VISUALLY BASED SPEECH ONSET/OFFSET DETECTION

Jacek C. Wojdeł and Leon J. M. Rothkrantz
Knowledge Based Systems Group,
Faculty of Information Technology and Systems,
Delft University of Technology,
Zuidplantsoen 4
2628 BZ Delft, The Netherlands
E-mail: J.C.Wojdel@cs.tudelft.nl, L.J.M.Rothkrantz@cs.tudelft.nl

KEYWORDS

Speech Recognition Systems, Lip-reading, Human-Computer Interfaces

ABSTRACT

In this paper we present a new model for video sequence processing that is aimed at extraction of the data describing the mouth shape in the image. The proposed technique can be used in any lip-reading related application. One of such applications, chosen for our experiments, is speech onset and offset detection. The experiments that are described here are aimed at detecting boundaries of the speech on the basis of visual modality only. This paper describes the data processing technique and the results of the experiments.

INTRODUCTION

In the recent years human-computer interaction becomes more and more intuitive and human-oriented. There is 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 modalities 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).

In order to construct a lip-reading system, some image processing technique must be used that will extract the important information concerning mouth movements from a video sequence. Techniques used for this task range from purely statistical analysis of the images (Revéret, 1997) to sophisticated model matching methods (Coianiz et al., 1995), (Luettin et al., 1995).

In this paper we will show another method, that allows extracting of the information from the mouth images. The method for processing image data will be described together with the results of the experiments.

SPEECH ON/OFF-SET LOCALIZATION

The processing interval of an automatic speech processing system is defined by the presence of speech. It is reasonable that such a system should be active and processing the incoming audio-data only when there is somebody actually speaking to it. Detection of this simple occurrence isn't always that easy. For example using the volume level to determine the beginning and the end of the utterance may be very misleading in a noisy environment.

In order to detect this activity time interval, another modality can be used. For example street information kiosks are often equipped with a weight sensor in front of the information panel. In this way the processing is activated only when somebody is standing in front of the display.

One of the possible modalities that can be used is of course the visual information about lip movements if they were tracked in the environment. In this way the activation of the speech recognition module would be triggered only when the registered movement of the lips appears to relate to the speech production process. We will further present the experiments that were conducted with this problem in mind.

DATA ACQUISITION

We did experiments with multiple speakers (only one of them was not native Dutch). The subjects were seated in front of the camera and asked to speak a set of five sentences. The camera was then adjusted so that only the lower part of the subject's face was in the field of view. The sentences were chosen from the POLYPHONE (Damhuis et al., 1994) corpus. The sentences in this corpus are sentences used in everyday life and grouped in sets of five so that all of the phonemes used in the Dutch language appear at least once in the set.

After acquisition of the video sentences, we manually labeled speech onset and offset points on the basis of the auditory signal only. Those boundary points were used to label all the frames in video sequences as either 1 for silence or 0 for speech. These labels were later used as target output values for the silence recognition system.

MODEL

One of the possible ways to describe the shape of the lips, is by using some measurements of their contour. In such a case, different lip-model definitions and different optimization techniques are 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)

We would like to propose a different approach that does not depend on the definition of the model, but uses an estimation of the shape parameters directly. At first we have to define the measurements that are needed in order to describe the shape of the mouth accurately. The measurements that we propose are as follows:

- the distance of the lips from the center of mouth,
- the thickness of the lips.

Both of them could be measured in any point around the mouth and they together represent the mouth shape accurately. Although both of the measures are not available directly from the image, they can be estimated and those estimations can be used as input data for the recognition system.

The first step in extraction of the data vector from the video frame was *hue-filtering*. This filtering was done as earlier presented in (Coianiz et al., 1995). 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. The

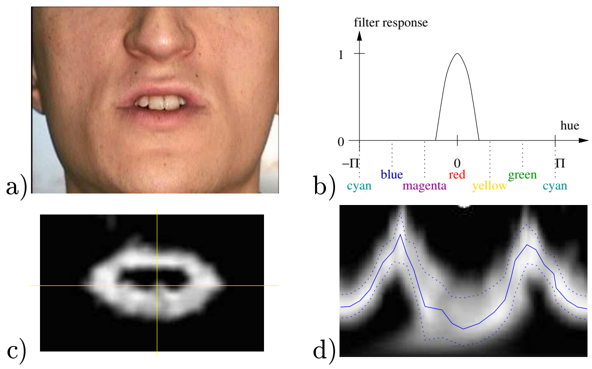

Figure 1: Data extraction process. (a) The original image, (b) hue-based filter, (c) filtered image with mean of the distribution superimposed on it, (d) the distribution in polar coordinates with mean and variance depicted.

