Large Vocabulary Audio-Visual Speech Recognition Using the Janus Speech Recognition Toolkit

Jan Kratt, Florian Metze, Rainer Stiefelhagen, and Alex Waibel

Interactive Systems Laboratories
University of Karlsruhe (Germany)
{kratt, metze, stiefel, waibel}@ira.uka.de

Abstract. This paper describes audio-visual speech recognition experiments on a multi-speaker, large vocabulary corpus using the Janus speech recognition toolkit. We describe a complete audio-visual speech recognition system and present experiments on this corpus. By using visual cues as additional input to the speech recognizer, we observed good improvements, both on clean and noisy speech in our experiments.

1 Introduction

Visual information is complementary to acoustic information in human speech perception, especially in noisy environments. Humans can disambiguate an acoustically confusable phoneme using visual information because many phonemes which are close to each other acoustically are very different from each other visually. The connection between visual and acoustic information in speech perception is demonstrated by the so-called McGurk Effect [1]. Visual information such as gestures, expressions, head-position, eyebrows, eyes, ears, mouth, teeth, tongue, cheeks, jaw, neck, and hair, could improve the performance of machine speech recognition [2,3]. Much research has been directed towards developing systems that combine the acoustic and visual information to improve accuracy of speech recognition [4,5,6,7]. Many of the presented audio-visual speech recognition systems work on a very limited domain, i.e. either only spelled digits [8,9,10] or letters [11,12,13] are recognized, or only a small vocabulary is addressed [14]. For large vocabulary audio-visual speech recognition, the work by Potamiamos, Neti et al. [15,16,17] has been presented. Their AVSR system is probably the most sophisticated system today.

In this work we also targeting the task of large vocabulary audio-visual speech recognition. Our approach is to use the Janus speech recognition toolkit [18,19], which was developed in our lab and to integrate visual speech recognition into this system. For our experiments we use the data that was used during the workshop on audio-visual speech recognition held at John-Hopkins University in 2000 [15]. In the experiments we observed improvements, both on clean and noisy speech, by using visual cues as additional input to the speech recognizer and hope to further improvements by an enhanced preprocessing and normalization of the data.

C.E. Rasmussen et al. (Eds.): DAGM 2004, LNCS 3175, pp. 488–495, 2004.

Fig. 1. Some pictures of the recorded faces taken from the videos.

2 Databasis

The data provided for our experiments consists of nearly 90 GB of video footage with an overall duration of about 40 hours. During recording of the videos, the speakers were placed in front of a light-colored wall and looked right into the camera. All videos have a resolution of 704 x 480 pixels and a frequence of 30Hz. Figure 1 depicts some sample pictures from the videodata.

Audio data was recorded at a sampling rate of 16kHz in a relatively clean audio environment. The utterances were made in a quiet office with only the noise of some computers in the background.

The utterances are composed of a vocabulary of about 10500 words. For the training we got about 17000 utterances from 261 speakers with a total length of about 35 hours. The testset is made of 26 speakers with about 1900 utterances. These utterances have a total length of four and a half hours. The exact numbers are given in table 1.

Table 1. Available audio visual stored utterances for training and test.

Set	Utter.	Duration	Spk.
Training	17111	34.9 h	261
Test	1893	4.6 h	26

In this work, not all speakers could be used for visual recognition, because the extraction of visual cues failed for some speakers. The biggest set we used consisted of 120 speakers for training a stream-recognizer (see section 4.3). For this system, 17 speakers were used for testing.

The speakers, for which the detection of the region of interest is most robust are selected by an automatic process which considers the variation in the position of the mouth, the variation in the width of the mouth and the number of frames, where the facial features could not be detected at all. Only those speakers were selected, where the respective values are below a certain thresholds.

3 Visual Preprocessing

In order to use the video images of a user's lips for speech recognition, the lips first have to be found and tracked in the video images. We use the program described in [20] to find eyes, nostrils and the lip corners in the pictures. The found lip corners are used to detect the mouth region for the visual training/recognition. For this purpose, a square around the corners is taken with them at the left and right border at about half of the height.

To compensate different illumination conditions and different skin tone of the subjects in the video images, we normalize the extracted mouth regions for brightness. A sample image is depicted in Figure 2.

Fig. 2. The effects of the normalization of the brightness.

