Visual Speech Recognition

Automatic System for Lip Reading of Dutch

Alin G. Chițu[1], Karin F. Driel[2], Leon J.M. Rothkrantz[3]

[1, 2,3]Man-Machine Interaction Group, Department of Mediamatica
Delft University of Technology, Delft, The Netherlands
[1]a.g.chitu@tudelft.nl, [2]karindriel@gmail.com [3]l.j.m.rothkrantz@tudelft.nl

Abstract

Visual speech recognition is the next step towards robust and ubiquitous speech recognition. This paper gives detailed information about the work done at Delft University of Technology towards a robust automatic lip reading system. We proudly introduce the data corpus for the Dutch language we have built for the present research, with details on its development and characteristics. In the current research we consider a large spectrum of visual features, namely motion description features based on optical flow, shape description features based on key point detection and a statistical approach, and appearance description features. This paper introduces the use of Active Appearance Models for the detection of landmarks on the speaker face for lip reading. The inference is made based on the Hidden Markov Models approach using the HTK Toolkit.

1. Introduction

Today's trend is to make the communication and interaction between humans and their artificial partners easier and more natural. Speech recognition technology has reached a maximum of performance and good recipes for building speech recognizers have been written. However, the major problems of background noise and reverberations due to the environment are still insurmountable. Therefore, inspecting other sources, other than sound, for complementary information which could alleviate these problems, is a necessity. Lip reading can therefore be seen both as a complementary process to speech recognition, and as a stand-alone process. The applications for lip reading as a stand-alone application are diverse: multimedia phones for the hearing impaired, mobile phone interface for public spaces (e.g. at the time of this writing, phone models that use lip reading are already being designed,) person identification, recovery of speech from deteriorated or mute movie clips, and, perhaps the most promoted applications, security by "video surveillance" (e.g. security cameras that are recovering what is being said without the need of a microphone, and, more importantly, from a large distance).

In this paper we introduce our work towards automatic and robust lip reading in detail. The introduction of our work is split into three obvious parts, namely data acquisition, data parameterization and inference.

Data acquisition is presented in Section 3. The amount of work involved with building an audio-visual speech corpus, is astonishing. However, the result is the largest and most complex bimodal data corpus for the Dutch language to date.

The second important aspect that needed to be covered is data parameterization, which should both reduce the dimensionality and capture the relevant information about the process being modeled. Many techniques can be employed here. In our current work we tried to approach a large pool of methods in order to understand their pros and cons and make an informed choice for our endeavor: automatic and robust lip reading. Section 4 introduces the image processing methods used: Active Appearance Models (AAM) [1] in Section 4.1, Lip Geometry Estimation [2] in Section 4.2 and Optical Flow [3] in Section 4.3. The visual features are introduced in Section 5.

For inference we used the Hidden Markov Model approach and the HTK Toolkit [4] developed at Cambridge University. Details about building the recognition models are found in Section 6.

Section 7 concludes this paper with the current status of our work and the future developments.

2. Related work

Lip reading literature has increased rapidly over the years. There are two possible directions of research, namely data parameterization and inference mechanisms.

Due to the temporal characteristics of both speech recognition and lip reading, the majority of techniques used for other machine learning areas are not suitable. However, for limited vocabulary applications it could still be appropriate. Wang et al. in [5] used b-spline functions to match the spoken utterances. Similar approaches make use of principal component analysis. For large vocabulary (continuous) speech recognition, a time series approach is more suitable. Due to its success in speech recognition, by far the most used method is the Hidden Markov Model and its derivatives.

For lip feature extraction many methods have been developed. They fit mainly in two broad classes: appearance-based methods and geometrical methods; combinations of the two are also not uncommon. Appearance-based methods consider (a transformation of) the raw image for feature extraction [6], [7]. The transformation of the raw image is employed in order to obtain some data reduction. The most popular method for this is Principal Component Analysis (PCA) [8], [9]. As an alternative to PCA, a discrete cosine transform [10] or discrete wavelet transform have been applied. However, these approaches give rise to a very high dimensionality of the feature vectors. On the other side geometrical methods aim to model the visual parts of the speech production apparatus, such as lips, teeth, and tongue, but also other parts of the face. Usually, specific points on the

face are detected and tracked. The detection process is often assisted by 2D or 3D geometrical models of the face [11]. Alternatively, statistical methods based on image filtering algorithms can be used to directly extract the shape of the lips. The dimensionality reduction obtained through the latter approach is very large. Performing optical flow analysis on the input video is used both as a measure of the overall movement (e.g. for onset/offset detection [12], [13], [15], [16]) and as feature generation engine [17].

