Classifying Visemes for Automatic Lipreading

Michiel Visser[2], Mannes Poel[1], and Anton Nijholt[1]

[1] University of Twente, Department of Computer Science,
P.O. Box 217 NL-7600 AE Enschede, The Netherlands
{anijholt, mpoel}@cs.utwente.nl
[2] Philips Medical Systems, The Netherlands

Abstract. Automatic lipreading is automatic speech recognition that uses only visual information. The relevant data in a video signal is isolated and features are extracted from it. From a sequence of feature vectors, where every vector represents one video image, a sequence of higher level semantic elements is formed. These semantic elements are "visemes" the visual equivalent of "phonemes" The developed prototype uses a Time Delayed Neural Network to classify the visemes.

1 Introduction

Automatic Speech Recognition (ASR) normally uses acoustic information as input. Often this information is too unreliable to get good recognition. This unreliability, indicated by the signal-to-noise ratio, is caused by noise of machinery, other people speaking, effects in the measurement, etc. In these cases visual information can be used to improve recognition, because speech production produces several visual side effects. This is normally referred to as lipreading, but is also called speech reading, to clarify that this information is not all located in the lips. Other relevant features for lipreading can be extracted from the jaw, tongue, teeth and skin, see for instance [2] and [1].

In this paper we will study the problem of mapping lip movements, contained in a digital video of the face, onto relevant semantic features, called visemes.

Section 2 describes the stages in the recognition process theoretically, while Section 3 describes a prototype that implements this decomposition. This prototype is evaluated in Section 4. The conclusions can be found in section 5.

2 Decomposition of the Lipreading System

The lipreading system is decomposed in three subsystems, First we have the liplocalization system, which localizes the lips in the digital input, then there is the feature extracting system, which computes the relevant features of the lips, and finally we have the classification system, which maps feature vectors to visemes.

Lip localization. The lip localizator must find the position of the lips in the digital image. This is a difficult task, due to the demands on robustness, accurateness, and the system should be person-independent.

V. Matoušek et al. (Eds.): TSD '99, LNAI 1692, pp. 349–352, 1999.

Most lip localization techniques are template- or edge-based. The template approach has as one of the disadvantages the lack of speed. A lot of templates must be tested on a lot of areas before an accurate position of the position of the mouth can be found. Lips can also be found using edge detection. Standard edge finders can find edges of the lips using a hard-crafted or learned model of the lip edges. One of the main disadvantages of the lip-edge approach is the lack of robustness.

Instead of finding the lips directly, the position of the lips can be calculated using the positions of other objects in the face, like the nose and the eyes. Using the nose and the eyes, the position of the mouth area can be predicted to a certain extent and therefore only a small piece of a video image has to be tested for lips.

Feature extraction. After the position, size and rotation of the mouth area are determined a set of features must be extracted using a parameter set of a lip edge model, pixel map of the mouth area, or a compressed pixel map. Using a lip edge model has the advantage that it gives a small set of well-defined parameters, that are more invariant in lighting and personal characteristics than pixel based parameters. The disadvantage is that it looses a lot of information and accuracy.

The set of parameters changes dynamically in time. Therefore often dynamical features are added to the feature set.

Translation to a semantic representation. To get a semantic representation of the spoken words the time sequence of the feature set must be translated into a sequence of semantic symbols. As semantic symbols we use visemes. Visemes are classes of phonemes which are indistinguishable from a lipreading point of view, see Figure 1.

viseme	phoneme class	viseme	phoneme class
0	<silence>	8	I e:
1	f v w	9	E E:
2	s z	10	A
3	S Z	11	@
4	p b m	12	i
5	g k x n N r j	13	O Y y u 2: o: 9 9: O:
6	t d	14	a:
7	l		

Fig. 1. Visemes as phoneme classes, using SAMPA notation.

However, the utterance of a phoneme does not generate exactly the same lip position and movement all the time. Various factors, such as person, situation and mood, causes these variances in phoneme production.

