ESANN'1998 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 22-23-24 April 1998, D-Facto public., ISBN 2-9600049-8-1, pp. 377-382

Hybrid Hidden Markov Model/Neural Network Models for Speechreading

Alexandrina Rogozan and Paul Deléglise

Laboratoire d'Informatique de l'Université du Maine
Université du Maine, 72085 Le Mans Cedex 9, France
E-mail: Alexandrina.Foucault@lium.univ-lemans.fr

This paper describes a new approach for visual speech recognition (also called speechreading) using hybrid HMM/NN models.
First, we use the Self-Organising Map (SOM) to merge phonemes that appear visually similar into visemes[1]. Then we develop an hybrid speechreading system with two communicating components: HMM and NN, to take advantage from the qualities of both. The first component is a classical continuous HMM, while the second one is the Time Delay Neural Network (TDNN) or the Jordan partially recurrent Neural Network (JNN). At the beginning of the recognition process the HMM component segments and labels the visual data. In the case of visemes which are often confused by using the HMM, but rarely with the NN, we use the NN component to label the corresponding boundaries. For the other visemes, the final response is given by the HMM component.
Finally, we evaluate the hybrid system on a continuously spelling task and we show that it outperform an HMM system and a NN one.

1. Introduction

Several researchers have already assessed the benefit of incorporating visual information (mostly lip shapes and movements) into an automatic speech recognition system. They have demonstrated that an audio-visual system is more robust than an audio one, over a wide range of acoustic conditions.

Our own work in this area [8] was focused on the integration of audio and visual sources for automatic speech recognition. The results we obtained show that a late integration (also called separate identification) is more promising than an early one. In order to improve the performance of the late-integration based system, 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;

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

ESANN'1998 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 22-23-24 April 1998, D-Facto public., ISBN 2-9600049-8-1, pp. 377-382

• appropriate visual pre-processing and classification approaches. Indeed, it is not well known which visual features carry the most relevant phonologically information and which recognition algorithms are more suitable for automatic speechreading.

There are two different approaches to visual pre-processing: in the model-based approach, a set of visual features concerning mainly the lip contour are extracted [1, 5]; in the image-based approach the image is usually preprocessed by filtering followed by dimension reduction and then used as a feature vector [6, 10]. On the other hand, visible speech may be learned by using either an HMM trained to maximise the likelihood [1, 5, 6] or a NN trained to minimise the error rate [10].

In this paper we will outline a SOM-based method to determine visemes. We will also describe an hybrid HMM/NN speechreading system. First, we use HMM to segment and label the visual data. Then the NN component (TDNN or JNN) is guided by these boundaries and chooses between a confusable set of visemes. We suggest the potential use of this hybrid system on a connected letter recognition task.

2. Visible speech pre-processing

2.1 Parametrisation of visible speech

We decided to use a geometric lip-shape based model for visible speech because it is insensitive to some environmental effects and its configuration could be described by a small set of parameters. This model, build on previous researches [4], 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 by computing their first and second derivatives. Each image frame is represented as a vector containing the values of these 9 visual features, of which the most kept pertain to the derivatives, according to our belief that the evolution of visual parameters is more significantly than their values.

2.2 Viseme determining

The grouping of visually similar phonemes into viseme is not straightforward, because of the coarticulation effects of adjacent sounds and articulatory differences among speakers. The last one affects the number of viseme categories and their respective constituents. By the way, the use of an automatic method to determine visemes suitable to our speaker becomes necessary.

We determine visemes from the training set of our audio-visual database. As the acoustic sentences were phonetically transcribed and segmented, 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 phonemes appear to be mod-

ESANN'1998 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 22-23-24 April 1998, D-Facto public., ISBN 2-9600049-8-1, pp. 377-382

elled with theirs respective transitions. This is particularly true for the consonant phonemes. Through this way, we take into account the coarticulation phenomenon.

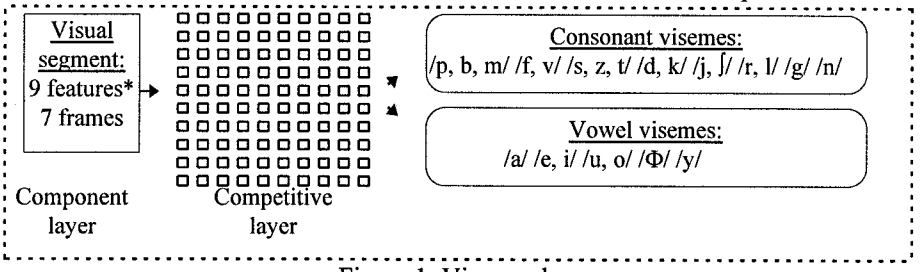

Figure 1: Viseme classes

In order to ease the clustering of visually similar phonemes in a homogeneous space, we first separate them into consonants and vowels. Then SOM [3] is used to construct topology-preserving mapping of training data, where the location of an unit carries semantic information. As learning for SOM is unsupervised, it does 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 figure 1. Most of visemes obtained for our speaker by computation appear to be consistent with those proposed by other lip-reading researchers. That confirms the appropriateness of this clustering algorithm for the phonemes grouping and the pertinence of previous visible speech parametrisation.

