SILENCE DETECTION AND VOWEL/CONSONANT DISCRIMINATION IN VIDEO SEQUENCES

Jacek C. Wojdeł and Leon J. M. Rothkrantz
Knowledge Based Systems Group,
Delft University of Technology,
Zuidplantsoen 4, 2628BZ Delft, The Netherlands
Phone: +31 15 278 8543
Fax: +31 15 278 7141
E-mail: J.C.Wojdel@cs.tudelft.nl, L.J.M.Rothkrantz@cs.tudelft.nl
WWW: `http://www.kbs.twi.tudelft.nl`

ABSTRACT: In this paper we present a set of experiments that were aimed at investigation of feasibility of using artificial neural networks (ANNs) in a lip-reading task. We present here the method for data extraction that is applied on video sequences containing lower half of the face of speaking subject. Further the data is used to evaluate the performance of ANNs in a task of classifying the frames in the video stream into three possible classes: vowel, consonant or silence.

INTRODUCTION

In the recent years human-computer interaction becomes more and more intuitive and human-oriented. That means a growing number of speech processing systems. Systems for automatic speech recognition have a strong limitation: the recognition rate of such systems is still below human recognition rate and deteriorates significantly with the growth of the noise level in the environment.

Human beings use other modality in noisy environment or in case of hearing disabilities: they read lips. It was shown that even well hearing subjects do lip-reading as a part of the speech recognition process. The influence of lip-reading is present even in perfect auditory conditions and it grows steadily with the degradation of the auditory signal (Adjoudani et al., 1997).

Automated lip-reading can be used in combination with audio-based speech recognition to improve the recognition rate of a system (Nakamura et al., 1997). In a little different approach, the visual information can be used to enhance the quality of the speech signal when transmitted on the distance (Girin et al., 1997).

PROCESSING MODEL

In order to perform an automated lip-reading we have to provide the recognition system with some form of the data extracted form the image. One of the possibilities to do so is to use some statistical analysis of the images (Revéret, 1997), another is to capture the geometrical properties of the mouth contour. In the latter case, many different lip-model definitions and many optimization techniques can be used, but the principal idea remains the same: “fit the model as good as possible on the image and measure the model.” (Coianiz et al., 1995; Luettin et al., 1995)

There is however a different approach possible, in which the geometry of the mouth can be represented by an estimation of some of its statistical properties (Wojdeł and Rothkrantz, 2000). In this case, the first proposed step in the process of data extraction from the video-sequence is a *hue-filtering*. The main idea behind this step is that the lips appear on the image as more reddish than the rest of the face and therefore the filter that gives high response in the red part of the color space can be used to determine which pixels in the image belong to the mouth area (see Fig. 2b,c). There is however no constraint on the nature of this filter as long as it properly highlights the lips in the image.

The image obtained after filtering is further treated as a distribution. The mean of this distribution: $[EX, EY]$ approximates accurately and in a stable way the center of the mouth. Using this value, we can