filter was realized using a simple paraboidal response for a hue component of each pixels color:

$$F(h) = \begin{cases} 1 - \frac{(h-h_0)^2}{w^2} & , \quad |h - h_0| \leq w \\ 0 & , \quad otherwise \end{cases} \tag{1}$$

The actual parameters of w and h_0 in Eq. 1 depend on the color reproduction of the used camera. We used $w = 0.23$ and $h_0 = 0$, which for our setup gave a fairly good segmentation of the mouth shape from the rest of the face. As a result of this filtering we obtain an image that can be seen as a distribution $I(x, y)$ of *lipness* in the image (see Fig. 1c). Furthermore, the mean of this distribution: $[EX, EY]$ approximates accurately and in a stable smooth way the center of the mouth. Using this value, we can transform the image into polar coordinates with the origin of the transformation defined by the mean value of the distribution:

$$J(\alpha, r) = I(EX + r\cos(\alpha), EY + r\sin(\alpha)) \tag{2}$$

In the current experiments, the orientation of the polar coordinate system was defined just by the horizontal axis of the image. This was done because of the assumption that the speaker does not significantly rotate his/her head during the recordings. However in case of the recordings without this constraint, the covariance matrix of the distribution can be used in order to find the most stretched direction of the mouth. The resulting mouth-shape representation can be seen on the Fig. 1d.

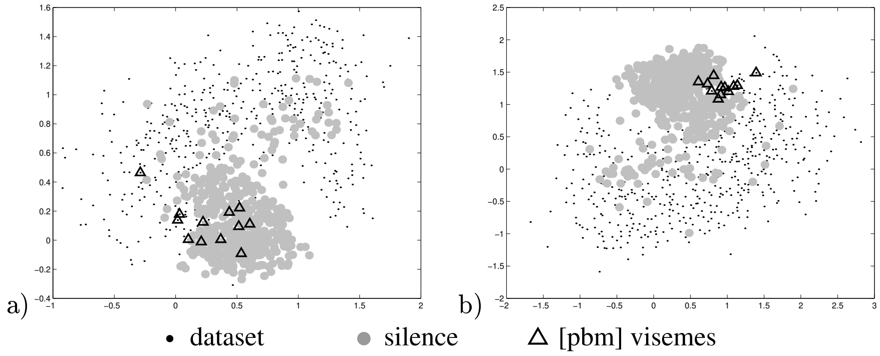

Figure 2: Sammon mapped dataset: (a) vectors extracted from single frames, (b) concatenated data vectors from three consecutive frames.

If we recall now the measurements that we need in order to represent the mouth shape accurately, one can see that they can be estimated by respectively mean value and variance of the distribution for a given angle:

$$\begin{aligned} \widehat{M}(\alpha) &= \frac{\sum_r J(\alpha,r)\cdot r}{\sum_r J(\alpha,r)} \\ \widehat{\sigma}(\alpha) &= \frac{\sum_r J(\alpha,r)\cdot(r-M(\alpha))^2}{\sum_r J(\alpha,r)} \end{aligned} \tag{3}$$

The mean $\widehat{M}(\alpha)$ of such a conditional distribution is an estimate of the distance of the lips from mouth center. The variance $\widehat{\sigma}(\alpha)$ on the other hand is directly related to the visible thickness of the lip.

DATA VISUALIZATION

In order to get an insight in the high dimensional data that is obtained from the video-sequence we have to use some dimension reduction techniques. We decided here to use the Sammon mapping (Ripley, 1996) as it gives a more accurate representation of the small distances in the dataset than other methods (see (Ripley, 1996)). Sammon mapping is a technique in which each (highly dimensional) vector from the dataset is explicitly mapped to some vector of a lower dimensionality space. The mapping is done in the way that minimizes the stress measure, which is defined as follows:

$$E = \sum_{i \neq j} \frac{\left(d_{ij} - d'_{ij}\right)^2}{d_{ij}} \tag{4}$$

where d_{ij} is the distance between i-th and j-th vector in the dataset and d'_{ij} is the distance between mappings of the i-th and j-th point. By minimizing the formula 4, the obtained mapping preserves distances and therefore also clustering properties of the initial dataset, as good as possible.