3. Hybrid architectures for speechreading

The speechreading system is composed of two communicating components and exploit the capacity of the HMM to treat continuous speech conjointly with the ability to discriminate visible information of the NN.

The first component is a classical continuous HMM based system. The viseme models, composed of three active states, are connected in a network representing lexico-syntactic rules. These models are learned with the Baum&Welch algorithm.

For the second component, as the classification of visual sequences has to deal with the temporal dimension of speech, the following NN have been chosen (cf. figure 2):

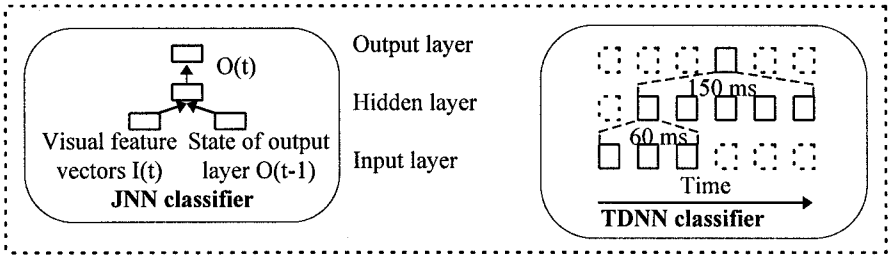

Figure 2: NN-based component

ESANN'1998 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 22-23-24 April 1998, D-Facto public., ISBN 2-9600049-8-1, pp. 377-382

• TDNN [9] treats with the temporal dimension of visible speech by introducing fixed delay: on the input layer 60 ms delay seems to be sufficient to represent low-level visual-phonetic event, whereas on the hidden layer 150 ms delay represents a higher-level contextual visual event;
• JNN [2] takes into account time through the fact that the state of each neurone depends on actual visual input vector, but also on the previous state of output layer.
These NN were trained to fit viseme targets with the backpropagation algorithm (for more details see [7]). After classification we use a time alignment algorithm to find the optimal path through the viseme-like state.

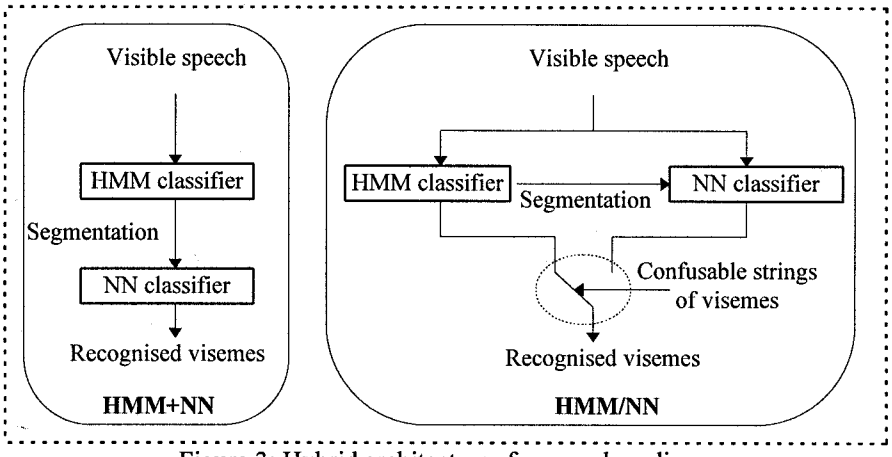

Figure 3: Hybrid architectures for speechreading

The first hybrid HMM+NN system uses the HMM component to segment the visual data and the NN component for classification purposes (cf. figure 3).

The second hybrid HMM/NN system, showed in figure 3, is based on the idea that if the recognition errors made by the two components are different, combining the recognition hypotheses would yield to performance improvement. The recognition process starts with the HMM decoding, which is supposed to be more reliable. The Viterbi algorithm is used to segment and label the visual data corresponding to the test set. Then we analyse the recognised viseme sequences. For the visemes which are often confused by using the HMM method, but rarely with the NN method, we furnish the corresponding boundaries to the NN classifier in order to remove the ambiguity. For the other visemes, the final response is given by the HMM based classifier. The confusable strings of visemes are previously find on the confusion matrix corresponding to the classification with the HMM or NN on the validation set.

