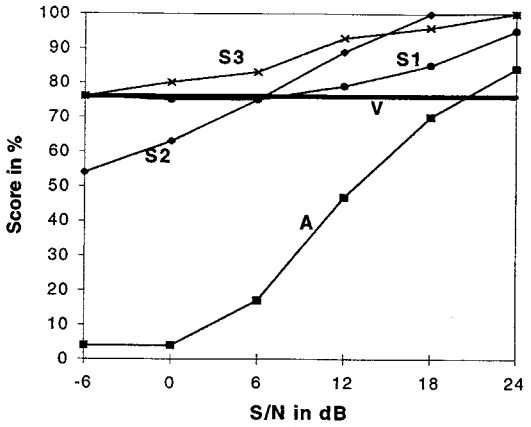

Figure 7: recognition scores of audio, video and audio-visual recognizer using the Separated Identification Model.

5. DISCUSSION

The *Direct Identification Model* must be rejected since our main challenge, i.e. AV>V and AV>A (figure 5) is not satisfied for all test conditions. However, AV scores are significantly higher than those of A alone. It is clear that this fusion technique is not optimal and visual cues can have better contribution in AV scores.

The *Separated Identification Model* offers better global scores according to the fusion strategy.

i. Results from the curve S1 (figure 7) satisfy our major constraint but are rather poor in high S/N. In fact, we expect better AV scores in high S/N test conditions since the two modalities are complementary.

ii. Results from the S2 curve show that this method offers an excellent contribution of visual data in high S/N conditions. However, when the audio input is highly degraded, the AV scores are lower than those in V mode.

iii. The S3 curve seems to be an intermediate level between the above mentioned techniques. First, it allows error corrections so the AV score can largely benefit of the integration of the two modalities. Second, when the acoustic input carries little information, the audio output decision is not reliable anymore and the system switches to the video HMM output. This prevents the global score from falling under the V alone score.

6. Conclusion

In order to better exploit the visual information in an audiovisual ASR, it is necessary to integrate both modalities in an optimal way.

Although the *Direct Identification Model* improves the audio recognition scores in noisy environments, it has shown its poor efficiency in highly degraded acoustic conditions. Our studies have shown that the *Separated Identification Model* can better take into account the bimodality of speech and hence can lead to a more efficient integration. The basic challenge "AV>V and AV>A" is reached only when a proper weight of A and V proportion is calculated.

References

[1] **Benoit**, C., **Mohammadi,** T., and **Abry**, C. (1992), "A set of French viseme for visual speech synthesis", *Talking Machines: Theories, Models and Design*, pp. 485-504.

[2] **Benoit,** C., **Mohammadi**, T., **Kandel,** S. (1994), "Audio-visual intelligibility of French speech in noise", *Journal of Speech Hearing Research .*

[3] **Bregler,** C., **Hild,** H., **Manke,** S., and **Waibel,** A. (1993), "Improving connected letter recognition by lipreading", *International joint conference of speech and signal processing, Minneapolis*, MN, Vol. 1, pp. 557-560.

[4] **Erber**, N.P. (1969), "Interaction of audition and vision in the recognition of oral speech stimuli", *Journal of Speech & Hearing Research*, 12, pp. 423-425.

[5] **Erber,** N.P. (1975), "Auditory-visual perception of speech", *Journal of Speech & Hearing Disorders*, 40, pp. 481-492.

[6] **Le Goff,** B., **Guiard-Marigny**, T., **Cohen**, M., & **Benoit,** C. (1994), "Real-time analysis-synthesis and intelligibility of talking faces ". *Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis* , New Paltz, New York, USA, Sept. 1994

[7] **Klatt**, D.H. (1979), "Speech perception: A model of acoustic phonetic analysis and lexical access". *Journal of Phonetics*, 7, pp. 279-312.

[8] **Lallouache**, M.T. (1991), Un poste "Visage-parole" couleur. Acquisition et traitement automatique des contours des lèvres. *Thèse de Doctorat de l'Institut Nationale Polytechnique de Grenoble*, 214p.

[9] **Mohammadi**, T., (1992), Synthèse du texte de visages parlants : réalisation d'un prototype et mesure d'intelligibilité bimodale. *Thèse de Doctorat de l'Institut Nationale Polytechnique de Grenoble.*

[10] **Petajan**, E. (1984), "Automatic lipreading to enhance speech recognition", *PhD Dissertation*, University of Illinois at Urbana-Champain.

[11] **Robert-Ribes,** J. (1995), Modèles d'integration audio-visuelle de signaux linguistiques : de la perception humaine à la reconnaissance automatique des voyelles. *Thèse de Doctorat de l'Institut Nationale Polytechnique de Grenoble.*

[12] **Stork**, D., **Wolff**, G., and **Levine**, E. (1992), "Neural Network lipreading system for improved speech recognition", *International joint conference of neural networks*, Baltimore.

[13] **Sumby**, W.H., and **Pollack**, I. (1954), "Visual contribution to speech intelligibility in noise", *Journal of the Acoustical Society of America*, 26, pp. 212-215.

[14] **Summerfield**, Q. (1987), "Some preliminaries to a comprehensive account of audio-visual speech perception". *In B. Dodd & R. Campbell (Eds.), Hearing by eye: the psychology of lipreading.*