PHONEME RECOGNITION OF MEETINGS USING AUDIO-VISUAL DATA

Petr Motlíček, Lukáš Burget, Jan Černocký, Igor Potůček

Brno University of Technology
Faculty of Information Technology
Božetěchova 2, Brno, 612 66, Czech Republic

ABSTRACT

The movement of speaker's faces are known to convey visual information that can improve speech intelligibility especially in case of somehow corrupted or noisy data. Therefore, availability of visual data could be exploited to enhance automatic speech recognition task. This paper demonstrates the use of visual parameters extracted from video for automatic recognition of context-independent phoneme strings from meeting data. Encouraged by the good performance of audio-visual systems utilized to work with "visually clean" data (limited variation in the speaker's frontal pose, lighting conditions, background, etc.), we investigate their efficiency in non-ideal conditions which are introduced by meeting audio-visual data employed in our experiments. A major issue is the phoneme recognition task based on combination of the audio and visual data so that the best use can be made of the two modalities together.

1. INTRODUCTION

Recently, the interest has grown in multimodal speech recognition system incorporating several audio sources as well as visual information. One of the applications is the automatic analysis of meetings (interaction between humans), where several audio and video sources are available. Automatic speech recognition (ASR) of meeting data is an extremely challenging task, due to nature of the audio capture and the conversational informality.

Information from the speaker's mouth region has been shown to improve the accuracy and noise robustness of Automatic Speech Recognition (ASR) systems [2, 4]. However, up-to-date systems use audio-visual speech features recorded under ideal lighting conditions. Those video recordings contain high-resolution video of subjects' frontal face, minimal changes in head positions and constant backgrounds. Insufficient research has been done in audio-visual ASR in real meeting conditions, where in addition to possibly noisy audio, the quality of visual channel is poor. But if visual information may have been used in ASR systems, we need to demonstrate its benefits in such non-ideal conditions. In this paper, we describe preliminary experiments with audio-visual ASR system with visually challenging meeting data from real multi-party conversations recorded at IDIAP [5]. Although, the original video recordings contain the whole meeting scenario, our experiments utilize only video streams containing tracked objects' heads of analyzed meetings.

This research has been partially supported by EC project Multi-modal meeting manager (M4), No. IST-2001-34485, by EC project Augmented Multi-party Interaction (AMI), No. 506811-AMI, by Grant Agency of Czech Republic under project No. 102/02/0124, and by post-doctoral grant of Grant Agency of Czech Republic No. GA102/02/D108.

2. EXPERIMENTAL DATA

The data used in our experiments is recorded in the IDIAP smart meeting room [5], equipped with synchronized multichannel audio-visual recording facilities. The meeting room is configured for full audio-visual recording of meetings with up to 6 participants. Recordings are generated by two cameras, each capturing front-on view of two participants including the table region used for note-taking. These cameras generate output PAL quality video signals,which are recorded onto separate MiniDV cassettes. Participants wear lapel microphones, and an eight-element circular equispaced microphone array is centrally located on the meeting table.

3. ACOUSTIC FEATURE EXTRACTION

In applications involving multi-party conversations, it may be possible to acquire the speech using microphone arrays, which provide the ability to discriminate between sounds based on their source location. This directional discrimination can enhance a signal from a given location. Such signal processing operation was done using beam-forming algorithm at IDIAP [1, 7].

The parameters of audio speech signal are well-known Mel-filterbank log energies (23 banks, window size 20ms, audio frame-rate $F_{afr} = 100$Hz). These parameters are extracted from beamformed audio recordings sampled at 16kHz.

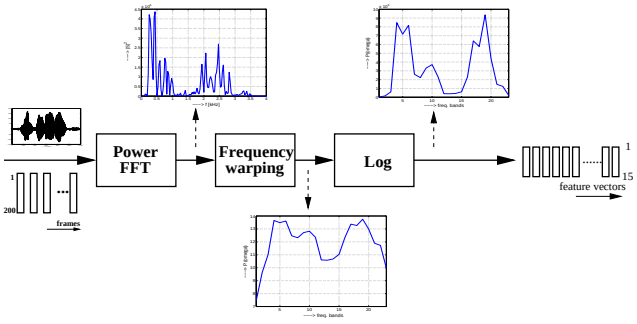

Fig. 1. Extraction of acoustic features.

