TOWARDS SPEAKER INDEPENDENT CONTINUOUS SPEECHREADING

Juergen Luettin

IDIAP
CP 592, 1920 Martigny, Switzerland
luettin@idiap.ch

ABSTRACT

This paper describes recent speechreading experiments for a speaker independent continuous digit recognition task. Visual feature extraction is performed by a lip tracker which recovers information about the lip shape and information about the grey-level intensity around the mouth. These features are used to train visual word models using continuous density HMMs. Results show that the method generalises well to new speakers and that the recognition rate is highly variable across digits as expected due to the high visual confusability of certain words.

1. INTRODUCTION

Current speechreading (or lipreading) systems have mainly been evaluated for small vocabulary, speaker dependent, isolated speech recognition tasks [8]. One of the main difficulties in speechreading, however, is to cope with the large appearance variability across subjects and to extract visual speech features which generalise well for new speakers. Appearance variability might for example be due to differences of lips, teeth, skin, facial hair, or due to different visual articulation. Additional variability can be introduced by different pose of the subject or by different lighting conditions. To be usable in real-world applications, a speechreading system should ideally be robust to all these factors.

These variabilities can cause severe problems in the extraction of visual speech features. Several researchers have therefore performed simplified speechreading experiments by painting the subject's lips with a reflective marker, by performing experiments on one subject only, or by using controlled recording environments. Most applications however require the image analysis to be performed on natural images and under different environmental conditions. This paper describes speechreading experiments where visual features are automatically extracted without the use of visual aids on a database of 37 subjects. Results are presented for a speaker independent continuous digit recognition task.

2. DATABASE

The M2VTS audio-visual database [7] was used for all experiments. It contains 185 recordings of 37 subjects (12 females and 25 males). Each recording contains the acoustic and the video signal of the continuously pronounced French digits from zero to nine. Five recordings have been taken of each speaker, at one week intervals to account for minor face changes like beards. For each person, the shot with the largest imperfection was labelled as shot 5. This shot differs from the others in face variation (head tilted, unshaved beards), voice variation (poor voice SNR) or shot imperfections (poor focus, different zoom factor). Additional imperfections apart from those of shot 5 are due to some people who were smiling while speaking. The database contains a total of over 27,000 colour images which were converted to grey-level images for the experiments reported here.

3. VISUAL FEATURE EXTRACTION

The method for visual feature extraction is based on a lip tracker which has been described in detail in [4, 6]. A point distribution model (PDM), also called active shape model (ASM) when used in image search [2], is used to model the shape of the lips. PDMs are flexible models which represent an object by a set of labelled points. The points describe the boundary or other significant locations of an object.

Shape deformation is modelled by decomposing a shape into a weighted sum of basis shapes using a Karhunen-Loéve (K-L) expansion. The basis shapes are obtained from the statistics of a representative training set using principal component analysis. This formulation of shape deformation constrains the shape model to only deform to shapes similar to the ones seen in the training set.

Similar to shape modelling, the texture around the mouth area is modelled by decomposing the intensities into a weighted sum of basis intensities using a K-L expansion. Intensity modelling fulfils two purposes: it is used as a mean for robust image representation for image search and for visual speech feature extraction. The basis vectors describe the in-

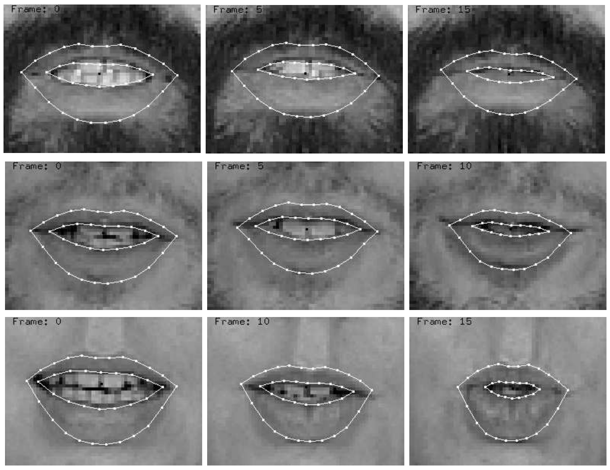

Figure 1: Examples of lip tracking results. The first row demonstrates the robustness of the algorithm for subjects with beard despite the reverse contrast between lips and skin.

tensities in the area around the shape points. The intensity space deforms with the shape of the model and therefore represents shape independent intensity information.

