Visual Speech Segmentation and Speaker Recognition for Transcription of TV News

Josef Chaloupka

SpeechLab, Department of Electronics and Signal Processing
Technical University of Liberec, Hálkova 6
461 17 Liberec, Czech Republic
josef.chaloupka@tul.cz

Abstract

This paper is about a method for visual segmentation of TV news. The TV news shows are segmented according to the visual stream from the video TV recordings in this method. Human faces are found in the single visual segments with the help of the fast algorithm for face detection. The found faces are compared with the visual GMMs, that have been trained from the video picture of the single broadcasters (anchors) from the TV news. The single visual segments, where the faces of the broadcasters have been found and recognized, have been compared with the acoustic segments from the acoustic segmentation. The speaker adapted HMMs have been used for speech recognition of these acoustic segments. The recognition rate is better for the use of this speaker-adapted HMMs compared to the use of the speaker independent HMMs. It is possible to use the methods for the speaker identification and verification from the acoustic signal in the acoustic segments. The results from the visual speaker identification will be better for smaller number of speakers and for the use of the video recordings of TV news with a lot of noise in the acoustic signal.
Index Terms: Visual speech segmentation, visual speaker recognition, TV news transcription

1. Introduction

A considerable improvement has been obtained in the area of the continuous speech processing and recognition recently. It is possible to recognize the continuous speech in the real time today. The recognition rate for the continuous speech recognition is over 90% for the English and it is over 80% for the inflectional languages that use large vocabularies (containing hundreds of thousands words). The technologies for the continuous speech recognition are applicable for the transcription of the broadcast programs, TV news, and so on. The acoustic signal is often segmented into smaller segments for the purposes of the transcription of broadcast programs in the first step. There is homogeneous information is in these segments – a single speaker is speaking, the music is playing, silence, noise, and so on. The segments are labeled as speech segments or as non-speech segments after the segmentation. The speech segments are recognized in the automatic speech recognition system. The Hidden Markov Models (HMMs) or the Artificial Neural Networks (ANNs) are used for the speech recognition most frequently today. The HMMs are used in this work. It is useful to find out who is speaking in the acoustic segment in the second step, i.e. to make speaker identification and verification. The speaker adapted HMMs can be used for the speech recognition from these acoustic segments if we have sufficient quantity of the acoustic recordings from the relevant human speaker. We can at least recognize if the speaker is a man or a woman (gender identification) if the speaker is not present in the speaker database, i. e. the speaker has not been verified. The recognition rate is better for the speech recognition where the speaker or gender adapted HMMs have been used. A large number of methods have been developed for the broadcast programs segmentation [1] and for the speech adaptation [2] till today. The methods and algorithm for the visual segmentation and for the visual speaker identification and verification from the visual stream of the TV news recordings are described in this work. The goal of this work has been the use of the visual signal as the supporting information for the transcription of TV news. Some methods for the visual broadcast segmentation have been described in previous papers [3]. But the single visual part is not useful for the broadcast programs segmentation. It is rather useful for the improvement of the information from the acoustic segmentation. The single visual segmentation is applicable providing that the single video parts from the broadcast video recording would be exactly cut and the video image would be rather static. These conditions are almost never satisfied. The next problem is that the visual stream is shifted compared to the acoustic stream by several hundreds of milliseconds in some television stations. The detected change of the visual signal would be ± 40 ms if the video sample rate would be 25 video images per seconds. These changes would be used for the improvement of the change boundaries from the acoustic segmentation. But it is not possible mostly because of the problems described above. Therefore the visual part of the video recordings has been used for the support of the transcription of TV news from the acoustic signal in this work. These video recordings of the TV news have been obtained from 3 Czech television stations. They are the news from the main news time from 7 till 8 o'clock p.m. The broadcasters comment short news from our country and from the world. These broadcasters are camera-scanned from the front and their utterances generate approximately 20-30% of the all video recordings. The video image is almost static if the broadcaster shows the news. It is good for audio-visual segmentation because the visual segments match with the acoustic segments. The human speakers (broadcasters) can be identified in these visual segments with the help of the image face detection and face identification.