4. VISUAL FEATURE EXTRACTION

Prior to extraction of visual features, we need to process the original video stream data in order to detect and track faces (heads) of humans (objects) in meetings. However, the design of multiple-people trackers in arbitrary backgrounds and stable in the long-term is to-date unresolved, although visual tracking has been intensively studied and considerable progress has been obtained. Detection and tracking of faces and body parts represent fundamental steps in our meeting analysis task. The requirements on visual tracking algorithm are the robustness against real-world conditions present in meetings, like variations in objects appearance and pose due to natural unrestricted motion and changing lighting conditions, and the presence of multiple self-occluding objects. Furthermore, the method has to be efficient in computational terms, given their extensive use by later recognition tasks. The method employed is based on a skin color detection [8]. In order to predict and identify the trajectory of moving objects in meeting recordings, the algorithm employs Kalman filtering. The skin color is known to be the key feature for hands and head detection. Such approach is mainly advantageous due to low computational cost. On the other hand, the reliability of correct detection of head poses is poor, due to dependence of the skin-tone color on the lighting conditions. Normalized RG-color space derived from RGB values provides good solution to the problem of varying brightness. The visual input in our experiments is a video stream which is supposed to contain sequence of head poses of one human object. The video frame-rate $F_{vfr} = 25$Hz and the input resolution 70×70 pixel region is obtained for every video frame. Practically, each video frame contains the whole speaker's head including the hair and neck.

4.1. Average brightness

In initial experiments, visual features based on average brightness of a region-of-interest (ROI) are extracted. The ROI is theoretically expected to be speaker's mouth, as seen

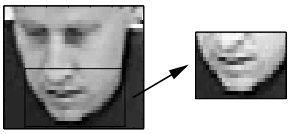

Fig. 2. Detection of ROI (mouth area) using correlation technique.

in Fig. 2, which should carry the most of variability caused by speech (it is of size 16×34 pixels). Such ROI is obtained using correlation-based mouth detector. This detector searches in each video frame the most similar pattern to the average mouth pattern. The correlation is performed on the level of video frames processed by standard edge detection algorithm. In this algorithm we applied the Sobel method [10], which is supposed to find edges using the Sobel approximation to the derivative. It returns edges at those points, where the gradient of intensity in the input image is maximum. From the ROI described by normalized brightness values, we compute the average intensity providing one visual feature for each video frame.

4.2. DCT coefficients

Subsequently, a two-dimensional, separable DCT is applied to the ROI and 16 lowest-order DCT coefficients are retained. The reasons for widespread use of DCT in feature extraction as well as in image compression [10, 11], are the high compaction of the energy of the input signal onto a few DCT coefficients and the availability of a fast implementation of the transform, similar to the FFT. However, the DCT is not shift invariant, thus performance depends on a precise tracking of the ROI.

4.3. Optical flow analysis

In the second approach, we attempted to avoid the use of the algorithm searching for ROI (position of mouth in an input image). The main difficulty is that the resolution of input images containing the whole speaker's head is low. Thus, algorithms detecting ROI are not sufficiently reliable.

The optical flow analysis assigns to each pixel of a video frame a flow velocity vector. This vector specifies a direction and a velocity of movement of objects in the scene at the position of given pixel. If we suppose that the variances in the optical flow vectors are mainly due to movement of speaker's mouth during the speech, optical flow analysis may do an interesting job [13]. In our experiments, we use the Horn-Schunck optical flow analysis [12]. Theoretically, optical flow is the distribution of apparent velocities of movement of brightness patterns in an image. It can give important information about the spatial arrangement of the objects viewed and the rate of change of this arrange-

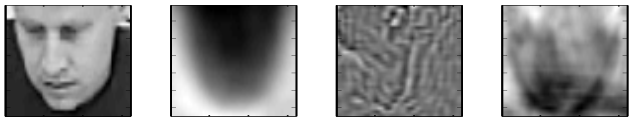

Fig. 3. Application of PCA-LDA analysis: from left-to-right an original video frame (tracked head position), first PCA basis, first PCA-LDA basis, reconstructed video frame from first 10 PCA-LDA basis.

ment [12]. Consequently, discontinuities in the optical flow can help in segmenting images into regions that correspond to different objects. The analysis takes as the input the sequence of video frames. Each frame is described by image brightness (denoted $E(x, y, t)$) at a point (x, y) in an image plane at time t. We assume that the brightness of each point is constant during a movement for a very short time. Thus, the equation is follows:

$$\frac{\mathrm{d}E}{\mathrm{d}t} \backsimeq \frac{\partial E}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}t} + \frac{\partial E}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}t} + \frac{\partial E}{\partial t} = 0. \quad (1)$$

If we let: $\frac{\mathrm{d}x}{\mathrm{d}t} = u$ and $\frac{\mathrm{d}y}{\mathrm{d}t} = v$, then a single linear equation is obtained:

$$E_x u + E_y v + E_t = 0. \quad (2)$$

The vectors u and v denote apparent velocities of brightness constrained by this equation. The additional constraint which minimizes the square magnitude of the gradient of the optical velocity is used, since we cannot use only Eq. 2 to determine flow velocity (u, v). Practically, optical flow velocities are calculated from a pair of connected images using several iterations.