The intensity weight vector represents the grey-levels near the lip contour and accounts for features like the intensity of the oral cavity, visibility of teeth and tongue, and finer details like protrusion. These parameters therefore contain important visual speech information. Figure 1 shows some lip tracking examples. Several subjects in the database have a beard which makes lip tracking more difficult and which increases the inter speaker variability of extracted features.

Lip tracking is based on a distance measure between the lip model and the image and a minimisation function which finds a minimum of this distance over the model parameters. Visual features can be recovered from the tracking results and are represented by the normalised weights of the basis shapes and the basis intensities. Much visual speech information is contained in the dynamics of lip movements rather than the actual shape or intensity. Furthermore, dynamic information might be more robust to linguistic variability, i.e. intensity values of the lips and skin will remain fairly constant during speech, while intensity values of the mouth opening will vary during speech. On the other hand, intensity values of the lips and skin will vary between speakers, but temporal intensity changes might be similar for different speakers and robust to illumination. Similar comparisons can be made with shape parameters. First order differential parameters (delta parameters) of the shape and intensity vectors were therefore used as additional features.

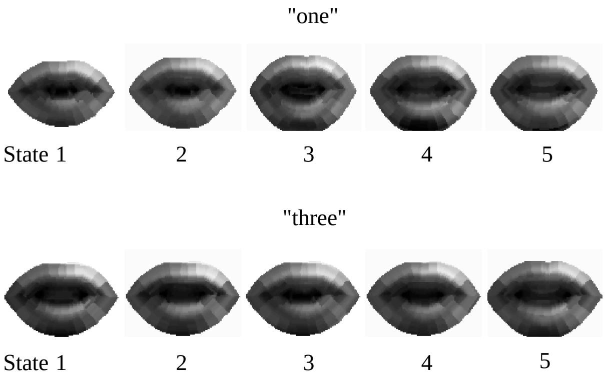

Figure 2: Sequence of HMM states for the words "one" and "two" learned from 11 subjects.

4. VISUAL SPEECH MODELLING

A visual observation of an utterance is represented by a sequence of visual feature vectors which are obtained from the lip tracker. The feature vector can consist of a combination of shape parameters, intensity parameters, and additional delta parameters. It is assumed that the feature vectors follow continuous probability distributions which are modelled by mixtures of Gaussian distributions. It is further assumed that temporal changes during speech are piece-wise stationary and follow a first-order Markov process.

As usually done in the case of small vocabulary ASR, whole word models were used, thus each word class was represented by one HMM. The HMMs only allowed self-loops and sequential transitions from the current to the next state. The models were trained using two training stages. In the first stage the models are trained using the word-segmented training data. Each HMM is initialised by linear segmentation of the training vectors onto the HMM states, followed by iterative Viterbi alignment and computation of the means and variances for each state. In the case of multiple mixtures, the vectors for each state are clustered by a modified K-Means algorithm. The models are further re-estimated based on ML estimation using the Baum-Welch procedure. The second stage consists in embedded training using the Baum-Welch algorithm on the whole training sentences. The word models are concatenated according to the transcriptions but no information about word boundaries is used.

Since no transcription of the speech data was available, the word boundaries of the training data were found by a HMM based speech recognition system which was used to segment and label the sentences [3]. The recogniser used the known sequence of digit word models, which were trained on the *Polyphone* database of IDIAP [1], and performed *forced alignment*. The training data of the visual features is therefore based on acoustic rather than visual segmentation. It is likely that acoustic and visual seg-

Table 1: Speaker independent recognition accuracy for different training and test procedures using HMMs with 8 states and 2 mixture components per state.

	Segmented Recognition	Continuous Recognition
Segmented Training	60.2%	51.3%
Embedded Training	58.7%	58.5%

mentation is not identical, e.g. the visual segmentation is more difficult to determine, and visual anticipation might precede the acoustic signal. The acoustic signal is however often more reliable than the visual signal, which favours the use of acoustic segmentation for visual speech recognition.