3. Data acquisition

The data corpus is the foundation of any successful speech research. After working for some time with a small data corpus we arrived at the conclusion that a new, larger and more complex data corpus was needed. We did extensive analyses to understand what the drawbacks of the existing corpora were. The paper [18] provides a rich comparison among some of the most used data corpora and introduces a set of guidelines to be followed when building a data corpus. Some of the important issues spotted were the resolution and frame rate of the video recordings, the richness of the language pool and most importantly, the size of the corpus. Following the analysis made in the paper on the influence of frame rate on the information retrieval [19], the conclusion was that in order to accurately capture all the speech information, we needed to record at higher frame rates than what standard consumer cameras currently provided. Figure 1 shows the poor coverage of the visemes in the case of fast speech rate. The results are based on the DUTAVSC data corpus [20]. Therefore, building a new data corpus became a new point on our agenda. The resulted corpus, which will be made available to the scientific community in the near future, is to our knowledge the largest and most complete bimodal corpus for the Dutch language to date.

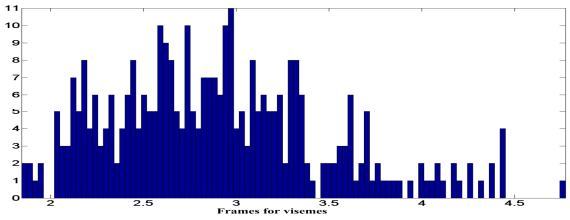

Figure 1. Viseme coverage in the case of fast speech. We have a mean of 3 frames per viseme. The recordings are made with a consumer camera at 25fps.

3.1. Recording settings

We considered recording in a controlled environment (i.e. reasonable noise levels and good illumination). The specific noise can be simulated or recorded in the required conditions and later superimposed on the clear audio data. An example of such database is NOISEX-92 [21]. This dataset contains white noise, pink noise, speech babble, factory noise, car interior noise, etc. On the video side we used a uniform background so that by employing chroma keying techniques one could immerse the speaker in any environment (e.g. weather forecast on television).

One more important improvement is adding the side view to the recordings. We devised two settings to achieve this effect. Firstly we used a mirror place at 45°, but we gave up on this idea due to image distortions and settled for a 2-camera setting. Figure 2 shows the recording environment.

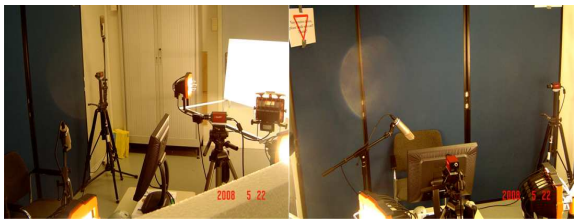

Figure 2. Recording environment.

We used two very sensitive condenser microphones; one for recording the speaker's voice and the second to record the background noise. The audio signal was sampled at 48 kHz and used a sample size of 16 bits. The video was recorded at 100Hz for both side and frontal view, half PAL resolution, and captured only the lower half of the face.

To systematize the recordings we built a prompter tool that provided the user with the next item to be uttered together with some instructions about the speaking style. The tool also controlled the video and audio devices. The resulting audio and video clips were synchronized and clipped at the exact length of the utterance. The speaker had the possibility to direct their own recordings and change the visual settings to maximize the visibility to improve the recording experience. Figure 3 shows a screenshot of the tool.