There is however one issue that needs to be inspected first, when applying this technique to the data coming from a video-sequence. As seen on the Fig. 1,

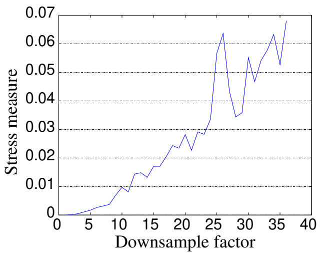

Figure 3: Distance matrix distortions

the last step of the image processing gives us the values of two functions; i.e. mean and variance of distribution for a given angle. Those functions must be sampled to represent the data vector that will be used in further processing. We have to chose an appropriate number of sampling points in order to achieve robust, yet sufficiently accurate representation of the data.

In order to find the appropriate sampling rate, we first extracted the feature vectors with 180 uniformly distributed sampling points for each of both functions. This dataset with its distance matrix was used as a reference set. The reference set was then downsampled with different rates and obtained distance matrices were compared with the reference matrix using formula (4) (see Fig. 3). As it can be seen, the distortions in the dataset grow smoothly up to the factor 10 (resulting in 36 dimensional feature vectors). After this point, the distortions become rather chaotic, which means that the distances in the dataset are not preserved properly. These results motivate our choice of 36 dimensional vectors (18 uniformly distributed sample points for mean and variance) as the data vectors for further processing.

After that, we applied the Sammon mapping on the dataset (see Fig. 2). Even straightforward mapping of vectors coming from the video-sequence reveals clustering of the frames that were labeled as silent (see

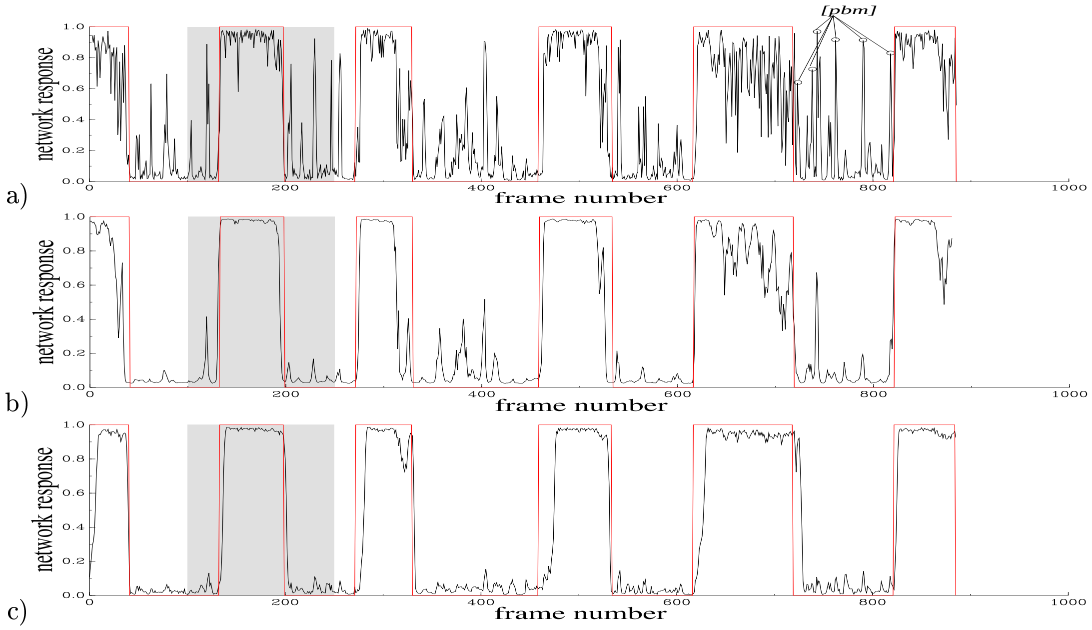

Figure 4: The output graphs for a different neural network architectures. (a) FFNN, (b) TDNN, (d) JNN. Straight lines depict target outputs, grey area represents the part of the sequence that was used as a test set.

Fig. 2a). Most of the *silent* frames contain the image of a closed mouth. There are however also such frames for the phonemes like p, b and m at some point of their pronunciation. Therefore there is no possible way of differentiation between *[pbm]* viseme and silence on the basis of a single frame. Fig. 2a shows therefore *[pbm]* frames scattered around in the silence cluster.