As the audio processing works with 100 timeslices every second it would be best to have a video stream with 100 frames per second, but the videos are recorded at a rate of 30Hz. To achieve a signal with 100Hz the existing frames are repeated three times and every third frame four times.

Once having a video stream with a frequency of 100Hz, the pictures are cosine transformed and the 64 coefficients with the highest summation over all training frames are searched as the best coefficients. During selection of the best 64 coefficients the first row and column is ignored because they consist of constant informations which gives no information for the shape of the mouth. Only the 64 best coefficients are used for training and recognition, they keep nearly all information about the video signal. The results are the same when taking all 4096 elements but the training and recognition takes much longer.

As a next step the video signal is delayed by 60ms to achieve better synchronization of acoustic and visual cues. This step is performed because the movements of the lips usually start some time before a sound is produced [12, 21].

4 Experiments

This section describes the audio-visual speech recognition experiments we performed. The first step in building the audio-visual speech recognizer was to train an audio-only recognition system. This was done by using an existing speech recognizer to label the transcriptions of the audio-visual data with exact timestamps. Once this was done, a new speech recognizer was trained on all 261 speakers in the audio-visual data set to get an audio-only reference system.

For the training of the visual recognizer, we use only up to 120 speakers. This was done because the visual lip-tracking module did not provide useful results for all of the video sequences and because training of the visual recognizer was quite time-consuming.

For the visual recognizer, we used a set of 13 visemes, as proposed by [15]. Twelve of these visemes were modeled with three states (begin, middle, end), resulting in 37 viseme states.

In the remainder of this section, we describe the different experiments that we performed. We then present and discuss the obtained results in Section 5.

4.1 Concatenation of Feature Vectors

As a first experiment, we simply concatenated the acoustic and visual input features and trained a speech recognizer on the combined feature vector. The acoustic part of the input vector consists of 13 cepstral coefficients per frame; as visual features, 64 DCT-coefficients are used. In order to provide context to the recognizer, five frames before and after the actual one are connected to the feature vector, which results in a feature vector with 847 elements. To reduce the dimensionality of the feature vector, a linear discriminant analysis (LDA) is calculated. The resulting 42 most significant coefficients are then used as input feature vector for the recognizer.

By concatenating the visual and acoustic input features, an acoustic speech recognition system can easily be adapted to perform audio-visual recognition with little changes. This approach, however, has several drawbacks: First, since the feature vector becomes large, training of the system becomes computationally expensive. A more severe disadvantage is that the importance or contribution of audio- and videodata to the recognition process gets unbalanced, since more features are used for the visual input than for acoustic input. Thus, important information in the audiodata might get lost by performing LDA.

4.2 Reducing the Feature Space

As the first case is not very flexible in changing the given weights for the video- and audio data a more flexible approach is needed to combine acoustic and visual cues. In [15] a hierarchical LDA approach (HiLDA) to reduce the audio-visual feature space was suggested. In this approach, LDA is performed on the visual and acoustic input vector separately. The resulting reduced vectors are then combined and again LDA on the combined audio-visual feature vector is performed.

This procedure has two advantages: First, the computational load is reduced. Now three matrix multiplications are needed for calculating the LDA transformation instead of one before, but the matrices are much smaller. As the needed operations for a matrix multiplication grow by $O(n^3)$ the overall needed number decreases. Second, this approach allows for better adjusting of the weights of the different modalities. We obtained good results when first reducing the visual

feature vector to only 10 coefficients and the acoustic vector to 90 coefficients. During the second LDA step a reduction to 42 coefficients is performed.

4.3 Stream Recognizer

While the hierarchical LDA approach gives much more flexibility than a simple concatenation, it still has some disadvantages that could be solved by a stream recognizer which processes acoustic and visual features independently. For building such a system, a separate classifier is trained to compute likelihoods for each of the input streams separately. Results are then combined at a later stage. This system has proven to give the best results. For the combination the possible hypotheses are scored for each stream and then combined by the given stream weights. For our system best results were achieved if the audiodata gets weighted by 70%.

As the weights for audio- and videodata are not combined before the recognition process it is not necessary to train a stream recognizer again because of changing the weights. This behavior can save a lot of time while testing different scenarios because only the test must be computed for each one. In the two cases described before the training must be computed again for each test.

