Integration of Speech Recognition and Automatic Lip-Reading

Pascal Wiggers, Leon J. M. Rothkrantz

Data and Knowledge engineering
Delft University of Technology
Mekelweg 4, 2628 CD Delft, The Netherlands
P.Wiggers@its.tudelft.nl
L.J.M.Rothkrantz@its.tudelft.nl

Abstract. At Delft University of Technology there is a project running on multimodal interfaces on the interaction of speech and lipreading. A large vocabulary speaker independent speech recognizer for the Dutch language was developed using Hidden Markov Toolkit and the Polyphone database of recorded Dutch speech. To make the system more noise robust audio cues provided by an automatic lip-reading technique were integrated in the system. In this paper we give an outline of both systems and present results of experiments.

1 Introduction

The topic of multimodal interfaces is attracting more and more attention in the research community. The focus is especially on the combination of automatic speech recognition with other modalities, either to improve recognition, as in the work done on audio-visual speech recognition by IBM [6, 10] or the provide a natural human-computer interface as in the German SmartKom project [9].

At Delft University of Technology there is a project running on multimodal interfaces. Systems have now been developed for several modalities. An automatic speech recognizer has been build [1] using Hidden Markov Toolkit (HTK) and a automatic lip-reader has been developed [3]. In this paper we report about the interaction between these two modalities and present the speech recognizer and an integrated audio-visual recognizer. Experiments with these systems were conducted under noisy conditions to show that including the visual modality in a speech recognizer leads to more noise robustness.

2 Development of the Speech Recognizer

Although a number of well-performing speech recognition systems is now available, we lacked a system that could easily be altered or adapted to fit our specific needs. Therefore we decided to build our own system. The final system uses context dependent models, but as this system was build and refined incrementally, actually a whole set of recognizers has been created ranging from a simple monophone recognizer to a sophisticated multiple mixture triphone system. The system was developed using a variety of tools, programs and scripts. By far the largest set of tools and software libraries was taken from the Hidden Markov Toolkit [7]. Furthermore a number of programs was written that filled the gaps in the development process not covered by any of the tools. This includes a program for data selection, some tools for creating an initial acoustic model set and a tool for creating scripts for triphone clustering.

2.1 Data Preparation

The training and testing data for this project was taken from the Dutch Polyphone database [4]. This is a rather large corpus containing telephone speech from 5050 different speakers in 222075 speech files. The speakers were selected from all dialect regions in the Netherlands and the ratio between male and female speakers is almost fifty-fifty. The utterances contain all Dutch phonemes in as many phonetic contexts as the designers of the database could find.

To ensure well-trained models that can be used for recognizing speech recorded for example using a PC microphone only utterances that did not contain any background noise or disturbing hesitations were selected. Mouth noises, like smacking and loud breaths however were allowed. Nine items per speaker were used, all read sentences selected from newspaper articles. This resulted in a training set containing 22626 utterances. This system was created as a baseline system for further research on the subject of speech recognition within our group. In particular we wanted to build a continuous audio-visual speech recognition system. A problem with bimodal speech recognition experiments is however the lack of multi-modal databases containing both audio and visual information. Of the few databases available, like M2VTS [8], 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 recorded 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 [2]. To be general purpose our prompt set includes isolated words, digit sequences, spelled words and phonetically rich sentences as well as commands from a telebanking application. Data from five of these subjects, male and female, was transcribed and used in the experiments described in this paper. A training set of approximately 500 utterances from all speakers and an independent test set containing 40 randomly chosen utterances from all speakers were formed.

The utterances in all data sets were encoded to Mel-frequency cepstral coefficient vectors. 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. A sampling rate of 10 ms was used and each vector was calculated over a segment of 25 ms. The SAMPA set [4] was adopted as phoneme set for the recognizer. Three special purpose phonemes were added to model (optional) silence and mouthnoise The resulting set contained 45 phones. For each of these phones a Hidden Markov Model was created. All models except the silence models share the same topology, which consisted of non-emitting start and end states and three emitting states using single Gaussian density functions. The states are connected in a left-to-right way, with self-loops and no skip transitions.

