MEDIUM VOCABULARY CONTINUOUS AUDIO-VISUAL SPEECH RECOGNITION

Pascal Wiggers, Jacek C. Wojdel, Leon J. M. Rothkrantz

Data and Knowledge Engineering Group
Delft University of Technology
P.Wiggers@its.tudelft.nl,
J.C.Wojdel@its.tudelft.nl,
L.J.M.Rothkrantz@its.tudelft.nl

ABSTRACT

This paper presents our experiments on continuous audio-visual speech recognition. A number of bimodal systems using feature fusion or fusion within Hidden Markov Models are implemented. Experiments with different fusion techniques and their results are presented. Further the performance levels of the bimodal system and a unimodal speech recognizer under noisy conditions are compared.

1. INTRODUCTION

The performance of large vocabulary continuous speech recognition systems has now reached acceptable levels. (Results between 5% and 10% word error rate on a 65k-word vocabulary have been reported for speaker independent systems [8].) But this is only the case for speech uttered under highly controlled laboratory conditions. The performance of these systems rapidly degrades in more realistic environments. One of the problems is noise, introduced either by the transmission channel, as in the case of telephone speech or by the environment.

As a result, audio-visual speech recognition has attracted a great deal of attention in the research community [1, 2, 3] because the visual modality is well known to contain some complementary information to the audio modality. What is even more important in this context, it is not affected by any background noise. But most results presented so far are limited to isolated word recognition or recognition of simple utterances like strings of digits. In this paper we present our experiments with a bimodal recognizer for continuous speech on Dutch language recordings with a vocabulary of over 1000 words.

In the next sections a brief overview of the baseline speech recognizer and the technique for extracting lip-features is given. Then the multi-modal data set that we collected is described. Subsequently a description of the experiments with different kinds of bimodal integration is given and the results are presented. Finally, the performance of a bimodal system and a speech recognizer under noisy conditions are compared.

2. THE SPEECH RECOGNIZER

The speech recognizer used in our experiments is a simplified version of our large vocabulary speaker independent recognizer described in [5]. It uses continuous density Hidden Markov Models to represent phonemes. Each model has three states connected in a left to right manner, with single Gaussian distribution functions attached to the states. There is a total number of 45 phonemes, which include the phonemes from the SAMPA set [7] and models for silence, optional pauses and mouth noises like, loud breath, sniffing or smacking.

The recognizer was trained on a subset of the Dutch Polyphone database [7]. This is a rather large corpus containing telephone speech from 5050 different speakers from all dialect regions in the Netherlands. The utterances contain all Dutch phonemes in as many phonetic contexts as the designers of the database could find.

A training set of 22626 utterances was selected from this database, all read sentences from newspaper articles. The utterances did not contain any background noise but mouth noises, like smacking, sniffing, loud breaths were allowed.
All utterances, that were originally in A-law 8-bit wave format, were encoded to Mel-frequency cepstral coefficient vectors, using a sampling rate of 10 ms and a segment window of 25 ms. Each vector contains twelve cepstral coefficients with log energy and delta and acceleration coefficients added, all scaled around zero by subtracting the cepstral mean from all vectors. This resulted in 39 dimensional feature vectors. The models were trained iteratively, using embedded Baum-Welch re-estimation and Viterbi alignment.

3. AUTOMATIC LIP-READING

The lip-reading part of our recognizer is based on the visual feature extraction technique described in [10]. It utilizes a lip-selective color filtering and allows for estimation of both geometric and intensity related features of the mouth. The geometry of the mouth initially extracted as a 36 dimensional vector is further simplified using principal component analysis (PCA) down to only 5 dimensions that cover 97% of the data variation. The intensity changes within the mouth image provided two groups of coefficients describing the visibility and position of teeth and mouth cavity (see [10] for further

details). The resulting 6 intensity parameters were used without further processing.
The visual data extracted in this way proved to be sufficiently accurate to allow for developing a limited vocabulary lip-reading system. This system performs with 70-80% word accuracy performance in a concatenated digit recognition task for a single person. As a natural step further, we tried to apply the same techniques to develop the automatic lip-reader for our large vocabulary data set. With such a system we could build a bimodal speech recognizer based on the late integration principle; by combining its outputs with the outputs of the speech recognizer using a Bayesian approach. However this method is not very well suited for continuous speech recognition as integration can only be done after an utterance is completely spoken, thus introducing a time delay. Furthermore N-best lists or output lattices are needed to combine hypothesis from both systems as they may output a different best hypothesis introducing even more overhead costs. This excluded the possibility of using late integration model for our bimodal speech recognizer. In section 5 we introduce an alternative method.