Another advantage of this additional flexibility is the possibility to automatically adapting the recognizer to a given environment, e.g. by measuring the signal-to-noise ratio of the audio-signal.

5 Results

Now the results of the different audio visual speech recognizers are presented. As the stream recognizer provides the best results the most detailed results are available for this system. Tests on small subsets show the advantage of the HiLDA and stream recognizer to the simple feature concatenation attempt. First we trained three audio-visual recognition systems and an audio only system with 14 speakers. Testing was done on five separate speakers. As can be seen in table 2 the audio only system performs best in this case, followed by the stream recognizer and the HiLDA approach. The concatenation is the worst of the tested scenarios.

The poor audio-visual recognition results in this case are likely due to the little amount of training data. As there is a high variability in the video, more training data is needed.

In our second experiment, we therefore trained the systems with 30 speakers (see Table 2). As you can see, now audio-visual recognition outperforms pure acoustic recognition, even on clean audio data. Testing was again done on five subjects.

In our last experiment we trained an audio-visual stream recognition system with a much bigger training set. Since the acoustic and visual parts of the stream recognizer can be trained independently, different amount of training data can be used for each modality. To train the acoustic part of the recognizer, we used

Table 2. Audio visual word error rates for a System trained on clean audiodata, five speakers are used for the test set.

	14 speakers	30 speakers
audio	48.28%	39.29%
concat	51.36%	40.11%
HiLDA	49.16%	38.48%
stream	49.28%	38.24%

Table 3. Audio visual word error rates for a System trained on 120 speakers for the video part and all 261 speakers for the audio case.

stream weights audio:video	clean audio	noisy audio
100:0 (audio-only)	**25.26%**	**53.94%**
90:10	24.78%	51.19%
80:20	24.30%	48.53%
70:30	**24.10%**	**47.37%**
60:40	24.29%	48.05%
50:50	25.52%	52.72%

all 261 speakers. For the visual part, we used only those 120 speakers, were the automatic tracking of the lips performed the best. Testing was done on 17 subjects.

Table 3 depicts the recognition results depending on the stream weights. It can be seen that weighting the acoustic stream by 70% led to the best recognition results, both on clean and on noisy audio. On clean speech, WER of the audio-visual system is 1% lower than the audio only system (5% relative improvement). In the case of noisy audio, the relative WER decreases by 12.5%: word error rate is dropped from 53.94% absolute to 47.37% absolute.

The noise level was selected to get a similar dimension of WER as in [15]. Figure 3 shows the progression of recognition rates for different SNR values. For rising noise levels higher improvements of the audio visual against the audio only recognition rates are achieved.

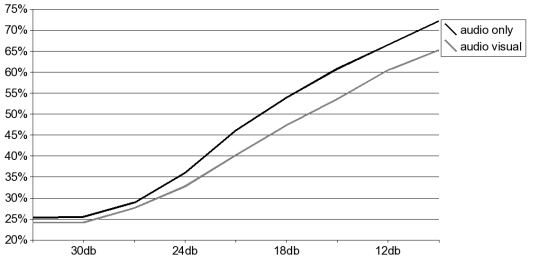

Fig. 3. Plot of word error rates for different SNR values.

6 Conclusion and Future Results

In this paper we have described how audio-visual speech recognition can be done with the Janus speech recognition toolkit, a HMM-based state of the art speech recognizer. Experiments were performed on a large vocabulary speaker independent continuous speech recognition task. We obtained good experimental results by training a stream recognizer, which first computes log likelihoods for each of the input modalities and then combines these hypotheses using the stream weights. With this approach, relative improvements on both clean and noisy speech were obtained. The achieved amount of improvements in relative WER is by now about half of the improvements reported in [15]. We think that this is mainly due to the fact that we could only use a fraction of the data used in [15].

The presented system provides a good basis for further audio-visual speech recognition research. We are now working on the improvement of the facial feature tracking approach in order to being able to use more speakers from the database. In fact, we are now already able to use 200 speakers instead of 120 used for the presented experiments.

Among the first things that we plan to improve is the visual preprocessing of the data. So far, only histogram normalization is done. Since we observed that some subjects tilted their heads quite significantly, we hope to improve the recognition results by appropriate rotation of the input images in the future. Adaptive adjustment of the combination weights for the input modalities should also improve the recognition results in the future.