4.4. Estimation of ROI using a lip detector

We were also dealing with an algorithm detecting lip positions (as the center of mouth) in the given image frame, which uses edge detection and color filtering for noise reduction and enhancement of the desired recognition of lips. First, bursts of red colored pixels are found. Then the largest red area, which is considered to be a lip position, is found using a "seed algorithm". The particular steps are [9]:

- Detection of red pixels of the face (this operation processes an input image in order to correctly locate mouth's pixels).
- Selection of the largest red area with a "seed algorithm". First, several erosions on the binary image are applied until a few white pixels remain. Then, the white pixels are used as a "seed" which has to be extended to all the surrounding white pixels on the binary image.

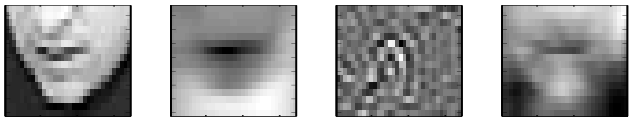

Fig. 4. Application of PCA-LDA analysis: from left-to-right a mouth area (ROI) detected using correlation technique, first PCA basis, first PCA-LDA basis, reconstructed video frame from first 10 PCA-LDA basis.

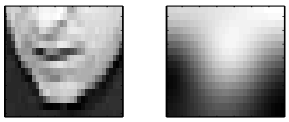

Fig. 5. Compression of image using DCT: from left-to-right a mouth area (ROI) detected using correlation technique, its reconstruction using the 10 lowest DCTs.

4.5. Methods for dimensionality reduction and better class separability

The challenging problem in visual processing is the necessity to process highly dimensional data compared to audio signal. Therefore, some kind of transformation (i.e., DCT) needs to be applied. We were inspired by previous work of Potamianos et al. [16], which attempts to reduce dimensionality of input video frames and improves discriminability between classes while taking into account video training data. Widely used method to improve discriminability among classes is Linear Discriminant Analysis (LDA). However, its application directly on input video frames requires computation of large statistics (within-class and across-class covariance matrices) that would have to be estimated (in case of 32×64 video frames, the covariance matrices would be of size 2048×2048). Thus, first dimensionality reduction is provided by Principal Component Analysis (PCA). Execution of the following LDA is then computationally much less expensive.

5. EXPERIMENTAL SETUP

The audio-visual speech database collected at IDIAP [14] has been used. Our main goal is the phoneme classification of audio-visual data based on Neural Nets (NN). For experimental purposes, data is split into three sets: training, (CV) cross-validation (together 41 minutes) and testing (9 minutes). First two sets are used to train NN. Then, testing data is forward passed through such NN.

Acoustic features (Mel-filterbank log-energies) are computed with $F_{afr} = 100$Hz. Derivation of visual features is the major part of our experiments. Each input video sequence is at the beginning processed by head tracking algorithm. Such visual data represents input to the following