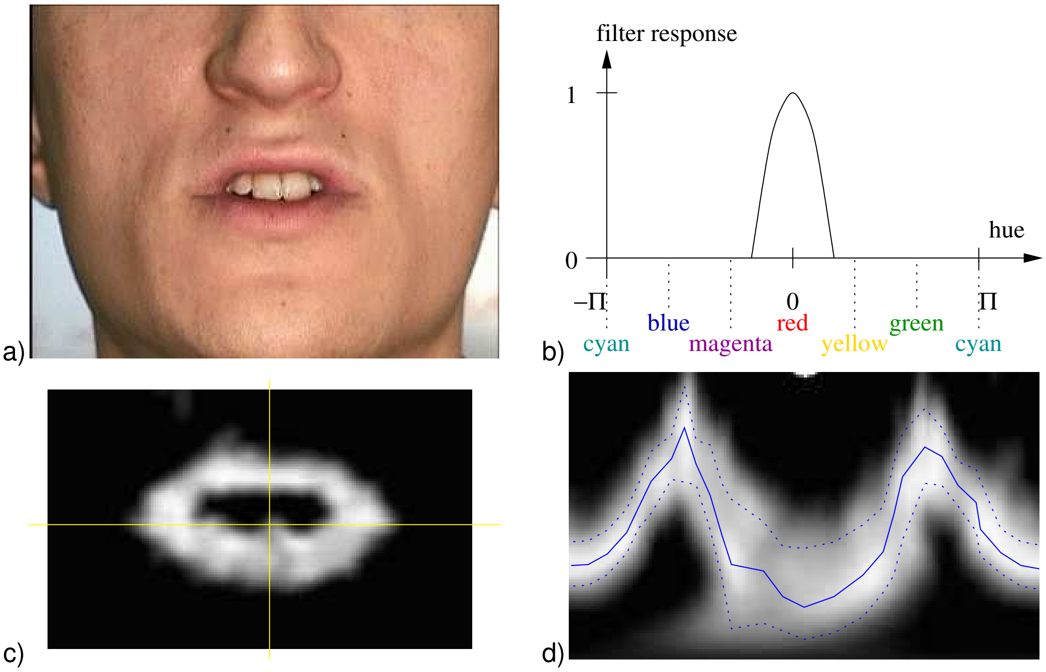

Figure 1: Data extraction process. (a) The original image, (b) hue-based filter, (c) filtered image with mean of the distribution superimposed on it, (d) the distribution in polar coordinates with mean and variance depicted.

transform the image into polar coordinates:

$$J(\alpha, r) = I(EX + r\cos(\alpha), EY + r\sin(\alpha)) \tag{1}$$

The resulting distribution can be seen on the Fig. 2d. From it we can estimate the characteristic values of mean and variance of the conditional distributions for a given angle α:

$$\begin{aligned} \widehat{M}(\alpha) &= \frac{\int_r J(\alpha,r)\cdot r}{\int_r J(\alpha,r)} \\ \widehat{\sigma}(\alpha) &= \frac{\int_r J(\alpha,r)\cdot(r-M(\alpha))^2}{\int_r J(\alpha,r)} \end{aligned} \tag{2}$$

The mean $\widehat{M}(\alpha)$ of such a conditional distributions estimates the distance of the lip from the mouth center. The variance $\widehat{\sigma}(\alpha)$ on the other hand is directly related to the visible thickness of the lip. Those two values can be sampled in a chosen number of angles and the resulting values can form a feature vector that is given as an input to the artificial neural network.

DATA ACQUISITION

As this research is done at Delft University of Technology in the Netherlands, we focused on gathering the data in Dutch language with both native and non-native speakers. In order to capture the wide spectrum of the language, we used a well known set of prompts that were gathered for developing speech recognizers. Those prompts form the POLYPHONE (Damhuis et al., 1994) corpus. Although the corpus contains only auditory data, we used its prompts together with their transcriptions and on this basis recorded our own set of audio-visual data.

In the experiments we asked three people to speak five phonetically rich sentences that together form a basic sets of prompts used in the POLYPHONE corpus. Those sentences are collected in such a way that in each set every phoneme used in Dutch language occurs at least once. Each of the subjects was asked to speak the sentences once at a normal speech rate. Then a subset of three sentences was spoken at a slower rate and finally the remaining two sentences were whispered by each of the subjects. In this way our dataset contains a total amount of 30 sentences of which 15 are spoken at a normal speech rate, 9 at a lower rate and 6 are whispered.