4. DATA SET

A problem with bimodal speech recognition experiments is the lack of multi-modal databases containing both audio and visual information. Of the few databases available, like M2VTS [9], most contain only single words or digits and as a result they are not very well suited for the development of continuous speech processing systems. Therefore we started to gather our own audio-visual data set. A number of respondents were asked to read prompts showed on the screen of a laptop in front of a digital video camera.

The video sequences were stored as MPEG1 stream, with a frame rate of 25 Hz. The audio was recorded using a sampling of 44 kHz with 16-bit resolution. For use in these experiments the audio files were converted to 8 bit A-law format, the format used by the speech recognizer.

A prompt collection divided in 24 sections is used. Each section of the prompt set contains a fixed number of different utterances. The utterances in a section are:

- 1 sentence containing 10 separate small words.
- 10 phonetically rich sentences
- 3 ten-digit sequences. These digits were randomly generated and have uniform distribution in the whole prompt set.
- 4 spelled words.
- 5 utterances from a telebanking application.

At the current stage, we have recorded 8 sessions with 8 different subjects. This gives in total over 4 hours of continuous recordings. The recorded subjects are all native Dutch speakers. Data from five of these subjects, including the only female speaker, was transcribed and used in these experiments. From each subject 4 or 5 sessions were used to create a training set of approximately 500 utterances from all speakers and an independent test set containing 40 randomly chosen utterances from all speakers.

Since the audio in this data set is recorded using a PC microphone and the speech recognizer was trained on telephone audio there is a distortion between the speech vectors produced by the HMMs and the actual data.

To get a baseline system suited for comparison with other systems that use this database the speech recognizer was adapted to the new data set using Baum-Welch re-estimation. Furthermore a bigram language model containing 1050 words was also trained on this data. This language model was used in all experiments described in this paper. The recognition results of the adapted recognizer are shown in table 1.

5. AUDIO-VISUAL RECOGNITION

Audio-visual fusion can be done at several stages in the recognition process, at the feature level, presented in section 5.1 or at the model level, presented in sections 5.2 and 5.3.

5.1. Feature fusion

In feature fusion the feature vectors from both modalities are simply concatenated to generate a single vector on which a regular HMM based recognizer can then be trained. For this and subsequent experiments video features were extracted for all utterances of the multi-modal data set using the technique described earlier. These features were then concatenated to the corresponding audio feature vectors using linear interpolation between video frames to get a frame rate of 10 ms.

Because our data set is currently to small to train a robust continuous speech recognizer, let alone a bimodal recognizer, a different approach was taken. The multi-modal recognizer was created by extending the 39 dimensional distribution functions of the (unadapted) baseline speech recognizer to 50 dimensional distributions. The additional 11 means and variances of all states of all models were initialized with the global means and variances calculated over the entire video data set. Thus the audio part of the system was already reasonably trained and needed only some adaptation to the new data set, but the visual part could only be trained on the multi-modal data set. But as all models initially have the same parameters for their visual features the distribution of the feature vectors during Baum-Welch re-estimation will be guided by the speech features. This way a continuous multi-modal recognizer can be obtained in a few training cycles with a limited amount of training data. The speech part of the models ensure robustness while the video part may give valuable cues to differentiate between homophones.

To train the video part of the system and adapt the audio part of the system, the combined models were re-estimated twice, using the bimodal training data. The models in this system thus received just as much training as the models in the adapted speech only system. The recognition results of this system on the test set are shown in table 1.

Table 1: Recognition results of ASR

system	word rec. %	accuracy %
speech recognizer	84.24	83.82
feature fusion	83.69	78.61

5.2. Multi-stream phoneme based recognition

The feature fusion system, although attractive because of its simplicity, was not able to improve upon the speech only system. This is not a surprise the model is very rigid. One of the problems with this approach is that is does not take into account the reliability of the separate streams. The audio stream is likely to be more reliable than the video stream in the setup described here, because of the clean audio and the well-trained speech part.

In model fusion two different data streams are used and these are combined within the Hidden Markov model. The multi-stream HMM explicitly models the reliability of its streams. In its simplest form, the state synchronous multi-stream model, it uses separate distributions for its streams in each state. The observation likelihood of the state is the weighted product of the likelihoods of its stream components, as shown in formula 1, where γ_s are the weights.

$$b_j(o_t) = \prod_{s=1}^{2} N(o_t, m_{sj}, \Sigma_{sj})^{g_s} \qquad (1)$$

A multi-stream recognizer was build using a similar approach as with the feature fusion model. The models from the baseline speech recognizer were used for the audio stream and the distributions in the video stream were initialized with the global mean and variance of the entire video data set. This system was also re-estimated twice. A number of recognition experiments was run on the test set using different weighting schemes, the results are shown in table 2. The video weights and audio weights add up to two in all cases.