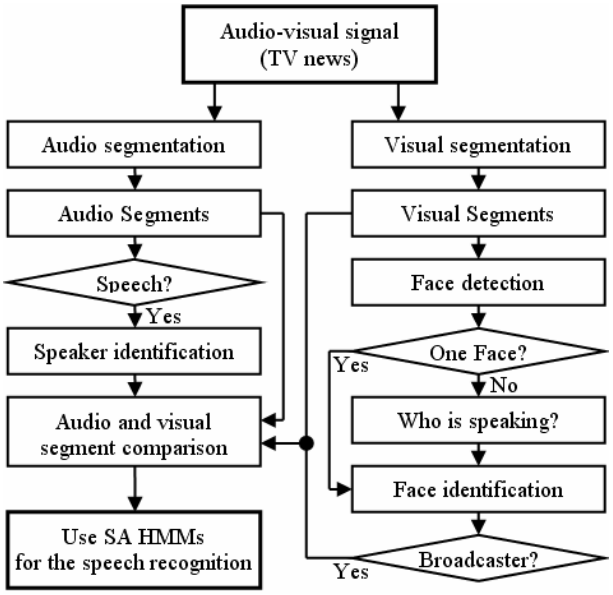

Figure 1*The principle of audio-visual segmentation.*

2. The Visual Segmentation of the Video News Recordings

A fast, simple and robust method has been found for the visual segmentation of video news recordings in our lab. The size of the processed visual data is several times bigger than the size of the acoustic data. For example: The visual data stream is 8 294 400 B per second for the video image with the sizes 384 x 288 pixels, where 1 pixel has 24 bits. RGB color value and the video sample frequency is 25 video images per second. The relevant acoustic data stream is 32 000 B per second for the sample frequency 16 000 Hz and when a single audio sample has a 16-bit value. Therefore the visual data stream is 259 x bigger than the acoustic data stream. The first method, that has been used, has been the following: The original RGB pixel values from the color video image have been transformed into the brightness. The sum from the all brightness values has been computed from the all gray-scale image. The difference of the brightness sums has been found between two subsequent images. The segment boundary has been marked in the visual signal when the value of the difference was bigger than some beforehand given threshold. The visual segments are between two following segment boundaries. Some errors have appeared in this simple and very fast method because the sum of the brightness is a global value that doesn't contain the informative content from the video image. Two very different video images can have the same value of the brightness sum. The video image has been divided into several regular areas. The sum of the brightness has been computed in each of these areas. The distance *VD* has been computed from the single corresponding areas between two subsequent video images:

$$VD_j = \sum_{i=1}^{N} \sqrt{\left(V_{ij} - V_{i(j-1)}\right)^2} \qquad (1)$$

where N is the number of the areas, j is the number of the image and V_{ij} is the sum of the brightness in the single area.

The segment boundary has been marked in the visual signal when the value of the distance has been bigger than some threshold as it was in the first method. The good results of the visual segmentation have been obtained if 16 areas have been created in the video image.

This method has the same or very similar problems if the image has smaller number of areas. A bigger number of areas (than 16) do not have significantly better results. This method has worse results if the number of the areas gets near to the number of pixels in the video image, because the smaller change of image produces a bigger value of distance. This method has been relatively as fast as the first method but sometimes the informative content has been changed but the segment boundary has not been detected.

Therefore the third method has been developed. This method is slower then previous methods but the results from the visual segmentation are more robust. The energy DCT visual features have been used in this method. These features are used for the visual speech recognition very often [4]. The original image has been transformed with the help of the 2D Discrete Cosine Transform DCT. The use of the DCT is very useful because we can use the algorithm of the Fast Cosine Transform FCT and the computation of this DCT is relatively fast. The energy has been computed from the single DCT coefficients. 20 highest energy DCT features have been then used for the purposes of the visual segmentation. The distance of the single energy DCT features has been computed between two subsequent video images and the segment boundary has been selected the same way as in the previous method. The visual segmentation of one hour of video news recording takes 1.5 hours on a 2.4 GHz PC.

The estimation of the success of this method for visual segmentation is very labor-intensive because the informative content changes very fast in some dynamic video scene and it is very difficult to manually segment this visual section for the comparison. Therefore only the segments with the speakers have been manually marked in the reference video recordings. These visual segments have been important for the comparison with the acoustic segment in the next step.

5 hours of video news recordings have been manually segmented image by image for the evaluation of the automatic visual segmentation. Only the segment boundaries where the important speaker (the anchor) was present have been manually selected and these boundaries have been compared to the automatically obtained boundaries.

The error has been detected if the automatically found segment boundary has not been the same as the manually selected boundary. Another kind of error has been detected if the automatically selected boundary has been between the manually selected boundaries. The resulting evaluation of the visual segmentation accuracy *VACC* was:

$$VACC = \frac{NV - DV - IV}{NV} \qquad (2)$$

where NV is the number of all manually selected segment boundaries, DV is the number of the automatically detected boundaries that have not been found compared to the manually found boundaries and IV is the number of the automatically detected boundaries that have been found between two manually selected segment boundaries.

