CONTINUOUS VISUAL SPEECH RECOGNITION USING GEOMETRIC LIP-SHAPE MODELS AND NEURAL NETWORKS

Alexandrina Rogozan and Paul Deléglise

Laboratoire d'Informatique de l'Université du Maine
Université du Maine, 72085 Le Mans Cedex 9, France
Tel: +33 02 43 83 38 64, Fax: +33 02 43 83 38 68
E-mail: Alexandrina.Foucault@lium.univ-lemans.fr

ABSTRACT

This paper describes a new approach for automatic speechreading. First, we use efficient, but effective representation of visible speech: a geometric lip-shape model.

Then we present an automatic objective method to merge phonemes that appear visually similar into visemes[1] for our speaker. In order to determine visemes, we trained SOM[2] using the Kohonen algorithm on each phoneme extracted from our visual database.

We go into the presentation of our visual speech recognition systems based on heuristics and neural networks (TDNN[3] or JNN[4]) trained to discriminate visual information. On a continuous spelling task, visual-alone recognition performance of about 37 % was achieved using the TDNN and about 33 % using the JNN one.

1. INTRODUCTION

Several researchers have demonstrated, through their models of visual or audio-visual recognition, the potential use of visual information (mostly lip shapes and movements) to improve the robustness and accuracy of speech recognition systems [1, 4, 5, 9].

Our own work in this area [2] was focused on the elaboration of an optimal integration strategy of audio and visual sources for automatic speech recognition. The results obtained show that separate identification and asynchronous integration (also called late integration) is more promising than direct integration. In order to improve its performance, we have to reinforce the purely visual identification by using:

- visual-specific recognition units: visemes. In fact, the phonemes are not suitable to label visual data because different sounds may be similar at the visual level;
- appropriate visual pre-processing and classification approaches. For example: whereas the parametrisation of acoustic data is well established, it is not the same for the visual data.

However, grouping of visually similar phonemes into viseme is not straightforward, because of:

- P1: coarticulation effects of adjacent sounds [3];
- P2: environmental effects (e.g. lighting);
- P3: articulatory differences among speakers [7].

Regarding visual speech modelling, works by [1, 8] have shown that a geometric model of the visible speech articulators gives satisfactory results for automatic speechreading.

This paper presents a variant of the previous geometric model of visible speech and a novel connectionist approach to determine visemes for our speaker. As concerns visual classification approach, we propose two purely-visual speech recognition systems using neural networks (TDNN and JNN) and heuristic rules.

We demonstrate the potential use of these systems on a connected word recognition problem. Without using any lexical, syntactic or acoustic rules, visual-alone recognition of 37 % is achieved using the TDNN and about 33 % using the JNN.

[1] Generally, the visemes are defined as distinctive units of lip-jaw shapes and movements.

[2] The Self-Organising Map was introduced by [6].

[3] The Time-Delay Neural Network was designed by [11].

[4] Work by [10] has proposed the Jordan partially recurrent Neural Network.

2. VISUAL PRE-PROCESSING TECHNIQUES

2.1 Parametrisation of Visual Data

Whereas the parametrisation of acoustic data is well established, it is not well known which visual features carry the most relevant speech information and which models of the visible speech are most suitable for automatic speechreading.

We decided to use a geometric lip-shape based model for visible speech. This choice was motivated on the one hand, by the fact that such a model is insensitive to some environmental effects, like lighting (P2), and on the other, because its configuration could be described by a small set of parameters. This model is build on previous researches [5, 8] and uses geometric measures on the internal lip shape of the speaker: height, width and area.

In addition to these static visual speech features (obtained by image processing each 20 ms), we investigate the dynamic of lip shape. For each feature, we compute the first derivative, its change between successive frames, and the second one. Each image frame is represented as a vector containing the values of these 9 visual features of which two thirds represent dynamic features and only one third represents static features.

Notice that most of the features kept pertain to the derivatives, according to our believe that the evolution of visual parameters is most significantly than their values.

2.2 Determining Visemes

The existence of articulator differences among various speakers (P3) affects the number of viseme categories and their respective constituents. By the way, the use of an automatic method to determine the viseme groups suitable for our speaker, becomes necessary.

We determine visemes from the training set of our audio-visual database. The acoustic sentences were phonetically transcribed and hand-segmented. Firstly, we use the projections of phonemic boundaries from acoustic signal on articulator signals to anchor fixed-size visual segments. In order to cover the visual realisation of any phoneme, each segment correspond to seven image frames.

While the evolution of a phoneme at visual level is analysed through 140 ms of signal, the phonetic transitions appear to be modelled with theirs respective phonemes. This is particularly true for the consonant phonemes. In this way, we take into account the coarticulation phenomenon (P1).