Table 2: Multi-stream fusion results

system	word rec. %	accuracy %
phonemes; equal weights	83.69	78.61
phonemes; audio weight 1.2	84.43	79.41
phonemes; audio weight 1.4	84.22	78.88

By setting the weights so, as to put more emphasis on the audio stream this system is capable of doing a little better than the stand-alone speech recognizer. A shortcoming of this system is that it uses phones as basic units but from a lip-reading point of view it is hard to distinguish between certain phonemes, because of similar lip movements.

Table 3: Visemes, using SAMPA notation

Viseme	Phonemes	Viseme	Phonemes
0	sil, sp	9	E, E:
1	f, v, w	10	A
2	s, z	11	@
3	S, Z	12	i
4	p, b, m	13	O, Y, y, u, 2:, o:, 9, 9:, O:
5	g, k, x, n, N, r, j	14	a:
6	t, d	15	h
7	l	16	Ei
8	I, e:	17	mn (mouth noise)

5.3. Multi-stream viseme based recognition

To solve the problem indicated above, it was decided to use visemes for the video stream. A viseme is basically a phoneme class; the visemes we adopted are shown in table 3.

The use of different units for the stream was realized by tying the distribution functions of corresponding states in the second stream for phonemes that are in the same phoneme class. The limited training data problem is also partially solved this way, because there is now more data per model in the second stream available. As with the previous systems this system was also re-estimated twice before recognition experiments were conducted.

Table 4 shows the recognition results of a number of viseme systems with different weights, once again the audio and video weights add up to 2. This system is capable of improving upon the speech recognizer even when both streams have equal weights. By giving the audio stream higher weight than the video stream the results show more improvement.

Table 4: Multi-stream models using visemes

system	word rec. %	accuracy %
visemes; audio weight: 0.9	84.76	80.48
visemes; audio weight: 1.0	85.56	80.28
visemes; audio weight: 1.1	85.92	82.09
visemes; audio weight: 1.2	85.03	80.75

6. NOISE ROBUSTNESS

In the experiments described in the previous section the improvements the bimodal system realized over the audio only system remained modest. This can be explained by the fact that both modalities encode similar information. If the video stream gives reason to belief that a plosive sound is uttered, and the speech recognizer had a hard time choosing between /p/ and /g/ then the bimodal system may correctly pick /p/. But if the speech recognizer already found that a /p/ was uttered then the additional information from the video data does not help much.

Since relatively clean audio was used in the experiments described so far, the speech recognizer did not need the additional information from the lip-data, most of the time. But in a more noise environment the cues given by the video stream may be more valuable. To verify this hypothesis the multi-stream viseme system was tested using noisy data. This was done by adding different levels of white noise to the audio samples in the test set. The performance of the systems was measured for signal to noise ratios between 20 dB and –5 dB.

The performance of the speech-only system degrades rapidly under these conditions as can be seen in figure 1. The results of the bimodal system are also shown in the figure. At low noise levels the multi-modal system performs slightly better than the speech recognizer, but as the noise level increases the bimodal systems clearly outperforms the unimodal system. Once again the multi-stream model with viseme models in the second stream shows the best results. At a signal to noise ratio of 5 dB the difference is 12%. As the noise level approaches –5 dB the audio recognition gets so poor that the visual cues can no longer provide adequate help.