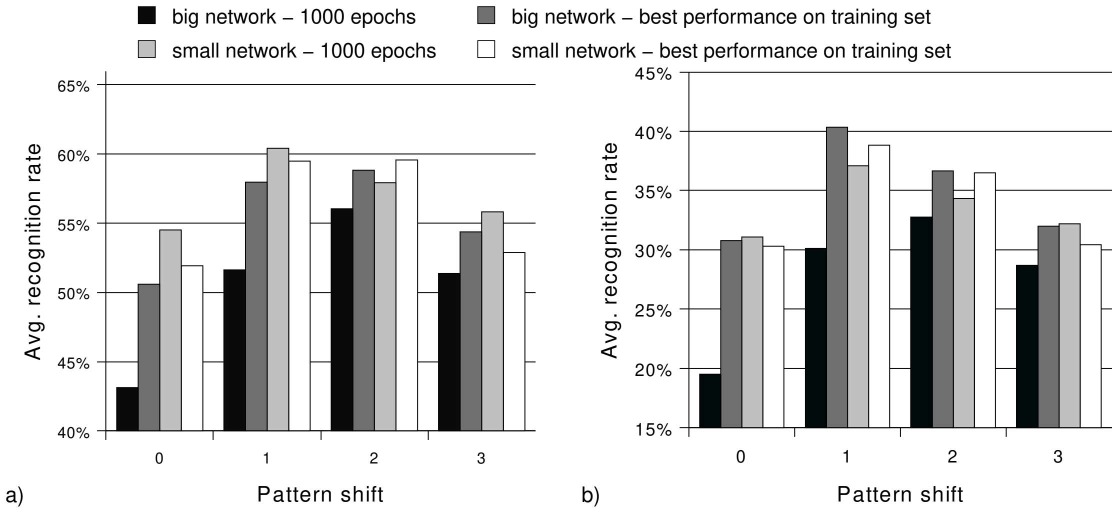

Figure 2: Changing recognition rate when shifting output patterns. a) recognition rate for all outputs, b) rate excluding silence detection output. Two Elman Hierarchical networks (`36-20-40-20-3` and `36-10-10-3`) were trained to either fixed number of epochs or to the best performance on training set.

All utterances were recorded with standard CCD video-camera in normal illumination conditions. The camera was focused on the lower part of the subject's face (see Fig. 2a). We didn't use any artificial mark-up or feature enhancements. After cutting out the fragments containing each of the sentences, we obtained a set of 5576 video frames which at 25 fps. frame rate gives about 220s of video recording. Although the amount of date gathered at this stage is not sufficient for developing a lipreading system, it is enough for our goal.

DATA LABELING

The labeling of the data was simplified by the fact that we already had the transcriptions of the sentences in our dataset. What we had to do was assigning the consecutive labels from the transcription to the appropriate video-frames. We used both visual and auditory information in order to place the labels appropriately, and followed a set of rules related to speech production process. All the vowels were labeled in frames that captured the mouth as being maximally stretched during phoneme pronunciation. The consonants with clear visual point of pronunciation (such as `p` or `l`) were labeled on the frames in which appropriate visual occurrence was seen. Other consonants were labeled to the frame in the middle of the audibly recognizable pronunciation. Example rules used in data labeling were in the following form:

- `[p]` - the last video-frame before opening the mouth
- `[l]` - the video-frame in which the bottom side of the tongue is visible between the teeth
- `[A]` - the video-frame in which the mouth is maximally stretched when pronouncing the phoneme
- `[t]` - the video-frame in the middle of an audibly recognizable phoneme `t`

In cases when the occurrence of the phoneme could not be objectively located from neither auditory nor visual channel (often for consonants such as `h` or `k`), the label was put in the middle frame between the neighboring labels. In cases when the same rule could have been applied to more than one frame, only the one in the middle was labeled (which happened often in case of vowels).

The rules that we used for labeling were designed in a way that simplifies the process of manual placing of the labels. There was however some concern that they may in fact produce a pattern set that would not be very suitable for training of a neural network. Some occurrences such as opening of the mouth span across several frames and if labeled too early would in fact force the network to predict given viseme instead of detecting it. One possible solution of this problem is to shift the resulting labels so

Table 1: Example recognition rates for different NN architectures

Architecture	Recognition rate	V/C only
Feed forward	48.9%	44.4%
TDNN	49.8%	47.7%
Elman	63.4%	70.6%

that they appear later in time. In this way phonemes labeled too early in the sequence can be shifted to appropriate places. The phonemes that were in appropriate places before shifting procedure would be then too late, but this is not a big problem as long as the input of the network contains some information on previous frames.

From earlier research we expected that recursive networks would perform better than other neural architectures, so to investigate the influence of pattern shifting we used four Elman neural networks with varying sizes and trained them with patterns shifted from 0 to 3 frames. The resulting recognition rates can be seen on Fig. 4. The results clearly indicate that shifting the patterns by one frame gives an optimal network performance.

TRAINING PROCEDURE

In order to do vowel/consonant discrimination, we prepared the dataset in the following way. From each video-frame a data vector was extracted using the technique described in the earlier section. In order to achieve mouth-size independency the obtained vector was scaled so that all of the values in it fit the $\langle 0, 1\rangle$ interval:

$$V_i^* = \frac{V_i}{\max\limits_{j=1\ldots N} V_j} \quad i = 1 \ldots N \tag{3}$$

The stream of those vectors formed an input for the neural network. We used 18-point sampling of the $\widehat{M}(\alpha)$ and $\widehat{\sigma}(\alpha)$ functions which proved to be adequate in our earlier experiments (Wojdeł and Rothkrantz, 2000). Therefore the input pattern contained in total 36 values for each video-frame ($N = 36$ in the above formula.)

