ASYNCHRONOUS INTEGRATION OF VISUAL INFORMATION IN AN AUTOMATIC SPEECH RECOGNITION SYSTEM

Mamoun Alissali, Paul Deléglise and Alexandrina Rogozan

Laboratoire d'informatique de l'Université du Maine
Le Mans, France
E-mail: alissali@lium.univ-lemans.fr

ABSTRACT

This paper deals with the integration of visual data in automatic speech recognition systems. We first describe the framework of our research; the development of advanced multi-user multi-modal interfaces. Then we present audio-visual speech recognition problems in general, and the ones we are interested in, in particular. After a very brief discussion of existing systems, the major part of the paper describes the systems we developed according to two different approaches to the problem of integration of visual data in speech recognition systems.

Section 3 presents the architecture of our audio-only reference and baseline systems. Our audio-visual systems are described in Section 2. We first describe a system we developed according to the first approach (called the direct integration model) and show its limitations. Our approach, which we call *asynchronous integration*, is then presented in Sectio 4.2. After the general guidelines, we go into some details about the distributed architecture and the variant of the N-best algorithm we developed for the implementation of this approach.

In Section 6 the performances of these different systems are compared, then we conclude by a brief discussion of the performance improvements we obtain and future work.

1. INTRODUCTION

The work we present here is part of the AMIBE project, which aims at the development of advanced multi-user multi-modal (mainly audio and visual) interfaces [8]. In this project, in addition to Human-Computer communication purposes, multi-modal audio and visual information are used to identify and continuously check the identity of the user.

In this project, the complementarity of the two modalities is used for:

- The identification and the continuous verification of user identity.
- The improvement of speech recognition system performance: classification errors may be reduced since they are modality-dependant. For example acoustic confusion between [b] and [d] may be easily solved with visual information.
- The improvement of the robustness of speech recognition: the two modalities are insensitive to each other's noise. This is particularly important in noisy environments and in real conditions where noise level is very variable.

As we work on audio-visual speech recognition, we are only concerned with the last two points. This paper deals with the integration of visual data in an automatic speech recognition system with varying test conditions.

In the next section we present the integration problems that we are interested in and how they are dealt with in existing systems. In section 3, we describe the general architecture of our recognition systems, via an acoustic-only baseline system, called S0. Then we develop the solutions we propose for the above-mentioned problems and, successively, their implementations in three different audio-visual systems; S1, S2 and S3. We particularly describe the N-best algorithm first used in S2, and the modifications done on this system to construct S3, which deals with all the problems we are interested in. The results as well as a brief discussion are presented in section 6. In the conclusions section we evaluate the present state of the art and future orientations.

2. INTEGRATION OF VISUAL INFORMATION

The integration of visual data in automatic speech recognition is not straight forward. Among the difficulties described in various works (see for example [10]), we try to deal with the following:

- temporal shift, due to complex articulatory phenomena, between visual and acoustic information [3];
- classification differences for the modeling of acoustic information and visual information;

- differences in sampling periods between the acoustic and the visual vectors due to acquisition conditions.

Existing systems implement various integration techniques; acoustic and visual information may be merged at system input [6]. Section 4.1 presents such a system and discusses the limitations of this approach. Another solution is to use two separate identification components (acoustic and visual) and to combine their results at system output. The solution we propose is based on this approach, but contrarily to published works, e.g. [1, 4], our systems deal with abovementioned phenomena of temporal shift between audio and visual sources.

3. BASELINE ARCHITECTURE

Our baseline system (S0) is a classic Continuous-Hidden-Markov-Models (CHMM) based system (for a detailed description of HMM see for example [7]. Unit models represent french phonemes. They are connected in a network representing lexico-syntactic rules.

The observations (system input) for S0 are vectors composed as follows. First, the analysis of the acoustic signal produces a parametric vector every 100 ms. Each vector is composed of 13 parameters: 12 MFCC and the total energy of the analysis window. An observation vector is obtained by the concatenation of a parametric representation vector and its first and second derivatives.

S0 is used as a base for the development of the audio-visual systems, and as a reference for system performance comparisons. In particular, learning in all systems is based on the Baum-Welch algorithm. While solution searching varies according to the system: in S0 and S1, we use the Viterbi algorithm. In S2 and S3 this same algorithm is used as a cost-evaluation function for the A^* algorithm, as explained in Section 5.

4. AUDIO-VISUAL SPEECH RECOGNITION SYSTEMS

4.1. Direct Integration

The first audio-visual system (S1) is constructed as an extension of S0, according to the direct integration model. In this model, hybrid data is merged at system input and the system deals with its input as if it were of the same nature. In our case, visual data is composed of three parameters which represent internal lip shape; height, width and area. These parameters are obtained by image processing each 20 ms. But, as we explained earlier (cf. 3), acoustic vectors are obtained each 100 ms, hence the sampling periods problem, which we solve by interpolating the visual parameters with a spline-under-tension function [5]. The observation vector is then obtained, as for S0, by the concatenation of the audio-visual data vector with its first and second derivatives.