Consonant visemes	Vowel visemes
/p, b, m/	/a/
/f, v/	/e, i/
/s, z, t/	/u, o/
/d, k/	/Φ/
/j, ʃ/	/y/
/r, l/	
/g/	
/n/	

Table 1: Viseme classes

In order to facilitate the clustering of visually similar phonemes in a homogeneous space, we first separate them into consonants and vowels. Then Self-Organising Map, also called Kohonen feature map, (SOM) [6] is used to construct topology-preserving mapping of training data, where the location of a unit carries semantic information. In contrast to most other algorithms for neural networks, learning for SOM is unsupervised and by the way do not require additional knowledge.

The SOM was trained using the algorithm of Kohonen on each phoneme of segmented sentences. Visually similar 22 French phonemes are clustered into 13 visemes, yielded table 1. Most of visemes obtained for our speaker by computation appear to be consistent with those discussed by other lip-reading researchers such as names [5, 7]. That confirm, on the one hand that our clustering algorithm is appropriate for the phonemes grouping and on the other the pertinence of our visible speech parametrisation.

3. VISUAL SPEECH RECOGNITION SYSTEMS

3.1 System Description

Previous pre-processing furnish a feature vector description of phonologically important visual information. Thus, we model the visible by representing each sentence as a sequence of visual speech vectors.

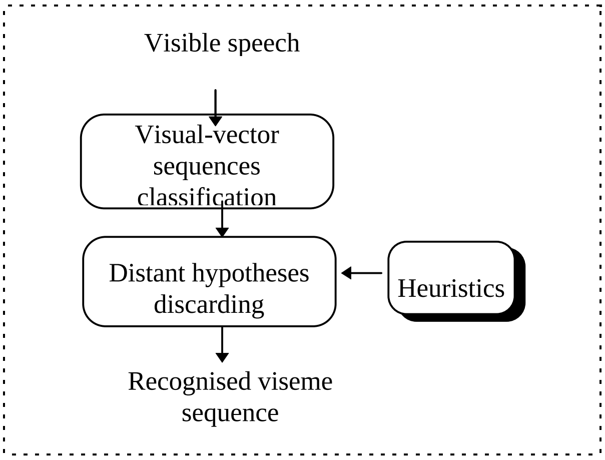
Visible speech
Visual-vector sequences classification
Distant hypotheses discarding
Heuristics
Recognised viseme sequence

Figure 1: Architecture of visual speech recognizer

Figure 1 depicts our visual speech recognition system based on neural networks and heuristics. The system learn to convert the input visual-vector sequence into a recognised sequence of visemes.

Classification of visual sequences has to deal with the inherent temporal dimension of visible speech. TDNN and JNN have been chosen as classifiers due to their capability to take into account time.

The classification of visual sequences uses no lexical or syntactic rules. In order to reduce recognition hypotheses, training sentences were used to extract heuristic rules including information about the global viseme duration and the per-frame score of each viseme-like state.

3.2 TDNN Based Classifier

The first purely-visual system is based on TDNN which seems to be very well suited for low-level viseme classification [11] and perform as time-shift invariant feature extractor.

Figure 2 shows the architecture of an TDNN and yields the manner in which it treat with the temporal dimension of speech: by introducing fixed delay on the input and hidden layers. On the input layer a fixed delay of 60 ms seems to be sufficient to represent low-level visual-phonetic events. The choice of a 150 ms delay for the hidden layer was made in order to represent a higher-level contextual visual event.

The backpropagation (thoughtlessly modified) algorithm was applied to train the network to fit viseme targets.