3 A Prototype of the Lip Reading System

For the above system a prototype is constructed, following the decomposition above.

Lip localization. For the liplocalization we used a non-standard approach. From a set of digital lip images (generated by hand) the principal components were computed. These principal components were used to compress the fixed regions of the picture. Since the principal components were fine tuned for lips, the lips will be in the region with the minimum loss of information.

Feature extraction. The feature extractor has as input a normalized rectangle containing the lips. Principal component analysis is used to generate a feature vector for the rectangler input frame. The principal components are computed using a single layer feedforward neural network and Sanger's learning algorithm [4].

Translation to visemes. The feature vectors computed by the feature extractor must be classified as (the produced) visemes. In order to incorporate temporal aspects in the classification process a Time-delay Neural Network (TDNN) [3,5], was used to classify the feature vectors.

In order to train the TDNN, a dataset was manually generated. Three different people uttered, under identical conditions, a list of non-existing words, which precisely contained all the viseme combinations. One testset was made, which consists of the utterances of a list of real words. These datasets were manually analyzed, and every feature vector was labeled with the corresponding viseme. These labeled sets were used to train and test the TDNN.

4 Evaluation of the Prototype

The succes rate of the prototype is given in figure 2. As can be seen from this table the average succes rate is 0.226. In order to gain more insight in the mistakes

viseme number	average succes rate	viseme number	average succes rate
0	0.260	8	0
1	0	9	0.667
2	0.167	10	0
3	0.389	11	0.419
4	0	12	0
5	0	13	0
6	0.867	14	0
7	0.600		

Fig. 2. Succes rate of the lipreading prototype.

made by the prototype a confusion matrix was constructed, figure 3. As one can see there is a lot of confusion between the first 7 visemes.

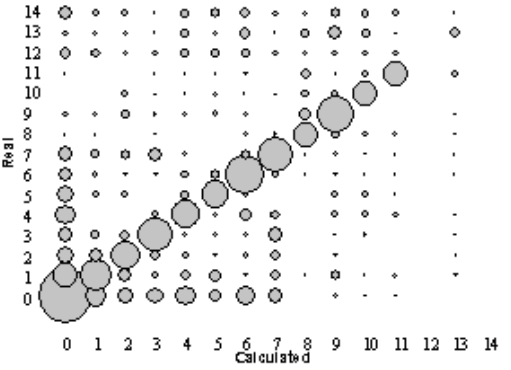

Fig. 3. Viseme confusion matrix

5 Conclusions

The prototype lacks speed and gives non-optimal results. The localization of the mouth is by far the slowest component in the prototype. More sophisticated techniques should be used to get a significant speedup. The following causes can be mentioned for the non-optimal classification results:

- The visemes, seen as a set of phonemes, are not correct, different visemes correspond to almost identical lip movements. An indication for this can be found in the confusion matrix in figure 3.
- Too little data: more instances of the trainings data should be generated and in the training sets natural words are preferable above synthetic ones.
- Differences in lighting: It turned out that differences in illuminations were larger than expected.

References

1. Benoît, Abry, Cathiard, Guiard-Marigny et all.: Read my lips: where? how? when? and so ... What? In: Proceedings of the 8th Int. Congress on Event perception and Action (1995).
2. Benoît, Guiard-Marigy, Le Goff, Adjoudani: Which components of the face do humans and machines best speechread? In: Stork, D. (eds.): Speechreading by Man and Machine. Springer-Verlag.
3. Lang, K.J., Hinton, G.E.: The development of the time-delay neural network architecture for speech recognition. Techical Report CMU-CS-88-152, Carnegie-Mellon University (1988).
4. Sanger, T.D.: Optimal unsupervised learning in a single-layer linear feedforward neural network. Neural Networks. **12** 459–473.
5. Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., Lang, K.J.: Phonemen recognition using time-delay neural networks. IEEE Transactions on Acoustic, Speech, and Signal Processing, **37** (1989) 328–339.