Since the importance of visual data depends on the noise level, the obtained audio-visual vector is split into two streams, audio and visual, the weights of which are adjusted according to noise level.

In addition to introducing the sampling periods difference problem, this architecture does not allow dealing with the articulatory phenomena and the classification differences. In the next section we first describe the S2 system which implements a solution to the first problem, then the S3 system which, in addition, deals with classification differences.

4.2. Asynchronous Integration

Studies on articulatory phenomena show that audio and visual modalities are not perfectly synchronous. Articulator movements may start before or after the beginning of sound production. These are known as, respectively, anticipation and retention phenomena. It is then necessary for the audio-visual system to be able to deal with slightly shifted temporal borders between the recognition units used for the two modalities.

One way to do this is to perform separate identifications. This is the case for our second audio-visual system (S2), which is composed of two separate subsystems (cf. Figure 1). The first subsystem is similar to S1 except that it uses an N-Best decoding algorithm, which we will describe in the next section. The second subsystem is a visual-only CHMM, guided by syntactic networks with time constraints. These networks are built after the N solutions proposed by the first subsystem. The time constraints define intervals inside which the visual subsystem may place its time borders. In the last step, a linear decision function is applied on the results of both subsystems in order to determine the best solution.

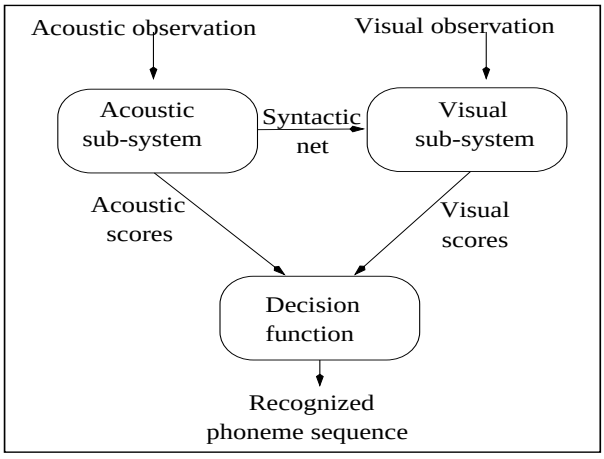

Figure 1: Architecture of the asynchronous integration systems.

For the remaining problem, classification differences, we constructed a third system (S3). It is identical to S2, but the vi-

sual subsystem uses visemes (classes of lip shapes and movements), instead of phonemes, as recognition units.

5. RESEARCH OF THE N BEST PHONETICALLY-DIFFERENT SOLUTIONS

This section described the algorithm used in the separate identification systems presented in the previous section. Derived from works such as [2], it extracts the N best phonetically-different solutions in an HMM-based speech recognition system. It is used in the two systems described in the previous section.

The Viterbi algorithm, used in S0 and S1, is time-synchronous in that it maintains the best solution for each state at each instant. In order to generalize this algorithm to obtain the N-best solutions it is sufficient to memorize the desired number of solutions at each state. But in this case the final solutions may be phonetically identical and differ only in their time alignment. A solution to this problem is to perform a backward research with an A^* algorithm.

A^* is a generic algorithm, which allows to obtain the best path (the path with the lowest cost) in a graph (see [9] for details). For a given problem, it requires the definition of:

- the graph, i.e. the nodes and their connections;
- an admissible cost-evaluation function;
- the nodes to be developed amongst a list (called the OPEN list) of all possible solutions. A choice is necessary to avoid combinatory explosion and to optimize research.

In our case, the nodes of the graph are partial solutions, i.e. sequences of (state, instant) pairs, where the states are those of all used Markov models. Their connections are defined by the lexico-syntactic constraints. Viterbi estimations are used for cost evaluation and nodes are selected according to a somehow complex strategy which we explain in Section 5.2.

5.1. The Cost Evaluation Function

As a cost-evaluation function we use the probabilities calculated and stored for each state and each instant by a forward pass of a time-synchronous algorithm: a first-order token-passing variant of the Viterbi algorithm. This function is at least admissible since the Viterbi algorithm finds the best (most probable) solution.

5.2. Node Selection

In order to avoid combinatory explosion, our algorithm selects the nodes to be developed according to the following criteria:

- Only final states of Markov models are considered in the construcution of the research graph.
- Only nodes whose cost is inferior to some threshold are developed. The threshold is determined proportionally to the final cost.
- We are only interested in phonetically-different solutions, similar partial solutions are discarded, except the first which is, in the same time, the one with the lowest cost.

Considering only final states yields less storage but makes it necessary to explicitly handle information about phoneme durations, which can not be done by the A^* algorithm since there is no cost evaluation during intra-model node expansion. The next section discusses two solutions we experimented to this problem.

5.3. Duration Modeling

In order to include information about phoneme durations we first implemented a per-frame heuristic cost function which is computed "on the fly". For each node expansion the cost is minimal for some well-adapted segment length, and is incremented beyond that length. The expansion is abandoned if the total cost exceeds a threshold, computed as a linear function of the minimal cost. Unfortunately, the heuristic cost function was observed to be well adapted only in some cases, especially when the tested Markov model corresponds effectively to the processed segment.