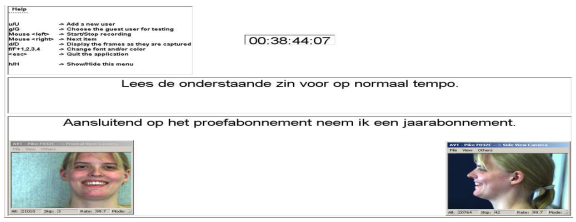

Figure 3. Prompter screen instructing the speaker about the required speaking style and the next utterance.

3.2. Statistics about the data corpus

The language pool contains 1966 unique words, 427 unique phonetically rich sentences, 91 context aware sentences (i.e. banking applications), 72 conversation starters and endings and 41 simple open questions (i.e. for these questions the user was asked to answer the first thing that came to mind. In this way we expected to collect spontaneous speech utterances). For each session, the speaker was asked to utter 64 different items (sentences, connected digit combinations, random words, spellings and free-answer questions) divided into 16 categories with respect to the language content and speech style: normal rate, fast rate or whispering.

The resulted corpus needs a storage memory of 473GB. It consists of recordings of 67 speakers, 20 female and 47 male, of which three recorded 2 sessions. Hence, the corpus has 70

sessions, each with over 60 utterances resulting in a total pool of more than 4200 recorded utterances, with a total recording time of over 6 hours.

For each speaker we recorded the gender, age, education level, native language (as well as possible multilingualism) and region of origin. The last aspect is used to identify possible particular clusters determined by the regional dialects. In [20][22] it is shown that the contextual background of a speaker influences both the content and the aspects of speech. The recognition of the context could enable us to build adaptive systems which can provide increased accuracy and reliability. The speakers were in large majority native Dutch speakers; hence we expect that the cultural impact is minimal.

4. Image Processing Models

This section briefly introduces the image processing techniques used in the current work.

4.1. Active Appearance Models

Active Appearance Models (AAMs) are employed to extract the location of specific points on the face from every frame of the video sequence. The AAM was introduced in [1] and is a generalization of the Active Shape Model (ASM) which combines both shape information (i.e. model based approach) and texture information (i.e. appearance based approach) in its searching scheme.

The AAM method creates a statistical model of shape variation and texture variation. The average shape is determined considering a training set of samples shapes. The sample shapes are aligned using a Generalized Procrustes Analysis. Each face sample is then warped so that the control points match the ones of the mean shape (Figure 4, left).

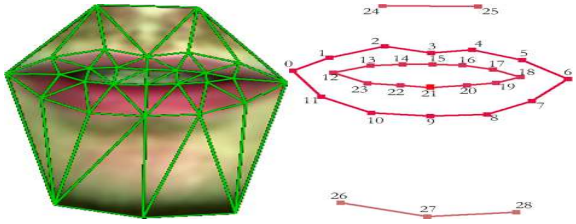

Figure 4. Left: the mean shape and the mean texture used for AAM; Right: the key point model used.

The search starts from the mean model and iteratively modifies the model parameters within the learned range, while minimizing the difference in appearance between the real image and the image synthesized based on the new model (*Figure 5*).

The required number of parameters is computed in both cases by using PCA. As with the majority of all searching schemes a good initial guess helps to speed up the process. The use of a face/mouth detection/tracking algorithm as a prior step was found to greatly speed up the search for the shape parameters during AAM-based processing. This enhancement made a real-time implementation of the algorithm possible.

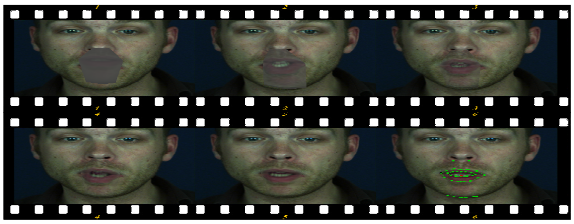

Figure 5. Example search sequence; the first five steps and the optimum found after 34 iterations

4.1.1. Landmarks and their definitions

The shape information extracted by the AAM from a face image is used to compute a set of suitable parameters that describe the appearance of the facial features. Therefore the first step is the selection of the key points that optimally capture speech related movements. The model used is shown in Figure 4 (right), and consists of 29 points distributed around the mouth, chin and nose. The next step is to transform the key point set into a set of representative parameters. This parameterization has the advantage of providing the inference engine with data that encapsulates the most important aspects of speech. This also acts as a dimensionality reduction procedure since the dimension of the feature space is lower than the dimension of the image space. As is described in Section 5.1, the parameters are computed as certain Euclidean distances and areas between the key points. Including a set of constraints to the points will also enable a more uniform annotation.