The resulting accuracy for visual segmentation for single methods has been followed: 53% for the first method, 74% for the second method, and 98% for the third method where the energy DCT features have been used. The human face has been found out in the beginning of the automatically detected visual segment. This face was then used for the visual speaker identification and verification.

3. The Face Detection in the Visual Stream

The face detection in the video image is a difficult task. The human speakers can be camera-scanned in various positions and they can have various skin colors. Some objects in the video image may also have very similar shape, color, and texture as the human face. A large number of methods have been developed for the face detection in the video picture up to now [5]. We have been focused in our video news recordings on the video pictures that contain one ore two faces (broadcasters of the TV news) and where the speakers are camera-scanned from the front. We have developed the fast and robust method for the face detection that is based on the color and shape segmentation of the video image. The original color RGB image has been converted to the H color part from the HSI color space [6]. This image has been color-segmented according to the beforehand given threshold T = 0.8901. The conversion to the HSI color space and color segmentation are made in one step with the help of the 3D look-up table that converts RGB color value directly to the values 0 and 1. The result is the binary image where the pixels from the image background have the value 0 and the pixels that create objects have value 1. The binary image is reduced 6 times to the sizes 64 x 48 pixels. This is possible because the speaker face from the important visual segments has relatively big surface in the video picture and the location exactness of the face object is still adequate in this smaller picture. The following shape segmentation is 200 x faster for this smaller image and that is why the image is reduced. The small binary noise is removed and the objects are simplified by means of the morphologic operations opening and closing [6] with the small rounded structure element with the sizes 5 x 5 pixels in the first step. The method of region identification [6] is applied to the binary image in the second step. The pixels from the single objects are transferred to the same value with the help of the region identification algorithm. The number of pixels in the single objects is computed and the proportion between the sizes in the single objects is determined too.

All objects that have a smaller number of pixels than some threshold are removed from the image. Also elongated objects are removed. These objects are most probably not faces. The result is a smaller number of objects that are probably human faces. These objects have been used for the separation of the face regions from the original color image. The objects of eyes in the expected positions would be good to find in the found objects for a better accuracy of the resolution if the object is a human face or not. The resulting method would be substantially slower. The case, that the found face region was not a human face, was not noted in our video news recordings. The face region would be removed with the biggest probability in the next step where the found face region is used for the visual speaker identification and verification. The face region(s) are found within 10 ms with the help of this method for the face detection.

4. The Speaker Identification and Verification from the Found Face Regions

The visual speaker identification is solved in the large number of papers and works [7]. The PCA or ICA transforms are used most frequently for these purposes. The computation of these transforms is very time consuming. The method of the GMMs (one mixture HMM) has been used in this work. 10 images with the faces have been collected for each human speaker that would be in our database. The energy DCT features have been computed from these images the same way as for the visual segmentation. 60 highest energy DCT features from each 10 images have been used for the training of the relevant GMM for single speakers. The objects from the face detection method have been recognized with the help of these GMMs. The best assignment between the unknown face region and GMMs is then found. The result is the maximum probability rate of the similarity. This rate is used for the rejection (if the rate is smaller than the selected value) of the unknown images that are not perhaps in the database. The visual speaker pseudo-verification has been used for the improvement of the visual speaker identification. The following algorithm has been developed for the visual speaker pseudo-verification: 3 video images from the 3 seconds of the important visual segments have been selected. The speaker is verified in the visual segment if the results from the 3 video images belong to the same speaker. The speaker is selected as unknown if this condition is not satisfied.

5. Who is speaking?

Two speakers (broadcasters) have been found in several visual segments. It has been necessary to find out which one of them is speaking. The objects of the lips have been found in the single face regions for these purposes. The lips object has been separated from the face region with the help of the simple color and shape segmentation [8]. The object of the lips is smaller than the object of the face. Therefore the special algorithm [8] has been used for the finding of the exact threshold for the color segmentation. The vertical expansion of the lips from single face regions is found out for each video image from the visual segment. The face is identified if the parameter of the vertical expansion of the lips is changed dynamically. The visual segment is labeled by the name of the speaker if the speaker is identified and verified.

