ASYNCHRONOUS INTEGRATION OF AUDIO AND VISUAL SOURCES IN BI-MODAL AUTOMATIC SPEECH RECOGNITION

Paul Deléglise, Alexandrina Rogozan and Mamoun Alissali
LIUM, University of Maine
Av. Olivier Messiaen, BP 535, 72017 Le Mans Cedex, France
Tel: +33 43.83.37.70; Fax: +33 43.83.33.66
e-mail: deleglise@lium.univ-lemans.fr

ABSTRACT

This paper presents our work on the integration of visual data in automatic speech recognition systems. We particularly aim at solving two problems:

- classifiation differences for the modeling of acoustic information (phonemes) and visual information (visemes);
- the phenomena of anticipation and retention of visemes on the corresponding phonemes.

We developed and tested three systems, each dealing with one or both problems and proposing a different integration strategy. The comparison of system performances show that some of the solutions we propose give satisfactory results, and suggest that further work on some others would lead to more performance improvement.

1 VISUAL SPEECH

1.1 Speech bi-modality

Psychological studies [14] have shown the bi-modal nature of speech communication; besides the acoustic signal, visual information (mostly lip shapes and movements) is involved in speech perception. Visible speech is especially effective in non-optimal communication conditions such as noise-degradation of the acoustic signal or because of the linguistic complexity of the message.

However, due to articulatory phonemena, conjoint automatic processing of the two modalities is not straight forward, because of:

- P1: the asynchronism between the visual and the acoustic information (retention and anticipation phenomena) [1];
- P2: classification differences for the modeling of acoustic information (phonemes) and visual information (visemes)[4];
- P3: differences in period between the acoustic vector (10 ms) and the visual vector (20 ms) due to acquisition conditions.

The work we present here aims at the development of an audio-visual speech recognition systems which deals with these three problems. It is focused on the elaboration of an optimal integration strategy.

1.2 Integration of visual source in ASR systems

Perceptual models were developed to explain the human behaviour on the above-mentionned phenomena [12], but, for the moment, these models can not be directly used to implement automatic audio-visual speech recognition systems.

Existing systems implement various integration thechniques; acoustic and visual information may be merged at system input [9], or results of two separate identification components (acoustic and visual) may be combined at system output [5, 2]. Some systems have more complex architectures, such as the master-slave scheme [8].

However, none of these systems deal with all three problems. We developed, implemented and

tested three different integration schemes. In a stepwise process, with each new system we incorporated processing of a new problem, reaching thus to a system that deals with all three problems.

2 ASYNCHRONOUS INTEGRATION OF SOURCES

2.1 Experiments framework

All of the systems we present here are based on Continuous Hidden Markov Models (CHMM). Unit HMM correspond to the recognition unit which is the phoneme or the viseme (the visual equivalent, as will be explained hereafter), according to the system architecture. Each unit model is composed of 3 states plus an initial and a final non-active states, and implements a duration model [13]. Word models are constructed by concatenation, with border adjustement, of unit models.

Learning and test data is extracted from a corpus of synchronized audio-visual recordings of a test person pronouncing continuous random four-letter sequences.

Acoustic signal is analysed at a frequency of 100 Hz, each sample is represented by 12 Mel-Frequency Cepstral Coefficients (MFCC) plus the total energy of the analysis window. Acoustic observations (ASR inputs) are composed of these vectors and their first and second derivatives.

In order to simplify image analysis and to obtain precise measurements, video signal is recorded under special conditions [10] at 50 images per second. The analysis process extracts a set of parameters which describe the shape of lips. Only three parmeters representing the internal lip shape were used in the experiments we present here. These parameters are : horizontal and vertical opening and opening surface.

2.2 System architectures

As a reference and a base-line system, we first developed an acoustic-only system, which we designate by S0.

2.2.1 Direct integration

The first audio-visual system (S1) is based on the direct integration model. It has the same architecture as S0, but its input is obtained by merging the visual and the acoustic parameters. Merging the two types of parameters raises the acquisition-periods difference problem (P3), which we solve, without loss of information on the acoustic level, by interpolating the visual parameters with a spline-under-tension function [6]. The vector we thus obtain is then split into two streams (visual and acoustic), the weights of which are adjusted according to the noise level.

2.2.2 Asynchronous integration

The direct integration model forces the same frame (intra-HMM) and label (inter-HMM) boundaries for acoustic and visual data, which makes it impossible to deal with anticipation/retention phenomena (P1).