Since the amount of data to be annotated is enormous, more than 2 millions of frames, we divided the data among multiple annotators. Each annotator was asked to train and supervise the AAM tool. In order to keep the annotation of the video frames uniform we strictly defined the key points.

4.1.1.1 Outer mouth contour

The points on the outer mouth contour are defined as follows:

- point 0 is the leftmost point of the lips (i.e. left mouth corner).
- point 6 is the rightmost point of the lips (i.e. right mouth corner).
- points 2, 3 and 4 are placed in accordance with the philtrum (infra-nasal depression), namely, 2 and 4 at the foot of the philtral columns, respectively, and 3 in the place where the philtrum meets the upper lip.
- points 8, 9 and 10 are correspond to points 4, 3 and 2, respectively.

Points 1, 5, 7 and 11 are placed such that the lip area is covered as closely as possible. However, their positions are preferred to be at equal distances from their neighboring points. It should be noted that the outer mouth contour contains much person-dependent information.

4.1.1.2 Inner mouth contour

In the case of the inner mouth contour the decision was that the stress should be placed on accurately describing the aperture of the mouth. A closed mouth is a special case. The

points on the inner contour are closely related to the ones on the outer contour and have similar definitions.

- point 12 is the leftmost point of the cavity of the mouth. However, in the case of a closed mouth this is not possible to observe. In that case this point should be placed such that it best describes the mouth line, but always to the left of points 13 and 23.
- point 18 is the rightmost point of cavity of the mouth. In a similar way as for point 12, in case of a closed mouth this point should be placed such that it best describes the mouth line, but always to the right of points 17 and 19.
- points 15 and 21 correspond to points 3 and 9 and follow the philtrum. Moreover all 4 points should always lie on the same line.

The last 8 points form pairs as follows: 13 and 23, 14 and 22, 16 and 20 and 17 and 19, and have similar definitions as the corresponding points on the outer contour.

4.1.1.3 Nose

The nose points are only a delimitation of the nose and are used as a reference to compute distances to other points (e.g. the distance from point 27 on the chin to the line formed by points 24 and 25 is used as a feature in our settings). The points are placed at the base of the nose.

4.1.1.4 Chin

Here we are interested in tracking the tip of the chin marked by point 27. Points 26 and 28 only support the detection of point 27 and should be placed symmetrical with respect to the 27 and describe the chin as closely as possible. Point 27 should be aligned to points 3, 9, 15 and 21.

4.2. Lip Geometry Estimation

The method of Lip Geometry Estimation (LGE) [2] uses a color filter to find all lip pixels. Then based on a statistical interpretation of the result the geometry of the mouth is directly estimated. This technique is unique because it does not rely on any a-priori geometrical lip model. Figure 6 exemplifies the usage of this method.

At first the Region of Interest (ROI) is detected by the use of a mouth detection/tracking algorithm. This is very important for at least two reasons: firstly the processing time is greatly reduced and secondly many possible unwanted artifacts are avoided. We include a Gaussian Mixture tracking model to track the mouth in time, and only use mouth detection when necessary. To detect the lip pixels we used several color filters: hue, hue-value, grayscale, pseudo-hue histogram, and neural networks on different color spaces. However, the best results were obtained with neural networks on the RGB values.

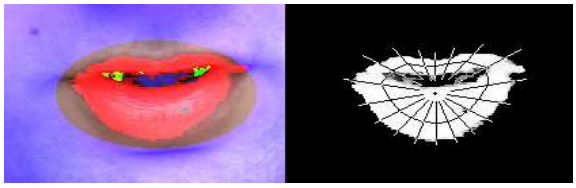

Figure 6. Lip Geometry Estimation approach and feature description.