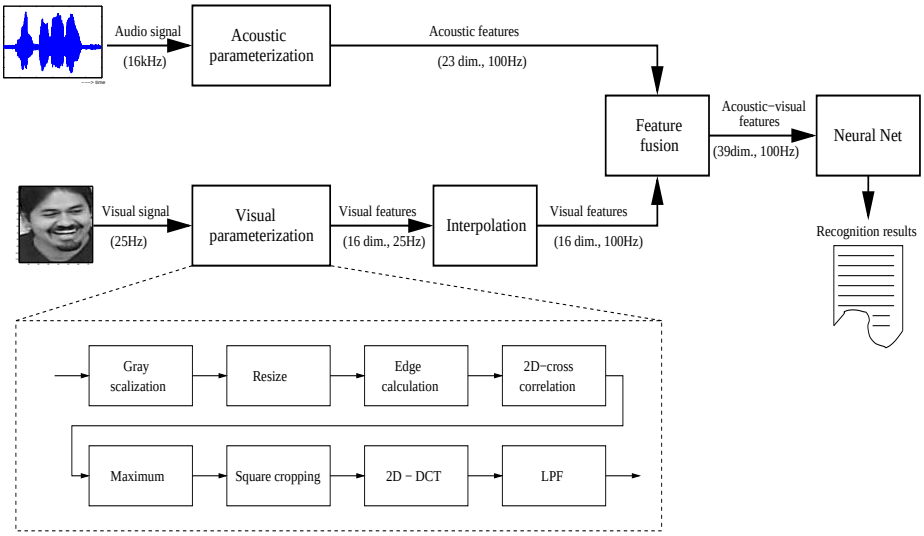

Fig. 6. Bimodal speech recognition system.

experiments:

1. One visual feature parameter of average brightness from the ROI is derived for each video frame.

2. 16 DCT coefficients from previously detected ROI are extracted (4 lowest DCTs in each dimension).

3. In the experiments with optical flow analysis, the ROI does not have to be detected. This analysis is applied directly on the sequence of the input video frames. Finally, three visual features are computed: horizontal and vertical variances of flow vector components and their covariance. These features are supposed to indicate the movement of speaker's mouth. They are especially useful for estimating silence periods.

4. The ROI has been also found by the lip locating algorithm based on edge detection and color filtering. From detected ROI, 16 DCT coefficients are derived.

5. We have performed several experiments with application of intra-frame LDA which is supposed to improve classification among speech classes and also provides the second-stage dimensionality reduction, whereas first-stage reduction is performed by PCA. Each video frame (32×64), containing lower half of tracked head position, is transformed into first 512 PCA basis. This data is used to estimate LDA statistics. Final transformation matrix is obtained by multiplying PCA and LDA matrices. Visual features used in recognizer are obtained by projection of input video frames onto first 45 PCA-LDA basis. Obviously, PCA and LDA statistics are estimated on training set. Fig. 3 shows original video frame, first PCA and PCA-LDA basis. For illustration, we also show a video frame reconstructed from first 10 PCA-LDA basis. Transformation obtained by LDA is not orthonormal, thus pseudo-inverse transformation is used for the reconstruction. It is not surprising that the first PCA basis return global head pose estimated over all training data. On contrary to the PCA basis, PCA-LDA basis attempts to emphasize the discriminability between speech classes, which are in video frames mainly caused by large variation around objects' mouths.

6. We have also experimented with application of PCA-LDA transform directly onto ROI (mouth poses detected by our correlation-based approach). In this case, mouth regions (16×34) are used for estimation of PCA (256 basis) and PCA-LDA statistics. Finally, projection onto first 45 PCA-LDA basis is performed. The first PCA and PCA-LDA basis are shown in Fig. 4. Compared to PCA, Fig. 5 shows the reconstruction of image from DCTs.

Fig. 6 shows audio-visual feature extraction system for the second set of experiments (extraction of DCT coefficients from ROI). The acoustic and visual features are combined into a single vector which is then used in training and recognition processes. In order to cope with the different acoustic and visual frame-rate, visual parameters are upsampled up to 100Hz by a simple linear interpolation. Finally, acoustic and interpolated visual features are merged to build *N*-dimensional audio-visual feature vectors.

The evaluation of different audio-visual features was done on phoneme set that consists of 46 phonemes. In addition, there were also two classes for silence and the

Feature extraction	CV [%]	FWD [%]	N
Audio only	28.9	31.0	23
Average brightness	29.0	31.52	23+1
DCT coefficients	28.33	31.33	23+16
Optical flow analysis	31.38	31.06	23+3
Seed algorithm	28.78	31.32	23+16
PCA-LDA (head)	26.74	31.80	23+45
PCA-LDA (mouth)	27.59	32.45	23+45

Table 1. Experimental results - **frame-based phoneme accuracies**: CV - cross-validation set, FW - forward passed testing data, N - vector size of extracted audio-visual features.

gap (a part of the speech recording belonging to a different speaker).

The recognition system is a simple NN, based on forward-backward algorithm, employing three layer perceptron with the softmax nonlinearity at the output. A Quicknet tool from the SPRACHcore package [15] was used. The size of input layer is determined by the length of feature vector. The hidden layer consists of 200 neurons with sigmoid non-linearities. The size of output layer is given by the number of phoneme classes. Outputs of the classifier are posterior probabilities of phoneme classes which we want to discriminate among.