In order to fully deal with these phenomena, these temporal constraints must be loosened on the inter-HMM boundaries. This is done in our second system, S2 (cf. figure 1), by processing audio and visual inputs in two separate communicating components, which, at the same time, avoids the period difference problem and eliminates the need for interpolation.

The first component is identical to S1, but it uses an N-Best decoding algorithm [3]. The best N solutions are converted to a decoding net with time constraints. These latter determine intervals within which the second component, a purely-visual CHMM, may vary the phoneme boundaries. For each solution proposed by the first component, a new (visual) score is computed by the second. The best solution is determined, by a decision function which is a linear combination of the two scores.

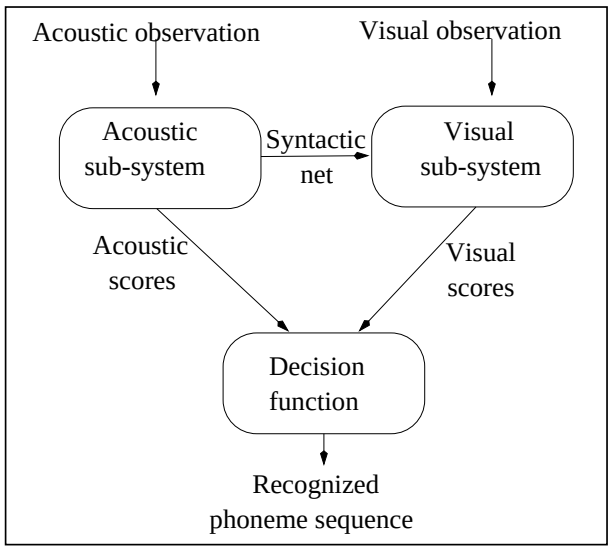

Figure 1: Architecture of the asynchronous integration systems

In the actual implementation, the coefficients of this function (the weighing factor) are determined empirically, according to the noise level, by selecting the value that seems to give to best results. For example, the weight of the visual component is 0.33 at -10dB and 0.22 without additional noise.

2.2.3 Use of visual-specific recognition units

The S2 system does not take into consideration classification differences (P2) since the two components use the same recognition unit, the phoneme. The importance of the use of an appropriate classification may be clearly seen by comparing auditory and visual confusion trees [7].

The third system (S3) uses an appropriate recognition unit for the visual component: the viseme (visual phonemes). Generally, visemes are defined as destinctive units of lip-jaw shapes and movements. In our case, visemes are used to label the visual data, which is a reduced representation since only internal lip openning measures are used. But this representation is still destinctive because of correlations between internal and external lip shapes.

The general architecture of S3 is identical to S2, except that, after N-Best decoding, the phoneme sequence hypotheses are mapped into viseme sequences for the visual CHMM entries. The decision on the recognized phoneme sequence is taken in a way similar to S2.

3 EXPERIMENTS

3.1 Test task and results

The four systems were experimented on the same task: recognition of connected letters in French. Utterances are synchronized audio-visual data of a test person pronouncing random four-letter sequences. The visual parameters are width, height and area of lip-opening, all measured on the internal lip boundary [10]. The corpus (described in 2.1) was realized at the ICP-Grenoble, it is composed of 200 utterances, of which two thirds were used for learning and one third for test. The acoustic signal is degraded with a dining-hall noise at a SNR of 10, 0 and -10 dB.

For each system and SNR, we obtain the following results, expressed in word accuracy (words correct minus insertion errors).

SNR System	clean	10Db	0Db	-10 dB
S0	90.85%	85.50%	62.32%	-44.37%
S1	95.42%	88.38%	75.00%	39.76%
S2	96.13%	89.08%	77.46%	44.36%
S3	95.77%	90.85%	79.23%	42.25%

3.2 Discussion

The results shown in this table correspond to the empirically-obtained optimal modality weights, i.e. the weights of the acoustic and visual modalities which give the best performances. As explained earlier, these weights are taken into consideration at system input in S1 and at system output in S2 and S3.

These results confirm the general hypothesis about the additive information included in lip shapes, especially in noisy environments. For example the weight of the visual stream in S1 is 0.5 at -10dB and 0.3 in clean conditions.