More statistical methods were used to further reduce the artifacts in the filter results (e.g. outlier removal – any pixel outside the stochastic ellipse in Figure 6 (left) are deleted).

At this time also information such as the visibility of the mouth cavity and teeth are extracted, based on the color intensity but also on the specific arrangement.

By interpreting the result of the image filter as a two-dimensional distribution, we can now describe the shape of the mouth, the mouth aperture and the thickness of the lips. This is obtained by simply analyzing the mean and variance of the given distribution. Figure 6 (right) shows a possible result of this method.

4.3. Optical flow

Optical Flow is a concept that is concerned with the notion of motion of objects within a visual representation. A common definition of Optical Flow is: "The distribution of apparent velocities of movement of brightness patterns in an image." The word "apparent" signals the fact that sometimes the optical flow does not correspond to the true motion field (e.g. "the rotating barber's pole illusion.") There are quite a few methods for optical flow detection, i.e. Lucas-Kanade, Horn-Schunck, phase correlation (i.e. the inverse of normalized cross power spectrum), gradient constraint-based methods, and block correlation methods. Of these, the most used are Lucas-Kanade and Horn-Schunck. We used the algorithm published in [3] by Bruce D. Lucas and Takeo Kanade.

We took the same approach as for LGE, namely first detecting the mouth and then the center of the mouth. Figure 7 (left) shows the optical flow.

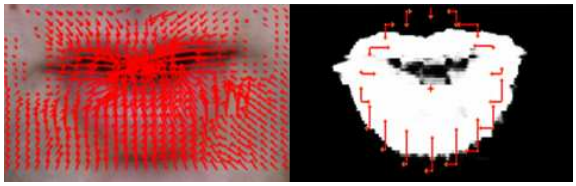

Figure 7. Optical flow approach and feature description.

5. Visual features

Data parameterization depends on what aspect of the data needs to be captured; for speech it should as closely as possible describe the curvature of the mouth and capture the transformations the mouth undertakes. Different image processing methods enable the capture of different aspects of the data.

5.1. Geometric feature

We use seven featured to describe the shape of the mouth: mouth and mouth aperture height, width and area, and nose to chin distance. For clarification we define them as follows:

- **Mouth height** is defined as the Euclidian distance between points 3 and 9.
- **Mouth width** is defined as the Euclidian distance between points 0 and 6.
- **Mouth area** is defined as the area inside of the outer lip contour.
- **Aperture height** is defined as the largest Euclidian distance between the pairs of points (13, 23), (14, 22), (16, 20), (17, 19) and (15, 21).
- **Aperture width** is defined as the Euclidian distance between the leftmost point (or coinciding pair of points) and the rightmost point (or coinciding pair of points) of the inner lip contour.
- **Aperture area** is the area covered by the mouth aperture, namely the inner lip contour.
- **Nose to chin distance** is the minimum distance between chin point 27 and the line defined by the nose points. This denotes the openness of the jaw.

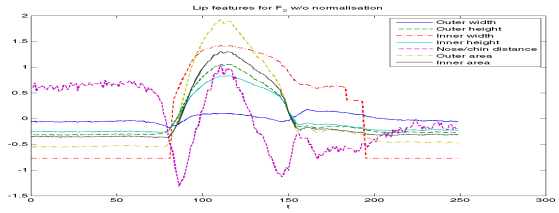

Figure 8. Geometric features for letter F (with the viseme representation "[E] [F]").

5.2. LGE features

LGE features are shown in Figure 6 (right). The space is split in 18 sectors centered in the middle of the mouth. The visual features are the mean and variance of the corresponding conditional distribution for each direction. The mean location describes the shape of the mouth, while the standard deviation describes the thickness of the lips in that specific direction. In these settings, to these features we also added color intensity features that provided a 20% performance gain.

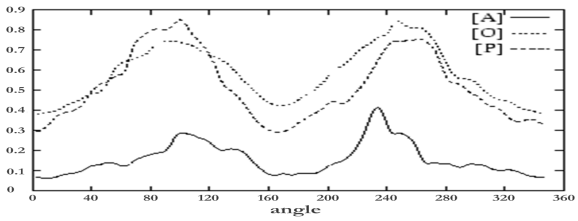

Figure 9. LGE features: lips thickness for an instance of the visemes [A], [O] and [P], respectively.