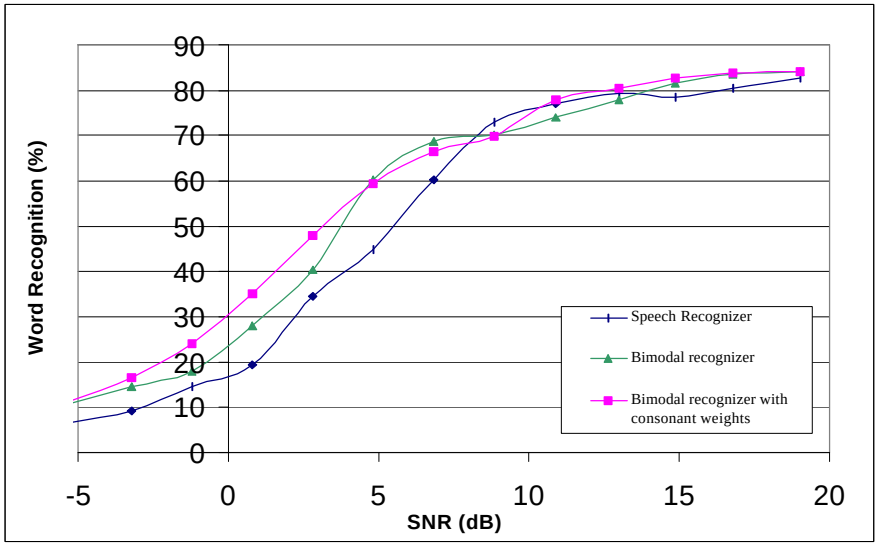

Figure 1: Recognition results for different SNR levels

From figure 1 it can also be observed that the audio stream is no longer more reliable than the video stream. But is also proved hard to do better than the system with equal weights for both streams, by giving the video stream a higher weight. Our lip reading technique seems to do especially well in discriminating between consonants (for example /f/ and /s/) therefore only the weights of the consonant visemes were increased. Figure 1 shows a system with consonant weights that were gradually increased as the noise level increased (up to a noise level of 0.8 dB, for higher levels the weights were decreased again). This system impressively outperformed all other systems.

7. CONCLUSIONS

Our experiments showed that adding visual cues to a continuous speech recognizer results in better performance. This is especially the case in noisy environments. The best results so far were obtained by using a multi-stream Hidden Markov model with phoneme units for the speech stream and viseme units for the video stream. In the case of clean audio the speech stream dominates the performance of the system. In the case of noisy audio the relative performance of the system get better as the weights of the video stream are gradually increased according to the noise level. Improvements up to 16% have been reached.

Although the bimodal system performs well there remain still a number of issues to be solved. Firstly the weights of the data stream were set in an ad hoc manner depending on outside knowledge of the noise levels. In a more practical system this should of course be done automatically on the basis of some reliability measurement from the incoming data. But as estimation of the SNR level is not a trivial task it is still an open question how to achieve this.

Secondly, the experiments described in this paper were performed using a simple speech recognizer and a small multi-modal data set. We believe that a robust bimodal system for continuous speech recognition can be build if a sufficiently large audio-visual speech corpus is available.

8. REFERENCES

[1] Neti, C. Potamianos, G., Luettin, J. Mattews I., Glotin, H., Vergyri, D., Sison, J., Mashari, A., Zhou, J., "Audio-Visual Speech Recognition", *IBM T.J. Watson Research Center, Summer Workshop 2000*, Final Report.

[2] A. Verma, T. Faruquie, C. Neti, S. Basu, A. Senior, "Late integration in audio-visual continuous speech recognition", *Automatic Speech Recognition and Understanding, 1999*

[3] Dupont, S. Luettin J., "Using the Multi-Stream Approach for Continuous Audio-Visual Speech Recognition", IDIAP Research Report 97-14.

[4] Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P., *The HTK* Book (for HTK version 3.0), Cambridge University Engineering Department

[5] Wiggers, P, Wojdel J., Rothkrantz, L. "A Speech Recognizer for the Dutch Language", *Euromedia 2002* Modana, Italy.

[6] Wojdel, J., Wiggers, P., Rothkrantz, L. "The Audio-Visual Corpus for Multimodal Speech Recognition in Dutch Language", *submitted to: ICSLP 2002, 2002*

[7] Damhuis M., Boogaart T., in 't Veld, C., Versteijlen, M.,W. Schelvis, W., Bos, L., Boves L., "Creation and Analysis of the Dutch Polyphone Corpus", *Proceedings ICSLP '94*, pp. 1803-1806, 18-22 September 1994,Yokohama, Japan

[8] Young, S. J., Chase, L. L., "Speech recognition evaluation: a review of the U.S. CSR and LVCSR programmes", *Computer Speech and Language (1998) 12, pp. 263-279*

[9] S. Pigeon and L. Vandendorpe, "The M2VTS multimodal face database" in *Lecture Notes in Computer Science: Audio- and Video- based Biometric Person Authentication (J. Bigun, G. Chollet and G. Borgefors, Eds.)*, vol. 1206, pp. 403--409, 1997

[10] Wojdel, J., Rothkrantz, L., *Using Aerial and Geometric Features in Automatic Lip-reading*, Proceedings of Eurospeech 2001, Scandinavia