4. Test task and results

We experiment the speechreading systems on a French spelling task. Utterances are visual data of a test person pronouncing nonsense four-letter sequences without

ESANN'1998 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 22-23-24 April 1998, D-Facto public., ISBN 2-9600049-8-1, pp. 377-382

pauses. The task might be equivalent to continuous recognition with small, but highly confusing vocabulary. The corpus realised at the ICP-Grenoble, is composed of 200 utterances, from which one third was used for training, another third for cross-validation and the last one for testing.

Accuracy	NN	HMM+NN	HMM/NN
TDNN	30.24 %	35.40 %	49.31 %
JNN	31.62 %	39.35 %	48.97 %

Table 1: System performances

Our system achieves 48.63 % viseme accuracy using the HMM and does not exceed 32 % using the NN. This difference is due to the fact that the NN require a large amount of data and our training set is not large enough for this purposes. On the other hand, there is not any efficient search technique to find the best scoring segmentation in continuous speech as for the HMM. It should be noted that a lot of errors are caused by insertion and deletion.

For the first hybrid HMM+NN system, we came to visual accuracy of 35.40 % using the TDNN and 39.35 % using the JNN (cf. Table 1). By the way, this system performs better than a baseline NN system, but worse than an HMM one. This result may be attributed to the fact that the label information furnished by the HMM component is discarded with such an architecture.

For the second hybrid HMM/NN system performance came up to 49.31 % using the TDNN and 48.97 % using the JNN (cf. Table 1). These results show performance improvement compared to an HMM based system and a NN based one. Even if it is not statistically important it suggests the appropriateness of hybrid models for speechreading. However these performances obtained without the aid of acoustic or syntactic guides are satisfactory if we take into account the fact that expert lipreaders will correctly recognise about 60 % of these nonsense isolated words.

5. Conclusion

In order to improve the accuracy of our previous audio-visual speech recognition system, we reinforced the purely-visual identification by using a viseme set suitable for our speaker and an appropriate classification technique.

The visemes we obtained by computation are coherent because most of them appear to be consistent with those proposed by other lip-reading researchers. On the other hand using these visemes as recognition units gives satisfactory results.

We described different classification approaches based on HMM, NN and hybrid HMM+NN or HMM/NN models for speechreading. Their performances are tested and compared on a connected letter recognition task. The comparison shows that the hybrid HMM/NN system is the most appropriate approach. Our results are satisfactory, comparable to those obtained by [1, 10] for an equivalent recognition task, but

ESANN'1998 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 22-23-24 April 1998, D-Facto public., ISBN 2-9600049-8-1, pp. 377-382

not as good as the ones reported in [5, 6]. One reason might be that the last results are obtained for a less complex recognition task.

Further work will concern the increase of our lip-reading database in order to improve neural network training. We are actually working on the integration of this hybrid speechreading system in an audio-visual system and the preliminary results show that it contributes significantly to the achievement of robust and accuracy speech recognition.

References

1. A. Goldschen, O. Garcia, E. Petajan : Rationale for Phoneme-Viseme Mapping and Feature Selection in Visual Speech Recognition. In Speechreading by Humans and Machines, 505-509 (1996)
2. M. I. Jordan : Attractor Dynamics and Parallelism in a Connectionist Sequential Machine. Proc. of the Eighth Annual Conference of the Cognitive Science Society, 531-546 (1986)
3. T. Kohonen : Self-Organization and Associative Memory. In Springer-Verlag (1988)
4. T. Lallouache : Un poste visage-parole : Acquisition et traitement des contours labiaux. Actes des Journées d'Eudes sur la Parole (1990)
5. J. Luettin, N. Thacker and S. Beet : Speechreading Using Shape and Intensity Information. Proc. of ICSLP, 58-62 (1996)
6. J.R. Movellan : Visual Speech Recognition with Stochastic Networks. Advances in Neural Information Processing Systems, vol. 7 (1995)
7. A. Rogozan, P. Deléglise : Continuos Visual Speech Recognition using Geometric Lip-Shape Models and Neural Networks. Proc. of Eurospeech, 1999-2003 (1997)
8. A. Rogozan, P. Deléglise, M. Alissali : Adaptive Determination of Audio and Visual Weights for Automatic Speech Recognition. Proc. of Audio and Visual Speech Processing Workshop, 61-65 (1997)
9. A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, K. Lang : Phoneme Recognition Using Time Delay Neural Networks. IEEE Transactions on Acoustics, Speech and Signal Processing (1989)
10. U. Meier, R. Stiefelhagen, J. Yang : Pre-processing of Visual Speech Under Real World Conditions. Proc. of Audio and Visual Speech Processing Workshop, 113-117 (1997)