5.3. Optical flow features

We employed optical flow analysis to directly recover the motion information apparent around the speaker's mouth. In a similar way as for the LGE feature extraction we consider 18 sectors around the speaker's mouth. We then compute the horizontal and vertical displacement for each sector. Figure 10 shows some results for optical flow, where the onset/offset points are clearly visible.

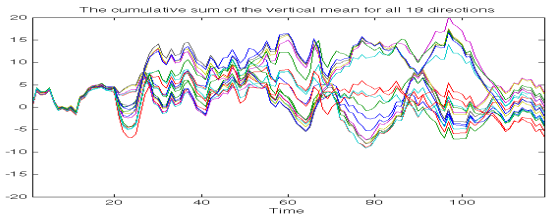

Figure 10. The distribution over time of the vertical mean of the optical flow for a fan with 18 distinct directions shown as cumulative sum.

6. Inference

For the actual recognition we used the Hidden Markov Model Toolkit developed at Cambridge University.

The recognition units were chosen to be visemes. The visemes are the phonemes' corresponding semantic unit of speech in the visual domain. For the Dutch language, most researchers use a set of 40 phonemes. However, since there is no one-to-one mapping from the phoneme set to the viseme set researchers define the visemes by clustering together the phonemes which have a similar visual outcome. The decision about the similarity of the outcome is based on the degree of confusion and distinction human subjects are attaining [23], [24]. We used a set of 16 visemes, as shown in Table 1 and defined in the paper [25]. To the set of visemes we added two extra models for *silence* and *short pause*, thus a number of 18 models were used.

Table 1. Viseme set for the Dutch language in SAMPA notation.

#	Viseme	Phoneme set	#	Viseme	Phoneme set
1	[F]	f v w	9	[I]	I e:
2	[S]	s z	10	[E]	E E:
3	[X]	S Z	11	[A]	A
4	[P]	p b m	12	[@]	@
5	[G]	g G k x n N r j h	13	[i]	i
6	[T]	t d	14	[O]	O Y y u 2: o: 9 9: O:
7	[L]	L	15	[a]	a:
8	[H]	H	16	[E+]	Ei

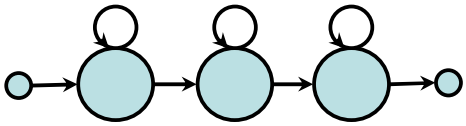

Figure 11. 5-State Left-Right HMM where only the middle states are emitting. The first and last state are used to link models together.

The main problem with this definition of visemes is that the separability of the words it decreasing since some of the words end up having the same transcription. In [10] it was computed that the decrease in word separability can go up to 10%. In the case of our task of connected letter recognition this effect appears for instance for letters: (P, B), (D,T), (G,J), (N,R), (O,U) and (V,W).

Each viseme is modeled by a 5-state Gaussian mixtures continuous density left-right HMM, of which only three states are emitting. The model is shown in Figure 11. Special attention is paid to the silence models since in the case of visual data even in the case when nothing is uttered the signal's energy can still be very high. The models are trained iteratively, using embedded Baum-Welch re-estimation and Viterbi alignment.

7. Results

Due to the large amount of work required to process all the data we decided to start with a speaker dependent approach. Even is this case we processed over 2 million frames from which features had to be extracted. However, by including a mouth detection/tracking system we managed to optimize the processing time in case of AAM and LGE to almost real time. The optical flow on the other hand is very resource intensive.

We analyzed different tasks with different degrees of difficulty such as connected digits, connected letters, connected words and random sentences. The latter is actually synonymous with continuous lip reading, which is the ideal application we hope to achieve. We also considered, where possible, "tri-viseme" (equivalent to tri-phone) constructions with inner word context. However due to the small size of the corpus there were many unseen tri-visemes.

7.1. Connected digit recognition

The results obtained for the digit recognition task are shown in Table 2.

Table 2. Percent correct and WRR for the digit recognition task.