In order to introduce the time dependencies in the feature vector, we can combine the data from some number of different frames. Figure 2b shows the result of a concatenation of 3 feature vectors from consecutive frames from the video-sequence (very nicely clustered *[pbm]* viseme; just on the edge of silence cluster).

SILENCE LOCALIZATION

Visualization of the data proved that the silence periods can be separated from the rest of the sequence. Therefore we conducted several experiments with Neural Network-based classification of those segments. All of the experiments were done using the Stuttgart Neural Network Simulator (SNNS) version 4.1 running on a Sun Ultra1 workstation.

For training and validation of the neural networks we used different video sequences with varying speech rates. All the extracted feature vectors were labeled 0 or 1 (*non-silent*, *silent*) according to audibly found boundaries of spoken sentences. From the full dataset we chose about 10% of the vectors to form the test set and used the rest for training. The training set was chosen so that it consisted of several intervals containing both beginning and ending of silence together with some non-silent frames. We measured the fitness of all the networks with mean square error of their response for a single epoch.

Fig. 4 shows the example responses of the networks for a single video-sequence containing 5 sentences. The video sequence was chosen so that it contained one of the sentences that were put in test set.

In the experiments we used three different NN models:

- **Feed Forward Neural Network** (FFNN)
- **Time Delayed Neural Network** (TDNN)
- **Jordan Recurrent Neural Network** (JNN)

A fully connected feed forward, back propagation trained neural network was fed with a single frame feature vector, therefore it had 36 input units and the single output unit. As it can be seen on Fig. 4a, the performance of the network is far from being acceptable. Most notably, there are a lot of frames within sentences that are recognized as silence. There is however one interesting fact that indicates that the proposed data processing technique is suitable for lipreading. The network cannot classify the silent intervals properly, but it tends to highlight *[pbm]* viseme in the sentences.

A time delayed neural network allows exploitation of time dependencies in the data set by using a so-called *perception window*. A perception window covers several consecutive input vectors and slides along the input sequence. The length of the window is called *delay*. Moreover, the hidden layers can also be time-delayed, which allows passing of the activation patterns from earlier time steps through the network. This network gives a reasonable results with very little variations of the output in silent periods and reasonable results for spoken ones (see Fig. 4b). Those improvements come directly from time dependency finding capabilities of TDNNs.

In a Jordan recurrent neural network the time dependency is represented by the recurrent nature of the network itself. The input to the JNN is only based on a single frame, the time related information is preserved within the network through *context neurons*. Context neurons are neurons that take as an input the output of the whole network and of themselves from a previous time step. The single hidden layer is fed with activation of both input and context neurons. In our case there is only one output neuron, and therefore also only one context neuron. This network architecture gave by far the most superior results (Fig. 4c). Not only it trained much faster than the rest of the networks, but it also gave the smoothest and most accurate output. There are only two possible disadvantages of using JNN in this task. One is that according to our experiences, JNN can be very easily overtrained in which case it starts jumping on the error surface. The other one is that it is always a little bit late when changing from speech to silence (as seen in Fig. 4c), which presumably comes from the fact that the output of the network heavily depends on the loop between context neuron and the output and the change in the activation of this neuron is not propagated in a single data pass.

CONCLUSIONS

As it was shown, using appropriate tools, the proposed mouth shape representation can be successfully used in problems related to lip-reading. There are some advantages of using the proposed method over classic contour tracking methods. Namely, the proposed method does not contain any iterative process which convergence could depend on starting conditions (contrary to active contour or template matching). Furthermore the method is noise insensitive thanks to the use of average features of the incoming data. However even with those capabilities, the proposed image processing technique would be useless if it didn't preserve the lip-reading related features of the incoming video. From data visualization it can be seen that the processed data preserves vital dependencies of the lip movements in the video. Moreover, using artificial neural networks in combination with our technique provides a speech onset/offset detection system.

For the future research, we will apply this image processing technique to the lip-reading itself. The next experiments that are planned are with viseme detection in the sequence and then further with visually based speech recognition. It will most probably enforce some adjustments to the technique together with use of more sophisticated recognition (classification) tools. Of them the most suitable seem to be Hidden Markov Models, which are already widely used in speech recognition.

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. In *Proceedings of the International Conference on Spoken Language Processing, ICSLP'94*, pages 1803–1803, Yokohama, Japan.

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., editors (1997). *Proceedings of ECSA, Eurospeech9*, Rhodes, Greece. ESCA.

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).

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*, chapter Multidimensional scaling, pages 305–311. Cambridge University Press.

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