Acknowledgments. This research has been funded by the European Communities project CHIL, Contract Number IST-506909.

References

1. H. McGurk and J. MacDonald. Hearing lips and seeing voices. Nature, 1976
2. G. Potamianos, C. Neti, S. Deligne. Joint Audio-Visual Speech Processing for Recognition and Enhancement. Proceedings of AVSP 2003, 2003
3. R. Goecke, G. Potamianos, C. Neti. Noisy Audio Feature Enhancement using Audio-Visual Speech Data. ICASSP 02, 2002
4. M.E. Hennecke, K.V. Prasad, D.G. Stork. Using deformable templates to infer visual speech dynamics. 28th Annual Asimolar conference on Signal speech and Computers.
5. A.J. Goldschen, O.N. Gracia, E. Petajan. Continuous optical automatic speech recognition by lipreading. 28th Annual Asimolar conference on Signal speech and Computers.
6. J.R. Movellan. Visual speech recognition with stochastic networks. NIPS 94, 1994
7. P. Duchnowski, U. Meier, A. Waibel. See me, hear me: Integrating automatic speech recognition and lip-reading. Internation Conference on Spoken Language Processing, ICSLP, pages 547-550, 1994

8. S. Deligne, G. Potamianos, C. Neti. Audio-Visual speech enhancement with avcdcn (Audio-Visual Codebook Dependent Cepstral Normalization), IEEE workshop on Sensor Array and Multichannel Signal Processing in August 2002, Washington DC and ICSLP 2002
9. S. Dupont and J. Luettin. Audio-visual speech modeling for continuous speech recognition. IEEE Trans. Multimedia, vol. 2, pp. 141-151, 2000
10. J. Huang, G. Potamianos, C. Neti. Improving Audio-Visual Speech Recognition with an Infrared Headset. Proceedings of AVSP 2003, 2003
11. Uwe Meier, Rainer Stiefelhagen, Jie Yang, Alex Waibel. Towards Unrestricted Lipreading. International Journal of pattern Recognition and Artificial Intelligence, Vol. 14, No. 5, pp. 571-785, 2000, Second International Conference on Multimodal Interfaces (ICMI99), 1999.
12. C. Bregler and Y. Konig. Eigenlips for robust speech recognition. Proc. IEEE Intl. Conf. Acous. Speech Sig. Process, pp. 669-672, 1994
13. I. Matthews, J.A. Bangham, S. Cox. Audiovisual speech recognition using multiscale nonlinear image decomposition. Proc. 4th ICSLP, vol. 1 pp. 38-41, 1996
14. A. Ogihara, S. Asao. An isolated word speech recognition based on fusion of visual and auditory information using 30-frames/s and 24-bit color image. IEICE Trans. Fund. Electron., Commun. Comput. Sci., vol. E80A, no 8, pp. 1417-1422, 1997
15. C. Neti, G. Potamianos et al. Audio-Visual Speech Recognition - Workshop 2000 Final Report. Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, 2000
16. G. Potamianos, C. Neti, G. Iyengar, Eric Helmuth. Large-Vocabulary Audio-Visual Speech Recognition by Machines and Humans, Proc. Eurospeech, 2001
17. G. Potamianos, A. Verma, C. Neti, G. Iyengar, S. Basu. A Cascade Image Transformation For Speaker Independent Automatic Speechreading. Proceedings of the IEEE International Conference on Multimedia and Expo, pp. 1097-1100, 2000
18. M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, M. Westphal "The Karlsruhe-VERBMOBIL Speech Recognition Engine", in Proceedings of ICASSP, Munich, Germany, 1997.
19. H. Soltau, F. Metze, C. Fügen, A. Waibel, "A One Pass-Decoder Based on Polymorphic Linguistic Context Assignment", in Proc. of ASRU, Trento, Italy, 2001.
20. Rainer Stiefelhagen and Jie Yang. Gaze Tracking for Multimodal Human-Computer Interaction. Proc. of the International Conference on Acoustics, Speech and Signal Processing: ICASSP'97, Munich, Germany, April 1997.
21. G. Gravier, G. Potamianos and C. Neti. Asynchrony modeling for audio-visual speech recognition. Proc. Human Language Technology Conference, 2002