	No Delta/Acc	With Delta/Acc
Monovisemes	69.86%;WRR=61.64%	78.08%;WRR=68.49%
Trivisemes	65.75%;WRR=60.27%	72.60%;WRR=50.68%

Inspecting the confusion matrix we found out the digits "3" and "4", "1" and "9" are the most confused.

7.2. Connected letter recognition

This task is slightly more difficult then the digit recognition task for two reasons: the letters are on average shorter then the digits when considering their viseme transcription (3.7 visemes per digit compared to 1.55 visemes per letter) and the number of classes to recognize is 26. Because of the one-to-many mapping between visemes to phonemes the number of classes is decreased to 20, making the distinction between letters troublesome. Table 3 shows the results for this case.

Table 3. Percent correct and WRR for the letter recognition task.

	No Delta/Acc	With Delta/Acc
Monovisemes	40.32%;WRR=-35.48%	44.62%;WRR=-25.27%
Trivisemes	37.10%;WRR=-44.62%	49.46%;WRR=-12.90%

While the percentage of the correctly recognized words decreased to a still acceptable level, the word recognition rate decreased to negative levels. This is due to the large number of inserted words which is probably a consequence of both the short length of words and the decreased separability.

7.3. Continuous speech recognition

A big disappointment was the fact that we were not yet able to train the system for continuous speech to a reasonable level. Therefore we decided to gather additional sessions of speaker data. When training a recognizer on random sentence samples without an imposed language model we attained more reasonable results (as seen in Table 4) which indicates that an under-trained language model was indeed part of the problem.

Table 4. Percent correct and WRR for the continuous speech task without a language model (i.e. the visemes played the role of words)

	No Delta/Acc	With Delta/Acc
Mono-visemes	29.47%;WRR=23.96%	34.05%;WRR=24.64%

8. Future work

Since our goal was to achieve continuous lip reading, we started new recording sessions which almost doubled the amount of data. Also since the language model was a bottleneck in our first experiment, we made plans to collect a large language corpus and construct a more reliable language model.

Currently we plan to exhaustively analyze the influence of the visual feature set on the results. Also, other aspects such as the exact influence of the frame rate and the influence of the facial expressions on the recognition will be investigated.

We envision a system that can process recordings in real time, take advantage of the context and adapt accordingly..

9. Acknowledgements

The research reported here is part of the Interactive Collaborative Information Systems (ICIS) project, supported by the Dutch Ministry of Economic Affairs, grant nr: BSIK03024 file.

We thank all the participants in our long and tiring recoding sessions, but especially Pegah Takapoui and Mathijs van Vulpen for their valuable help in setting up the recording environment and supervising the recordings.

10. References

[1] Cootes, T.F., Edwards, G.J., Taylor, C.J., "Active Appearance Models", In *Proceedings of the 5th European Conference on Computer Vision* – Volume 2, pp. 484–498, Springer, 1998.

[2] Wojdel, J. C., and Rothkrantz, L. J. M., "Visually based speech onset/offset detection", in *Proceedings of 5th Annual Scientific Conference on Web Technology, New Media, Communications and Telematics Theory, Methods, Tools and Application (Euromedia2000)*, Antwerp, Belgium, pp. 156–160, 2000.

[3] Lucas, B. D., and Kanade, T., "An iterative image registration technique with an application to stereo vision", in *Proc. Seventh International Joint Conference on Artificial Intelligence*, pp. 674–679, 1981.

[4] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X. A., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P., "The HTK Book (for HTK Version 3.4)", 2005.

[5] Wang, S. L., Lau, W. H., and Leung, S. H., "Automatic Lipreading with Limited Training Data.", In *Proceedings of the 18th international Conference on Pattern Recognition* Vol. 3, Washington, DC, pp. 881-884, 2006.

[6] Li, N., Dettmer, S., and Shah, M., "Lipreading using eigen sequences", In Proc. International Workshop on Automatic Face- and Gesture-Recognition, (Zurich, Switzerland), pp. 30–34, 1995.

[7] Li, N., Dettmer, S., and Shah, M., "Visually recognizing speech using eigen sequences", *Motion-based recognition*, 1997.