6. EXPERIMENTAL RESULTS

To evaluate our various audio-visual feature extraction algorithms, we observe the following results:

- the best frame-based phoneme accuracy on CV sets,
- a frame-based phoneme accuracy on forward passed test data.

Experimental results are given in Tab. 1. During all experiments, the acoustic features were not touched. The vector size of visual features (as given in Tab. 1 together with acoustic features) did vary due to different kind of algorithms used to extract these parameters.

7. CONCLUSIONS

This paper proposes a bimodal speech recognition scheme using visual parameters extracted from meeting recordings. The main goal is to combine such features with classical acoustic parameters in order to increase robustness of ASR. However up-to-date, there was not any significant contribution which would bring sufficient solution. On the other hand, automatic analysis of meeting data is nowadays challenging task, and the effort to utilize a visual information that is free and most of the time available is natural. Experimental results related to the use of multimodal features are compared to the acoustic parameters (baseline). Although obtained results expressed by frame-based phoneme accuracies show small absolute improvement over the baseline, they are not negligible. Many problems appearing while processing video meeting data have not been properly solved yet. We need to improve the used head tracking algorithm as well as mouth detection method, which are still not very reliable, mainly due to real image conditions of visual data (low resolution of objects appearing in meetings, varying lighting conditions, etc.). We also expect that the temporal information from sequence of video frames, which has not been taken into account in our experiments yet, can play an important role.

8. REFERENCES

[1] I. McCowan, C. Marro and L. Mauuary. “Robust speech recognition using near-field superdirective beamforming with post-filtering.” *In Proceedings of ICASSP 2000*, volume 3, pages 1723-1726, Istanbul, Turkey, May 2000.

[2] P. Duchnowski, U. Meier, A. Waibel. “See me, hear me: Integrating Automatic Speech Recognition and Lip-reading.” *In Proceedings of ICSLP 1994*, pp. 547-550, Yokohama, Japan, September 1994.

[3] G. Potamianos, J. Luettin, C. Neti. “Hierarchical discriminant features for audio-visual LVCSR.” In *Proceedings of ICASSP 2001*, pp. 165-168, Salt Lake City, USA, May 2001.

[4] G. Potamianos, C. Neti. “Automatic Speechreading of Impaired speech.” *In Proceedings of Conf. Audio-Visual Speech Proc.*, pp.177-182, Aalborg, 2001.

[5] I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner, H. Boulard. “Modeling Human Interaction in Meetings.” *In Proceedings of ICASSP 2003*, Hong Kong, May 2003.

[6] D. Moore. “The IDIAP smart meeting room.” *IDIAP Communication 02-07*, 2002.

[7] I. McCowan, D. Moore, S. Sridharan. “Near-field Adaptive Beamformer for Robust Speech Recognition.” *In Digital Signal Processing 12*, pp. 87-106, 2002.

[8] I. Potucek. “Tracking movement objects in sequence pictures.” *In ElectronicsLetter.com journal*, <http://www.electronics.com>, January 2003.

[9] J. Kadlec. “Lips Detection in Low Resolution Images.” *To appear in EEICT 2004*, <http://www.feec.vutbr.cz/EEICT>, April 2004.

[10] J. C. Russ. “The Image Processing Handbook.” *CRC Press, Inc.*, 2nd Ed., USA, 1995.

[11] M. Heckmann, K. Kroschel, Ch. Savariaux, F. Berthommier. “DCT-based Video Features for Audio-Visual Speech Recognition.” *In Proceedings of ICSLP 2002*, pp. 1925-1928, Denver, USA, September 2002.

[12] B. K. P. Horn, B. G. Schunck. “Determining Optical Flow.” *In Artificial Intelligence*, vol. 17, nos.1-3, pp. 185-203 (1981-8).

[13] K. Iwano, S. Tamura, S. Furui. “Bimodal Speech Recognition using Lip Movement Analysis measured by Optical Flow Analysis.” *Int. Workshop on Hands-Free Speech Communication (HSC 2001)*, pp. 187-190, Kyoto, Japan, April 2001.

[14] Multimodal Media File Server, <http://mmm.idiap.ch>.

[15] “The SPRACHcore software package”, <http://www.icsi.berkeley.edu/∼dpwe/projects/sprach/>.

[16] G. Potamianos, Ch. Neti. “Audio-Visual Speech Recognition in Challenging Environments.” *In Proceedings of EUROSPEECH 2003*, pp. 1293-1296, Geneva, Switzerland, September 2003.