Multimodal Phoneme Recognition of Meeting Data

Petr Motlíček and Jan Černocký

VUT Brno, Faculty of Information Technology
{motlicek,cernocky}@fit.vutbr.cz

Abstract. This paper describes experiments in automatic recognition of context-independent phoneme strings from meeting data using audio-visual features. Visual features are known to improve accuracy and noise robustness of automatic speech recognizers. However, many problems appear when not "visually clean" data is provided, such as data without limited variation in the speaker's frontal pose, lighting conditions, background, etc. The goal of this work was to test whether visual information can be helpful for recognition of phonemes using neural nets. While the audio part is fixed and uses standard Mel filter-bank energies, different features describing the video were tested: average brightness, DCT coefficients extracted from region-of-interest (ROI), optical flow analysis and lip-position features. The recognition was evaluated on a sub-set of IDIAP meeting room data. We have seen small improvement when compared to purely audio-recognition, but further work needs to be done especially concerning the determination of reliability of video features.

1 Introduction

Information from the speaker's mouth region has been shown to improve the accuracy and noise robustness of Automatic Speech Recognition (ASR) systems [2]. However, up-to-date systems use audio-visual speech features recorded under ideal lighting conditions. Those video recordings contain a high-resolution video of subjects' frontal face, minimal changes in head positions and constant backgrounds. Insufficient research has been done about audio-visual ASR performance in real meeting conditions, where in addition to possibly noisy audio, the quality of the 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 our first experiments with an audio-visual ASR system with visually challenging data from real multi-party conversations recorded at IDIAP [4]. The method for extracting acoustic speech features used in our experiments is fixed and we focus mainly on the visual part processing algorithms. Although, the video recordings contain the whole meeting scenario,

in our experiments we already work with video streams generated by a head tracking algorithm. However, we also have to take into account such method processing sequences of head poses, that would be able to reliably indicate the position of the mouth in each video frame.

2 Experimental data

The data used in our experiments for training and testing purposes is recorded by the IDIAP smart meeting room [4], [5], equipped with synchronized multi-channel audio-visual recording facilities. The used recordings are generated by two cameras each capturing front-on view of two participants including the table region used for note-taking. All participants wear lapel microphones, and an eight-element circular equi-spaced microphone array is centrally located on the meeting table. Multimodal meting recordings used in our experiments are split into three parts: training, cross-validation (together 41 minutes) and testing (9 minutes).

3 Acoustic feature extraction

In applications involving multi-party conversations, it may be possible to acquire the speech using microphone arrays. Microphone arrays 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], [6], and separates speech signals of each speaker recorded by each lapel microphone.

The parameters of audio speech signal are provided by set of well-known Mel-filterbank log energies (23 banks) generated for each 20ms long speech frame, with audio feature sampling frequency F_{afs} = 100Hz. These parameters are extracted from beamformed audio signal recordings sampled at 16kHz.

4 Visual feature extraction

Prior to the extraction of visual features, we need to process the original video stream data in order to detect and track faces (heads) of humans. The method employed to track heads of human objects in meeting recordings is based on color detection. In order to predict and identify the trajectory of moving objects in meeting recordings, the algorithm employs Kalman filtering. Color is known to be a key feature for hands and head detection. Such an 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. The normalized RG-color space derived from RGB values provides a good solution to the problem of varying brightness. The visual input in our experiments is a video stream which is supposed to contain a sequence of head poses of one human object appearing in given meeting. The

video feature sampling frequency $F_{vfs} = 25$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 very initial experiments, the visual features are based on average brightness of a region-of-interest (ROI). The ROI is theoretically expected to be speaker's mouth, which should carry most of the variability caused by speech. Such a ROI is obtained using a correlation-based mouth detector. This detector searches the most similar pattern to the average mouth pattern in each video frame. The correlation is performed on the level of video frames processed by standard edge detection algorithm. In this algorithm we applied the Sobel method [7], 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 [7], [8], 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 considerable approach to derive visual features, we attempt to avoid the use of an algorithm searching for the 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. If we suppose that the variances in the flow of video frames are mainly due to movement of the speaker's mouth during the speech, optical flow analysis may do interesting job. In our experiments, we use the Horn-Schunck optical flow analysis [9]. Theoretically, optical flow is the distribution of apparent velocities of movement of brightness patterns in an image. It can provide an important information about the spatial arrangement of the objects viewed and the rate of change of this arrangement. 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 brightness of each

point is constant during a movement for a very short time. Thus, the equations are as follows:

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

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:

$$E_x u + E_y v + E_t = 0. \tag{2}$$

The vectors u and v denote apparent velocities of brightness constrained by this equation. Practically, optical flow velocities are calculated from a pair of connected images using several iterations.

4.4 Algorithm detecting lip positions (seed algorithm)

Finally, we were dealing with an algorithm detecting lip positions in the given image frame and which uses edge detection and color filtering for noise reduction and enhancement of the desired recognition of lips. The particular steps are:

- Detection of red pixels of the face (this operation processes an input image in order to correctly locate mouth's pixels), followed by its transformation to the binary form.
- Selection of the largest white area (the binary representation) 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.

5 Experimental setup

The audio-visual speech database collected at IDIAP [10] has been used. Acoustic features generated from beamformed speech recordings are Mel-filterbank log energies, as described in Sect. 3. These acoustic features are computed with F_{afs} = 100Hz. Derivation of visual features is the major part of our experiments. Each input sequence of video frames is at the beginning processed by a head tracking algorithm. Such visual data represents inputs in the following 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 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 indicate whether the speaker's mouth is moving or not, they are especially useful for estimating silence periods.

	Acoustic			Acoustic+Visual		
Feature extraction	CV [%]	FW [%]	N	CV [%]	FW [%]	N
Average brightness	28.9	31.0	23	29.0	**31.52**	24
DCT coefficients	28.9	31.0	23	28.33	**31.33**	39
Optical flow analysis	28.9	31.0	23	31.38	31.06	26
Seed algorithm	28.9	31.0	23	28.78	31.32	39

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

4. In last experiments, the ROI has been found by the lip detecting algorithm based on edge detection and color filtering. From detected ROI, 16 DCT coefficients are derived.

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 feature sampling frequencies, visual parameters are upsampled from 25Hz 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 a phoneme set that consists of 46 phonemes.

The recognition system is a simple Neural Network (NN) employing a three layer perceptron with the softmax nonlinearity at the output. A Quicknet tool from the SPRACHcore package [11] was used in all experiments. The size of the input layer is determined by the length of the feature vectors. In all experiments, the hidden layer consists of 60 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

For experimental purposes, the training data is split into training and cross-validation (CV) sets. These two sets are used to train NN. Then, testing data are forward passed through such NN. To evaluate our various audio-visual feature extraction algorithms, we observe the following results: (a) the best frame-based phoneme accuracy on CV sets, (b) a frame-based phoneme accuracy on forward passed testing data.

Experimental results are given in Tab. 1. In all experiments, the acoustic features were kept constant. The vector size of visual features (as given in Tab. 1 together with acoustic features) did vary due to different kind of methods used to extract these parameters.

7 Conclusion

This paper presents preliminary experiments with automatic recognition of phonemes in meeting recordings. Although, obtained results expressed by frame-based phoneme accuracies show small absolute improvement over the baseline, they are not negligible.

Algorithms extracting visual features are mainly influenced by the quality of the used head detection algorithm, which needs to work with low resolution video frames, under varying lighting conditions of meeting data, etc. In further work, we will concentrate on increasing the robustness of the visual feature extraction algorithms (e.g., the mouth detection needs to be replaced to track more reliably the mouth region) and on different modalities of combination with the acoustic part.

8 Acknowledgments

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

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] 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.

[5] D. Moore. "The IDIAP smart meeting room." *IDIAP Communication 02-07*, 2002.

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

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

[8] 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.

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

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

[11] "SPRACHcore", <http://www.icsi.berkeley.edu/~dpwe/projects/sprach/>.