Figure 2 displays the visualised HMM states of the word models "one" and "two" trained on 11 speakers [5] on the Tulips1 database of isolated digits. Each state is visualised by synthesising a lip instance using the mean shape and mean intensity vector of that HMM state.

Recognition was performed based on the maximum posterior probability in which the prior probabilities for all word classes were assumed to be equal. The Viterbi algorithm was used to calculate the most likely state sequence.

5. EXPERIMENTS

Two recognition tasks were performed: one task was continuous word recognition on the whole sentences and the second task was defined as word recognition on the segmented sentences. Although the first task is continuous speech recognition, the sequence of the digits was always the same from "zero" to "neuf". The words were therefore always spoken in the same context, which usually simplifies continuous speech recognition. The second task can be considered as isolated word recognition but where the words were spoken continuously. It was mainly performed to obtain the recognition accuracy, given the acoustic segmentation. All experiments were performed for speaker independent tests on the first four shots of the database using the leave-one-out procedure. One experiment therefore consisted of 37 leave-one-out tests, each made up of 1440 training words and 40 test words. This resulted in a total of 1480 test words spoken by 37 subjects.

Visual features were obtained from lip tracking results and consisted of 14 shape parameters, 10 intensity parameters, scale, and temporal difference parameters. This resulted in a 50-dimensional feature vector. Different HMM architectures were investigated by varying the number of states (1 - 10) and the number of mixture components (1, 2, 4,

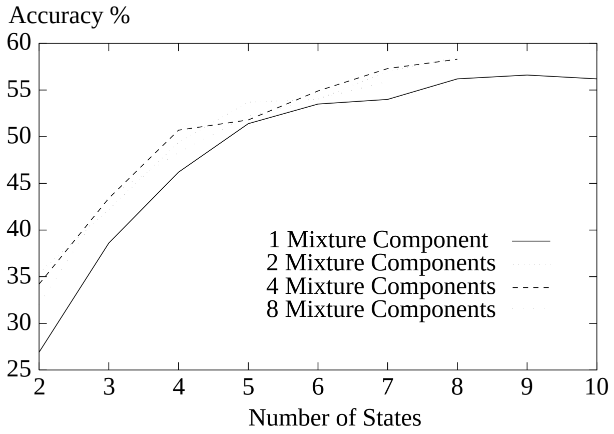

Figure 3: Continuous speaker independent digit recognition rate rate for different number of mixtures as a function of the number of HMM states using visual features only.

8). Since the HMMs only allowed sequential transition probabilities, those training segments where the number of frames was below the number of HMM states were excluded in the training and recognition of segmented digits.

Highest continuous recognition results at 58.5% accuracy were obtained using HMMs with eight states and two mixture components per state. This performance seems relatively high, considering the high visual confusability between several words and the difficult task of speaker independent continuous speech recognition. It is unlikely that perfect recognition for such a task using visual information only can be obtained, whether for humans or for machines.

The results for this HMM architecture using different training and test procedures are shown in Table 1. Although continuous speech recognition is usually much more difficult than isolated speech recognition, the error rates for the procedure of segmented training and testing are not much smaller than for embedded training and continuous recognition. This suggests that the acoustic segmentation might not be adequate for visual word segmentation. This assumption is supported by the fact that the segmented recognition results for embedded training were lower than for segmented training. Continuous recognition results on the other hand improved considerably after embedded training. The relatively high performance for continuous word recognition might however be due to the fact that the words were always spoken in the same context.

Results for continuous digit recognition and embedded training for different numbers of states and mixtures are summarised in Fig. 3. The accuracy increased substantially with the number of HMM states. This suggests that the visual speech signal might not contain quasi-stationary segments which

Table 2: Confusion matrix for continuous digit recognition using HMMs with 8 states and 2 mixture components per state.