2.2 Training

As only unsegmented training data was available we initially set the mean and variance of all the Gaussians of all the models to the global mean and variance of the complete data set. These models were then trained iterative using embedded Baum-Welch re-estimation and Viterbi alignment. The short pause model, that is very susceptible to picking up vectors belonging to neighboring phones because of its optional nature, was added only after a few training cycles. The resulting single Gaussian monophone system was tested using a Viterbi decoder. An analysis tool, fully compatible with the standard US NIST scoring package, was used to calculate word recognition percentage and the word accuracy. In this test and in all other tests conducted during development the development test set was used. A backed-off bigram language model containing 1060 words was used in these tests. The first row of Table 1 shows the recognition results from the monophone system. To capture coarticulation effects word internal triphones were introduced. In total there were 8570 different triphones in the training data, but for many of these models there only was a single example in the training data making reliable estimates for these models difficult. To find the right balance between the number of models and their modeling accuracy data-driven state clustering was used to obtain a smaller set of generalized triphones. A system containing 2526 triphones provided a good balance between the number of parameters and the modeling accuracy. It contained less than one third of the original triphones and had at least three models per cluster and in most cases more. Each cluster had at least four examples in the training data. The transition matrices of the triphones that corresponded to the same monophone were tied. Models that ended up having their states in the same clusters and which had the same transition matrix were tied to obtain generalized triphones. To overcome the problem of unseen triphones, without resorting to knowledge based techniques, we introduced a technique that can be described as 'backed-off triphone approach'. In this approach the original monophone models augment the triphone model set. During recognition the word network is constructed by inserting the HMMs in the language model, triphone models are used whenever available, otherwise the corresponding monophone is used. This is implemented by tying all triphones that have no model of their own to the corresponding monophone. Essentially the monophones become generalized triphones. Of course they are less specialized than the other generalized triphones because they are trained on all corresponding triphones but the ones they represent, but being monophones they are general enough to cover the unseen triphones. For both the triphone model set and the monophone model set the number of Gaussian mixtures in the distribution functions was incremented to get more realistic distribution function. The results are shown in Table 1.

2.3 Evaluation Experiments

The final set of acoustic models consisted of the 17-mixture generalized triphone set and the 15-mixture monophone set. This system was tuned using the development test set. A grammar scale factor was used to regulate the relative influences of the language model and the acoustic model. These results presented so far were obtained on the test set that was used during the development process to tune several, to test the robustness of the system an evaluation test was performed using 100 sentences, each of which was spoken by a different person. The bigram language model used contained 5017 different words. The result is shown in Table 1. The experiment showed that the recognizer generalizes very well to speakers it was neither trained nor tuned on even when a large word network is used.

Table 1. Recognition results of ASR

System	Percentage of words recognized	Word accuracy
monophone	38.34%	-28.42%
15-mixture monophone	56.81%	15.51%
17-mixture triphone	71.00%	39.57%
Development set	95.27%	89.59%
Evaluation set	93.55%	88.76%

3 Audio-Visual Speech Recognition

As is clear from Table 1 the recognizer performs well on relatively clean audio. But as with most speech recognizers performance degrades severely when background noise is involved (Fig. 1). The visual modality is well known to contain some complementary information to the audio modality and, what is even more important in this context, it is not affected by any environmental noise. Thus, to get a more noise robust system the visual modality was taken into account by extending the speech feature vectors with features that were extracted using the lip-reading technique described in [3], resulting in a bimodal continuous speech recognizer. For these experiments the multi-modal data set we recorded was used. To get a baseline speech recognition system suited for comparison in these experiments the monophone system described above was adapted to the audio part of the data set by a number of re-estimation cycles. The results are shown in Table 2.

3.1 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 mentioned earlier. These features were then concatenated to the corresponding audio feature vectors. Because our data set is currently to small to train a robust bimodal recognizer from scratch, a different approach was taken. The multi-modal recognizer was created by extending the 39 dimensional distribution functions of the (unadapted) monophone 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 ensures robustness while the video part may give valuable cues to differentiate between similar sounding phonemes. 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 are shown in Table 2.

3.2 Multi-stream Phoneme Models

The feature fusion system, although attractive because of its simplicity, was not able to improve upon the speech-only system. 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. 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.

$$b_j(o_t) = \prod_{s=1}^{2} N(o_t, \mu_{sj}, \Sigma_{sj})^{\gamma_s} \tag{1}$$

The observation likelihood of the state is the weighted product of the likelihoods of its stream components, as shown in formula 1, where the γ_s are the weights. 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 initialize with the global mean and variance of the entire video data set. The system was re-estimated twice. Recognition experiments using different weighting schemes were performed, the results are shown in Table 2. The video and audio weights add up to two in all cases. 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 2. Recognition results bimodal systems

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

3.3 Multi-stream Viseme Models

To solve the problem indicated above, it was decided to use visemes for the video stream. A viseme is basically a phoneme class; the following viseme classes were defined:

{sil, sp}, {E, E:}, {f, v, w}, {A, a}, {s, z}, {@}, {S, Z}, {p, b, m}
{O, Y, y, u, 2:, o:, 9, 9:, O:}, {g, k, x, n, N, r, j}, {t, d}, {I, e:}

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 2 shows the recognition results of a number of viseme systems with different weights. This system is capable of improving upon the speech recognizer even when both streams have equal weights.