The appropriate output pattern was formed from three values each representing one of the classes:

0 - silence the frames that contain no utterances should be classified in this way,

v - vowel the frames with labels `I e: E E: A @ i O Y y u 2: o: 9 9: O: a:`

c - consonant the frames with labels `f v w s z S Z p b m g k x n N r j t d l`

As typically done, we constructed the outputs in such a way that the value 1 is put in the video-frame where the given class was labeled, with neighboring frames taking lower values of 0.5 and 0.25 depending on the distance to the labeled one. An example output data can be seen on Fig. 6 (note that for clarity reasons, one output; *silence* is not depicted there). The output patterns prepared in this way were then shifted one frame forth in order to avoid the problems of too early labeling as it was discussed earlier in this paper.

From the whole set of 30 sentences we chose 3 of them to form our test set and used the remaining 27 for training the networks. The test sentences were chosen so that we had one sentence from each of the subjects, with all *normal*, *slow* and *whispered* sentences present in this set.

In all our experiments we used the Stuttgart Neural Network Simulator (SNNS) version 4.1 running on a Sun Ultra1 workstation. Using this tool we trained several different neural network architectures. The architectures used here varied from typical feed forward, back propagation trained neural network, through Time Delayed neural network to simple recursive neural networks such as Jordan Neural Network and Elman Neural Network.

RECOGNITION RESULTS

The results of comparison between different NN architectures were consistent with our previous experiments in this field (Wojdeł and Rothkrantz, 2000). The typical feed forward neural networks are not

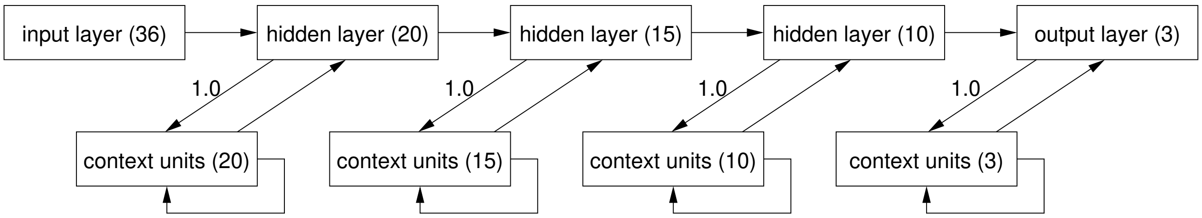

Figure 3: Extended Hierarchical Elman network used in our experiments that reached the best performance

Table 2: Recognition results for `36-20-15-10-3` Elman network

	Network response			
Target	0	c	v	-
0	70.5%	14.3%	3.8%	11.4%
c	2.6%	72.2%	19.5%	5.2%
v	0.0%	26.5%	67.3%	6.1%
-	1.7%	45.2%	12.2%	40.9%

really suitable for this task and so they score poorly in overall recognition rate. The Time Delayed neural networks perform a bit better, but still their recognition rate is significantly lower than that of recurrent neural networks (see Tab. 6)

The Elman Hierarchical Neural Network constructed from three hidden layers with respectively 20, 40 and 20 neurons in each of them (see Fig. 6) proved to be the most efficient in vowel/consonant discrimination. The overall results of recognition are summarized in Tab. 6. Together with the three classes that can be recognized by the network (*0*, *v*, *c*) this table contains also a fourth column labeled (–), which corresponds to the output patterns that should not be classified as any of the three classes. Such patterns occur in our dataset for example in slow speech between some phonemes that last for a long time. According to our labeling method, if a phoneme lasts for longer than 5 frames ($\sim$200 ms), it will have a single output peak associated with it in the middle of the utterance and some unspecified frames around it. In the experiments we assumed that the network response was unspecified if none of the network's outputs had a higher activation than a threshold value of 0.1

In case of the *silence* detection, the achieved recognition rate of about 70% does not give the real impression on how well the speech is being detected. It can be inspected that almost all of the incorrectly recognized frames appear at the speech onset/offset points, which means that the actual silence detection is either to early or too late in correspondence to the labeled data. The amount of 30% incorrectly recognized frames is equivalent to about 0.5s (12 video frames) average mismatch in detecting onset or offset of the speech.