[8] Hong, X., Yao, H., Wan, Y., and Chen, R.,"A PCA Based Visual DCT Feature Extraction Method for Lip-Reading", pp. 321–326, 2006.

[9] Bregler, C., and Konig, Y., ""Eigenlips" for robust speech recognition", in Acoustics, Speech, and Signal Processing, ICASSP-94, IEEE International Conference on, 1994.

[10] Duchnowski, P., Hunke, M., Büsching, D., Meier, U., and Waibel, A., "Toward Movement-Invariant Automatic Lip-Reading and Speech Recognition", in International Conference on Acoustics, Speech, and Signal Processing, (ICASSP-95), vol. 1, pp. 109–112, 1995.

[11] Essa, I. A., and Pentland, A., "A Vision System for Observing and Extracting Facial Action Parameters", in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 76–83, June 1994.

[12] Tamura, S., Iwano, K., and Furui, S., "A Robust Multi-Modal Speech Recognition Method Using Optical-Flow Analysis", in Extended summary of IDS02, Kloster Irsee, Germany, pp. 2–4, June 2002.

[13] Mase, K., and Pentland, A., "Automatic Lipreading by Optical-Flow Analysis", in Systems and Computers in Japan, vol. 22, pp. 67–76, 1991.

[14] Wojdel, J.C., "Automatic lipreading in the Dutch Language", PhD thesis, 2003

[15] Fleet, D. J., Black, M. J., Yacoob, Y., and Jepson, A. D., "Design and Use of Linear Models for Image Motion Analysis", International Journal of Computer Vision, vol. 36, no. 3, pp. 171–193, 2000.

[16] Martin, A., "Lipreading by Optical Flow Correlation", tech. rep., Compute Science Department University of Central Florida, 1995.

[17] Chițu, A. G., Rothkrantz, L. J. M., Wiggers, P., and Wojdeł, J., "Comparison between different feature extraction techniques for audio-visual", *Journal on Multimodal User Interfaces*, vol. 1, no. 1, pp. 7–20, Springer, March, 2007.

[18] Chițu, A. G., and Rothkrantz, L. J. M., "Building a Data Corpus for Audio-Visual Speech Recognition", *Euromedia2007*, pp. 88–92, April, 2007.

[19] Chițu, A. G., and Rothkrantz, L. J. M., "The Influence of Video Sampling Rate on Lipreading Performance", *12-th International Conference on Speech and Computer (SPECOM'2007)*, pp. 678–684, Moscow State Linguistic University, Moscow, October, 2007.

[20] Wojdeł, J.C., Wiggers, P., and Rothkrantz, L.J.M., "An audio-visual corpus for multimodal speech recognition in Dutch language" In *Proceedings of the International Conference on Spoken Language Processing (ICSLP2002)*, Denver CO, USA, pp. 1917–1920, September, 2002.

[21] Varga, A., and Steeneken, H., "Assessment for automatic speech recognition II: NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems." *Speech Communication*, Vol. 12, no. 3, pp. 247–251, July, 1993.

[22] Wiggers, P., Rothkrantz, L. J. M., "Exploring the Influence of Speaker Characteristics on Word Use in a Corpus of Spoken Language using a Data Mining Approach", *Specom 2007*, pp. 633–638, Moscow State Linguistic University, Moscow, October 2007.

[23] Williams, J. J., Rutledge, J. C., Garsteckiy, D. C., and Katsaggelos, A. K., "Frame rate and viseme analysis for multimedia applications", *Proc. IEEE Works. Multimedia Signal Process*, pp. 13–18, Princeton, 1997.

[24] Williams, J. J., Rutledge, J. C., Aggelos Katsaggelos, K., and Garstecki, D. C., "Frame Rate and Viseme Analysis for Multimedia Applications to Assist Speechreading", *The Journal of VLSI Signal Processing*, Volume 20, no. 1–2, pp. 7-23, November 29, 2004.

[25] Visser, M., Poel, M., and Nijholt, A., "Classifying visemes for automatic lipreading", In *Proceedings of TSD'99*, 1999.