However, acoustic-only systems perform generally better than audio-visual systems in the special case of clean conditions, because of the additional noise brought in by the visual information. Our audio-visual systems perform always better than the acoustic-only system due to the variations of the weighing factors according to the noise level.

This is particularly true in the case of S1, where interpolation and weighing allow to integrate visual information without loss of information on the acoustic level.

Interpolation is avoided in the two other audio-visual systems, since data fusion takes place at system output. Both systems show better performances than S1, due to the fact that they take into consideration the asynchronism between the two modalities.

Use of an appropriate classification for visual data should yield more improvement. Unexpectedly, this is not always true, since S2 shows better performances than S3 in some cases (clean and -10dB). However, the differences not being statistically significant, these results are yet to be confirmed on more important corpus.

4 CONCLUSION

The results we obtain prove the importance of the chosen method when integrating of visual data in an ASR. The approach we present here consists of incrementally building a system that deals with the various problems.

The solutions we propose benifit from perceptive studies, but do not necessarily implement perceptive models. In particular we take into consideration the asynchronism and show that dealing with it improves system performances under all test conditions. The integration method used here (syntactic and temporal constraints and rescoring) is derived from negotiating multi-agent systems [11].

The preliminary results are promising. Regarding the unexpected results (cf. 3), in addition to what was here-above, they may be explained by the fact that our viseme set is not well suited for the task, since viseme classes are constructed on the symbolic level by grouping phoneme classes. Moreover, since no much work was done on visemes for speech recognition, this grouping was done on the basis of perceptive studies. We are actually working on corpus construction and recognition-oriented viseme sets.

Further work is also to be done on the decision function. Particularly we are working on automatic learning of such a function, possible approaches are probabilistic (maximisation of likelihood) and neural networks (recognition-error minimisation).

References

[1] C. Abry and T. Lallouache. Audibility and stability of articulatory movements, deciphering two experiments on anticipatory rounding in french. *Proc. of the 12th ICPS*, 1:220–225, 1991.

[2] A. Adjouani and C. Benoît. Audio-visual speech recognition compared across two architectures. *Eurospeech'95*, pages 1563–1566, 1995.

[3] S. Austin, R. Schwartz, and P. Placeway. The forward-backward search algorithm. *Proc. IEEE Int. Conf. on Acoustic, Speech and Signal Processing, Minneapolis*, pages 697–700, 1993.

[4] C. Benoît, T. Mohamadi, and S. Kandel. Effects of phonetic context on audio-visual intelligibility of french. *Journal of Speech and Hearing Research*, 37:1195–1203, October 1994.

[5] C. Bregler, H. Hild, S. Manke, and A. Waibel. Improving connected letter recognition by lipreading. *Proc. IEEE Int. Conf. on Acoustic, Speech and Signal Processing, Minneapolis*, 1:557–560, 1993.

[6] A. Cline. Scalar and planar valuated curve fitting using splines under tension. *Communications of the ACM*, 17(4):218–225, April 1974.

[7] G. Fisher. Confusions among visually perceived consonants. *Journal of Speech and Hearing Disorders*, 40:481–492, 1968.

[8] B. Jacob, R. Andre-Obrecht, N. Parlangeau, and C. Senac. Fusion des donnees acoustiques et articulatoires en reconnaissance de la parole. *15 colloque GRETSI*, 1:365–368, 1995.

[9] P. Jourlin, M. El-Bèze, and H. Méloni. Integrating visual and acoustic information in speech recognition system based on HMM. *ICPhS*, 4:288–291, 1995.

[10] T. Lallouache. *Un poste visage-parole : acquisition et traitement1automatique des contours labiaux*. PhD thesis, Institut National Polytechnique, Grenoble, France, 1991.

[11] S. Lander. *Distributed Search and Conflict Mangement among Reusable Heterogeneous Agents*. PhD thesis, University of Massachusettes, Departement of Computer Science, 1994.

[12] J. Robert-Ribes. *Modèles d'intgration audiovisuelle de signaux linguistiques*. PhD thesis, Institut National Polytechnique, Grenoble, France, 1995.

[13] N. Suaudeau and R. André-Obrecht. Sound duration modeling and time-variable speaking ratein a speech recognition system. *Proceedings of EuroSpeech'93*, pages 307–310, 1993.

[14] Q. Summerfield and A. MacLeod. Quantifying the contribution of vision to speech percepion in noise. *British Journal of Audiology*, 21, 1987.