Figure 6 presents an example flow of network outputs for words “mijn vleugel” from one of the test sentences. The first two rows of numbers in this figure represent the target outputs with boxes around the dominating values. The second two rows show the output of the network for the same frames. The lowest row of symbols is the transcription of the words in SAMPA notation. As it can be seen the

Target output

Cons.	0.50	1.00	0.50	0.50	1.00	0.50	1.00	0.50	1.00	0.50	0.25	0.50	1.00	0.50	0.25	0.50	1.00	0.50
Vowel	0.25	0.50	1.00	1.00	0.50	0.25	0.00	0.00	0.25	0.50	1.00	0.50	0.25	0.50	1.00	0.50	0.25	0.00

Network response

Cons.	0.49	0.37	0.14	0.24	0.47	0.48	0.56	0.55	0.51	0.15	0.06	0.23	0.28	0.48	0.52	0.46	0.50	0.37
Vowel	0.09	0.17	0.56	0.54	0.30	0.24	0.17	0.24	0.26	0.46	0.49	0.64	0.36	0.16	0.21	0.39	0.19	0.17
	-	m	E	i	n	-	v	-	l	-	2:	-	G	-	@	-	l	-

Figure 4: Example recognition flow for words “mijn vleugel” (“my piano”) in one of test sentences.

network output is not ideal and for example phonemes [G] and [@] are not recognized properly. The rest however seems to be recognized fairly well.

CONCLUSIONS

The proposed method of data extraction from the video sequence proves to be feasible in vowel/consonant discrimination. From the presented results we may conclude that the proposed approach in combination with partial recurrent neural networks can be used in developing a lip-reading system. Despite the small size of the dataset used in the experiments, the recognition rate for the test set is relatively high, which indicates that the proposed data representation is suitable for lipreading tasks.

In the near future, bigger corpus of video recordings must be gathered in order to proceed with the development of the system for continuous lip-reading. Having a significantly larger set of recordings (and so also prompts) will allow also investigation in how much the results depend on the contextual information extracted by Elman network.

REFERENCES

Adjoudani, A., Guiard-Marigny, T., Goff, B. L., Reveret, L. and Benoit, C. (1997), A multimedia platform for audio-visual speech processing, *in* Kokkinakis et al. (1997).

Coianiz, T., Torresani, L. and Caprile, B. (1995), 2D deformable models for visual speech analysis, *in* Stork and Hennecke (1995).

Damhuis, M., Boogaart, T., In 't Veld, C., Versteijlen, M., Schelvis, W., Bos, L. and Boves, L. (1994), Creation and analysis of the Dutch polyphone corpus, *Proceedings of the International Conference on Spoken Language Processing, ICSLP'94*, Yokohama, Japan, pp. 1803–1803.

Girin, L., Feng, G. and Schwartz, J. (1997), Noisy speech enhancement by fusion of auditory and visual information: a study of vowel transitions, *in* Kokkinakis et al. (1997).

Kokkinakis, G., Fakotakis, N. and Dermatas, E. (eds) (1997), *Proceedings of ESCA, Eurospeech97*, ESCA, Rhodes, Greece.

Luettin, J., Thacker, N. A. and Beet, S. W. (1995), Active shape models for visual speech feature extraction, *in* Stork and Hennecke (1995).

Nakamura, S., Nagai, R. and Shikano, K. (1997), Improved bimodal speech recognition using tied–mixture HMMs and 5000 word audio–visual synchronous database, *in* Kokkinakis et al. (1997).

Revéret, L. (1997), From raw images of the lips to articulatory parameters: A viseme–based prediction, *in* Kokkinakis et al. (1997).

Stork, D. G. and Hennecke, M. E. (eds) (1995), *Speechreading by Humans and Machines*, Vol. 150 of *NATO ASI Series, Series F: Computer and Systems Sciences*, Springer Verlag, Berlin.

Wojdeł, J. C. and Rothkrantz, L. J. M. (2000), Visually based speech onset/offset detection, *Proceedings of Euromedia2000*, Antwerp, Belgium.