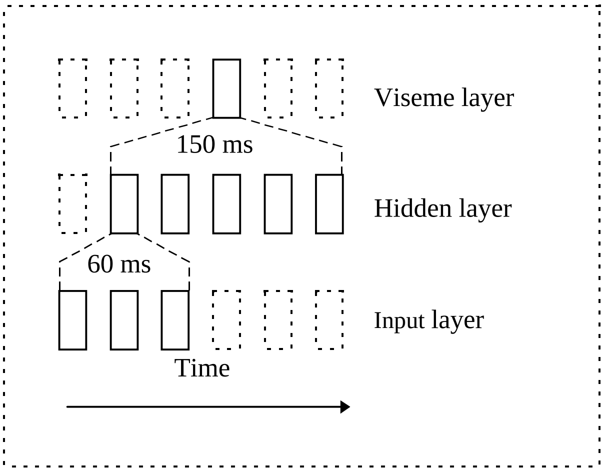

Figure 2: TDNN based classifier

3.3 JNN Based Classifier

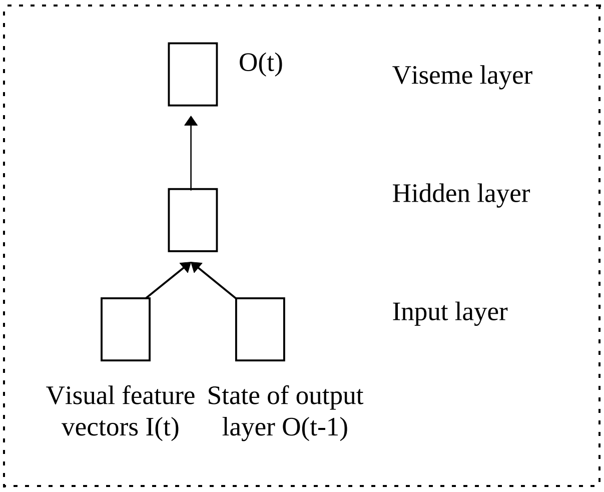

Figure 3: JNN based classifier

Figure 3 yields pre-processed visual data feeding the JNN inputs. Time is taken into account by the fact that the state of each neurone depends on actual visual input vector, but also on the previous state of output layer. The standard backpropagation algorithm for partial recurrent networks was used to train this neural network.

4. TEST TASK AND RESULTS

We experiment our system on French spelling task. Utterances are visual data of a test person pronouncing nonsense four-letter sequences without pauses. The task might be equivalent to continuous recognition with small, but highly confusing vocabulary.

Classifier	TDNN	JNN
without viseme boundaries	37 %	33,33 %
with known viseme boundaries	57 %	53 %

Table 2: System performances

The corpus realised at the ICP-Grenoble, it is composed of 200 utterances, of which one third was used as training data to set the weights of neural networks, one third for cross-validation and the last one for test.

As we yield table 2, our system achieved 37 % viseme accuracy using the TDNN based classifier and only about 33 % using the JNN one. This difference is due to the fact that the JNN seems to need more training data. These performances were obtained without the aid of acoustic, lexical or syntactic guides, and confirm the importance of visual cues in automatic speech perception. It should be noted that a lot of errors are caused by insertion and deletion. When we presented the visual sentences with known viseme boundaries, we came to visual accuracy of up to 57 % using the TDNN and 53 % using JNN.

5. CONCLUSION

In order to improve the accuracy of our previous audio-visual speech recognition system [2], we have to reinforce the purely-visual identification by using a viseme set suitable for our speaker. Thus, the use of an automatic method to determine the viseme groups becomes necessary. Then, a continuos visual speech recognition system is used to test the appropriateness of the obtained viseme set.

This paper presents a continuos visual speech recognition system based on geometric lip-shape models and neural networks. The preliminary results are promising, comparable to those obtained by [4] and [5] for equivalent recognition task, but not as good as the ones reported in [9]. One reason might be that last results are obtained for a less complex recognition task.

Our visual recognition system can be further refined by using an appropriate pre segmentation technique of visual data. We are also on the way to largely increase our corpus in order to improve neural network training. The visual system will contribute significantly to the achievement of robust and accuracy speech recognition at the time of integration with an acoustic one.

6. REFERENCES

[1] A. Adjouani and C. Benoît, « Audio-Visual Speech Recognition Compared Across Two Architectures », in *Proceedings of Eurospeech*, pages 1563-1566, Madrid, 1995.

[2] M. Alissali, P. Deléglise and A. Rogozan, « Asynchronous Integration of Visual Information in an Automatic Speech Recognition System », in *Proceedings of ICSLP*, pages 34-38, Philadelpy, 1996.

[3] C. Benoît, T. Mohamadi and S. Kandel, « Effects of Phonetic Context on Audio-Visual Intelligibility of French », in *Journal of Speech and Hearing Research*, Vol. 37, pages 1195-1203, 1994.

[4] P. Duchnowski, U. Meier and A. Waibel, « See me, Hear me: Integrating Automatic Speech Recognition and Lip-reading », in *Proceedings of ICSLP*, Philadelphy, 1994.

[5] J. Goldschen, « Continuous Automatic Speech Recognition by Lipreading », *Ph.D. Dissertation*, George Washington University, 1993.

[6] T. Kohonen, « Self-Organization and Associative Memory », in *Springer-Verlag*, 1988.

[7] P. B. Kricos, « Differences in Visual Intelligibility Across Talkers », in *Speechreading by Humans and Machines*, Stork & Hennecke eds., 150:43-55, 1996.

[8] T. Lallouache, « Un poste visage-parole : Acquisition et traitement des contours labiaux », *Actes des Journées d'Eudes sur la Parole*, 1990.

[9] J. Luettin, N. Thacker and S. Beet, « Speechreading Using Shape and Intensity Information », *in Proceedings of ICSLP*, pages 58-62, Philadelphy, 1996.

[10] M. I. Jordan, « Attractor dynamics and parallelism in a connectionist sequential machine », in *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*, pages 531-546, Hillsdale HJ, 1986.

[11] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano and K. Lang, « Phoneme Recognition Using Time Delay Neural Networks », in *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1989.