	0	1	2	3	4	5	6	7	8	9	Del
zero	132	0	2	1	0	0	0	0	2	0	11
un	0	95	1	2	6	3	3	6	2	2	28
deux	3	0	92	1	0	1	3	2	8	5	33
trois	0	2	0	108	2	0	2	0	6	4	24
quatre	0	1	3	0	72	2	4	4	7	0	55
cinq	1	4	2	0	3	59	10	6	2	3	58
six	0	2	1	1	9	3	61	1	6	1	63
sept	1	3	3	0	4	5	4	80	3	10	35
huit	5	2	3	9	1	3	1	1	102	5	16
neuf	3	1	8	1	1	3	0	3	2	117	9
Ins	1	1	2	1	2	5	11	8	9	6	

extend beyond a large number of feature vectors. The smaller error rates could however also be due to the smaller number of deletion errors resulting from HMMs with more states. The performance also generally increased with the number of mixture components. For HMMs with a large number of states, only a few mixture components could be trained due to the limited training data.

The confusion matrix for continuous word recognition using HMMs with eight states and two mixture components is shown in Tab. 2 (rows represent the actual digits, columns the recognised digits, Del stands for deletion errors and Ins for insertion errors). The word recognition rate of the system varied considerably across different words. Visually more distinct words like "zero", "trois", and "neuf" obtained high recognition rates, whereas visually less distinct words like "quatre","cinq", "six", and "sept" were harder to distinguish. These visually less distinct digits are subject to very little facial movements which therefore often resulted in *deletion errors*. For the 1480 test words, the recognition results consisted of 918 correctly recognised words, 332 deletion errors, 230 substitution errors, and 53 insertions. Deletion errors therefore accounted to over half of the total errors.

6. CONCLUSION

The described continuous digit recognition experiment represents one of the largest speechreading experiments with regards to the number of speakers and the size of the database. It also represents one of the first speaker independent continuous speechreading tests. The system obtained a word accuracy of up to 58.5% for the given task, which seems relatively high considering the difficult task. Digits which are visually highly confusable caused most of the errors but visually more distinct words obtained high recognition rates, even for new subject and continuous speech. About half of the recognition errors were due to deletion errors as a result of the high visual similarity between certain words. Some errors might also be due to the small visual frame rate of 25 Hz which is about 4 times lower than typical acoustic frame rates and which might not capture all important speech events. Results suggest that the extracted visual features and their modelling approach generalise well to new speakers and enable successful continuous speechreading.

7. ACKNOWLEDGEMENTS

This work has been performed under the European ACTS-M2VTS project with the financial support of the Swiss Office for Education and Science (BBW).

8. REFERENCES

[1] G. Chollet, J. L. Cochard, A. Constantinescu, and P. Langlais. Swiss French Polyphone and Polyvar : Telephone speech databases to study intra and inter speaker variability. Technical report, IDIAP, Martigny, 1995.

[2] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models - their training and application. *Computer Vision and Image Understanding*, 61:38–59, Jan 1995.

[3] P. Jourlin, J. Luettin, D. Genoud, and H. Wassner. Acoustic-labial speaker verification. In *Proceedings of the First International Conference on Audio- and Video-based Biometric Person Authentication*, Lecture Notes in Computer Science, pages 319–326. Springer Verlag, 1997.

[4] J. Luettin, N. A. Thacker, and S. W. Beet. Locating and tracking facial speech features. In *Proceedings of the International Conference on Pattern Recognition (ICPR'96)*, volume I, pages 652–656. IAPR, 1996.

[5] J. Luettin, N. A. Thacker, and S. W. Beet. Visual speech recognition using active shape models and hidden Markov models. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'96)*, volume 2, pages 817–820, 1996.

[6] J. Luettin and N. A. Thacker. Speechreading using probabilistic models. *Computer Vision and Image Understanding*, 65(2):163–178, February 1997.

[7] S. Pigeon and L. Vandendorpe. The M2VTS multimodal face database. In *Proceedings of the First International Conference on Audio- and Video-Based Biometric Person Authentication*, Lecture Notes in Computer Science. Springer Verlag, 1997.

[8] D. G. Stork and M. E. Hennecke, editors. *Speechreading by Humans and Machines*, volume 150 of *NATO ASI Series, Series F: Computer and Systems Sciences*. Springer Verlag, Berlin, 1996.