The second solution, the one we carried on, is to include a duration model. Duration modeling has been shown to improve the performance of HMM-based recognition systems [11]. In particular, such a model is essential for visual speech recognition because of the small amount of information.

The general idea is to assign to each HMM a duration probability function, whose parameters are extracted statistically from the corpus. This probability is taken into consideration when computing the scores in the backward pass.

When testing the S2 and S3 systems, we observed that in some cases the best solution is not selected by the backward pass, especially in the visual subsystem. This is due to the fact that time constraints are included only in this pass, which sometimes yields large differences between the scores assigned by the two passes to the same solution.

We solved this problem by modifiying the Viterbi algorithm to take duration constraints into consideration. This modification improves the performance of the algorithm, since it is certain to extract the best solution and to make better global estimations in the forward pass.

We notice that durations comparisons can be performed only on final states of Makov models. The choices made here are thus perfectly coherent with those done for node selection (cf.Section 5.2).

6. TESTS

The four systems were tested, under various noise conditions, on the same task : sequences of four french letters pronounced by a native speaker. Two thirds of the corpus, composed of 200 utterances, were used for learning and one third for tests. The acoustic signal is artificially degraded with a dining-hall noise at SNRs of 10, 0 and -10 dB. System performances are shown in Table 1 and plotted in Figure 2.

SNR System	clean %	10 dB %	0 dB %	-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

Table 1: System performances.

In general, these results confirm that including visual information improves the performance and the robustness of automatic speech recognition systems. They also show that separate identification and asynchronous integration is more promising than direct integration, since S2 and S3 performe better than S1.

However, according to our hypotheses, S3 should have yielded better results than S2, because it also deals with classification differences. This is only partially true (for 10 and 0 dB SNRs). This unexpected result may be attributed to an inappropriate choice of the viseme set we used. Also, a more complex decision making function may be required.

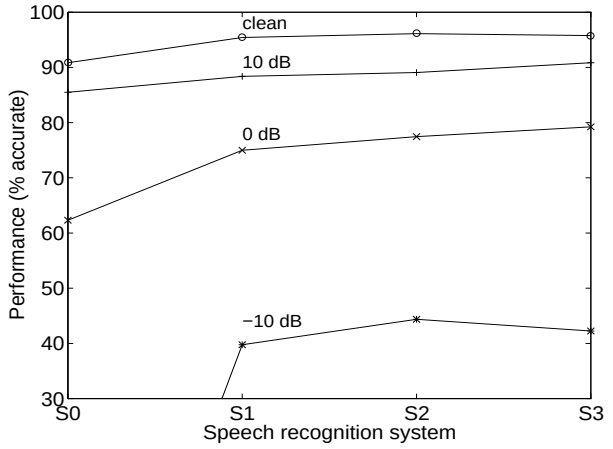

Figure 2: System performance progression.

7. CONCLUSIONS

In this paper we presented our work on audio-visual speech recognition. Compared to existing systems, our approach has the particularity of dealing with the time delays that can be observed in some cases between the visual and the acoustic aspects of speech production. In order to do so, the two systems we developed according to this approach combine a variety of techniques such as N-best algorithms and duration modeling. The obtained results are satisfactory, but they are still to be confirmed on more complex tasks.

Solutions to the problems mentioned in the previous section are being sought using more sophisticated probabilistic models and neural networks. We are also working on the definition and the realization of a relatively large audio-visual corpus.

8. REFERENCES

1. A. Adjouani and C. Benoît. Audio-visual speech recognition compared across two architectures. *Eurospeech'95*, pages 1563–1566, 1995.
2. S. Austin, R. Schwartz, and P. Placeway. The forward-backward search algorithm. *Proc. ICASSP*, pages 697–700, 1991.
3. 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.
4. 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.
5. A. Cline. Scalar and planar valuated curve fitting using splines under tension. *Communications of the ACM*, 17(4):218–225, April 1974.
6. 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.
7. Kai-Fu Lee. *Automatic Speech recognition. The Development of the SPHINX System.* Kluwer Academic Publishers, Boston, 1992.
8. C. Montacié, M.J. Caraty, R. André-Obrecht, L.J. Boë, P. Deléglise, M. El-Bèze, I. Herlin, P. Jourlin, T. Lallouache, B. Leroy, and H. Méloni. Applications multimodales pour interfaces et bornes évoluées. In H. Méloni, editor, *École Thématique "Fondements et Perspectives en Traitement Automatique de la Parole*, pages 155–164, Marseille, France, Juillet 1995.
9. J. Nillson. *Principles of Artificial Intelligence.* Tioga, 1980.
10. J. Robert-Ribes. *Modèles d'intgration audiovisuelle de signaux linguistiques.* PhD thesis, Institut National Polytechnique, Grenoble, France, 1995.
11. N. Suaudeau and R. André-Obrecht. Sound duration modeling and time-variable speaking rate in a speech recognition system. *Proceedings of EuroSpeech'93*, pages 307–310, 1993.