4 Noise Robustness

In the experiments described in the previous section the improvements the bi-modal system realized over the audio only system remain 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 has 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 Fig. 1. The multi-stream bimodal system with visemes is also shown. For low noise levels the multi-modal system performs slightly better than the speech recognizer, but as the noise level increases the bimodal system outperforms the unimodal system. At an SNR level of 5 dB the difference is 12%. Recognition results for different SNR levels

From Fig. 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 weighs of the consonant visemes were increased. Figure 1 also 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.

5 Conclusion

In this paper we described the development of a large vocabulary speech recognizer for the Dutch language. The final system uses backed-off triphones that solve the problem of unseen triphones without the need for specific linguistic knowledge. The final system performs well, more than ninety percent of all words are recognized correctly. The recognizer can cope with noises like smacking or loud breathing and the system is speaker independent. It has been tested with vocabularies of more than 5000 words, the performance decreased only slightly in this case.

Our experiments with bimodal recognition showed that adding visual cues to our 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% word recognition have been realized.

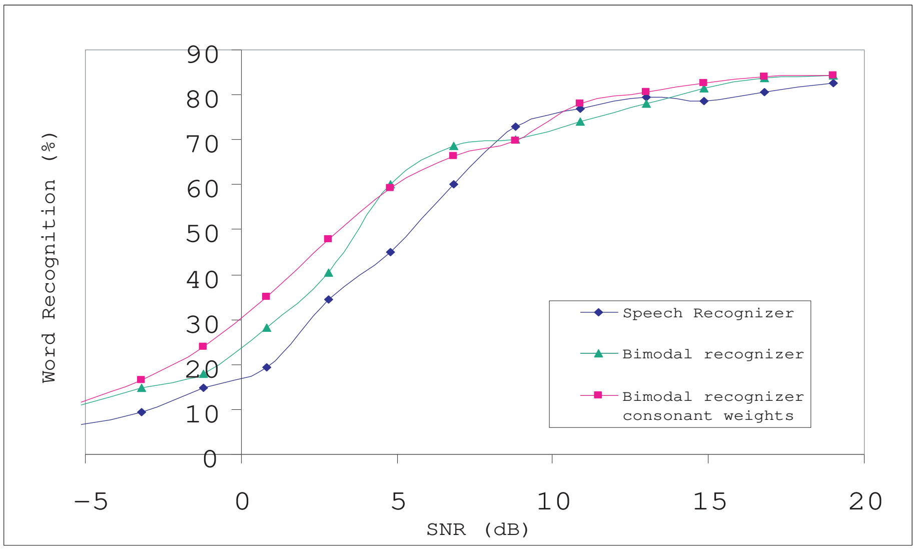

Fig. 1. Recognition results for different SNR levels

References

1. Wiggers, P, Wojdel J.,Rothkrantz, L. *A Speech Recognizer for the Dutch Language*, in proceedings of Euromedia 2002, Modena, Italy.
2. Wojdel, J., Wiggers, P, Rotkrantz, L. *The Audio-Visual Corpus for Multimodal Speech Recognition in Dutch Language*, Submitted to ICSLP 2002, 2002
3. Wojdel, J., Rothkrantz, L., *Using Aerial and Geometric Features in Automatic Lip-reading*, Proceedings of Eurospeech 2001, Scandinavia
4. 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
5. Dupont, S. Luettin J., *Using the Multi-Stream Approach for Continuous Audio Visual Speech Recognition*, IDIAP Research Report 97-14.
6. 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.
7. Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P., *The HTK Book (for HTK version 3.0)*, Cambridge University Engineering Department
8. 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
9. Wolfgang Wahlster, Norbert Reithinger, Anselm Blocher: *SmartKom: Multimodal Communication with a Life-Like Character* , proceedings of Eurospeech 2001, scandinavia
10. A. Verma, T. Faruquie, C. Neti, S. Basu, A. Senior, *Late integration in audio-visual continuous speech recognition*, Automatic Speech Recognition and Understanding, 1999