6. Experimental Tests

Our system for the automatic transcription of broadcast programs [9] has been used for the testing of the visual signal efficiency. This system contains the module for the acoustic segmentation (the speech detector and the speaker change detector) of the video news recordings, the module for adaptation and the module for the continuous speech recognition. The module for the continuous speech recognition is focused on the processing and recognition of the Czech language. The Czech language belongs to the inflectional languages. A big vocabulary is necessary for the coverage of the independent text for the inflectional languages. The Czech language doesn't have a fixed order of the words in the sentences as it is for example in English. We use the vocabulary

with 312 000 Czech words at present (2005/06). This vocabulary covers 96.5% independent text [10] (English language needs only 65 000 words for the 99% coverage of the independent text). The recognition rate (accuracy) with the use of this 312 000-word vocabulary is 79.8% for continuous speech recognition where the speaker independent (SI) HMMs have been used and 83.6% where the speaker adapted (SA) HMMs have been used on the data [11] that have been created in the project COST298 [12]. These results – the recognition rates are not as good as for the English. Therefore the alternative methods are tested for the improvement of the recognition rate. The use of the visual part is one of these methods. We are focused on the improvement of our system for the transcription of TV news in this work. For the TV news the video news recordings from 3 Czech TV stations have been selected. The acoustic segmentation according to the results of the visual segmentation is not a good idea but the results from the visual segmentation can be used as the supporting information. Therefore the faces of the TV news broadcasters have been found in the visual segments from the visual segmentation. The visual segments where the speaker has been detected and verified have been compared with the relevant acoustic segment from the acoustic segmentation. If at least 90% of the visual segment was found in the acoustic segment, then this acoustic segment was declared as narrated by the speaker who was identified in the visual part. This acoustic segment has been recognized with the help of speaker-adapted HMMs. This experiment has been focused on the comparison of the improvement for the acoustic segment where the SA HMMs will be used to the use of the SI HMMs. The information who is speaking in the acoustic segment has been obtained from the visual signal. 173 acoustic segments (6873 words) have been selected and separated from this audio-visual segmentation. 16 different TV broadcasters (8 women and 8 men) in these segments have been present in our database. The resulting recognition rate (accuracy) has been 87.04% for the use of the SA HMMs and 85.81% for the SI HMMs. So, the relative improvement in the recognition rate was 8.7% when the additional visual information has been used. These results are very good because the video recordings of the TV broadcasters are obtained in the TV studio and therefore only a low level of noise is in the recordings. The second reason is that TV broadcasters have very good pronunciation.

7. Conclusion and Future Work

The possibilities of the visual signal use for the transcription of TV news are described and discussed in this work. The relatively fast, simple and robust methods for the visual segmentation and for the speaker identification and verification from the visual segments have been developed and created. The experiments have been focused on the acoustic segment speech recognition where the TV broadcasters from the TV news are present. This test is relatively simple because the video recordings are obtained in the TV studio. The human faces are very well spotlighted and the speakers are camera-scanned from the front. The visual part is useful for the broadcast programs where a small number of known people is found very often or periodically, for example in various talk shows, in the news shows, in the parliamentary discussions, and so on. The use of the visual part attains good results for the speaker identification where the noise is present in the acoustic signal and the speaker identification and verification from the acoustic signal is not possible. We would like to test this algorithm for the visual segmentation for a wider range of broadcast programs in the near future. The database of GMMs of speaker faces (for the speakers that are found in these broadcast programs) will be extended in the first step.

8. Acknowledgements

The research reported in this paper was partly supported by the Czech Science Foundation (GACR) through the project No. 102/05/0278.

9. References

[1] Zdansky, J., Nouza, J., “Detection of Acoustic Change-Points in Audio Records via Global BIC Maximization and Dynamic Programming”, In: Interspeech 2005, September, 2005, Lisboa, Portugal, pp. 669-672, ISSN 1018-4074

[2] Ariki Y, Shigemori T., Kaneko T., Ogata J., Fujimoto M., "Live Speech Recognition in Sports Games by Adaptation of Acoustic Model and Language Model”, in Proc. of Eurospeech2003, Geneva, 2003

[3] Tsekeridou, S., Krinidis, S., Pitas, I. "Scene Change Detection Based on Audio-Visual Analysis and Interaction", In Multi-Image Analysis: 10th International Workshop on Theoretical Foundations of Computer Vision, Dagstuhl Castle, Germany, March 12-17, 2000

[4] Scalon, P., Reilly, R., B., De Chazal, P.: "Visual Feature Analysis for Automatic Speechreading", In Audio Visual Speech Processing Conf., France, 2003

[5] Yang, M., H., Ahuja, N. "Face Detection and Gesture Recognition for Human-Computer Interaction", In Kluwer Academic Publishers, USA, 2001, ISBN 0-7923-7409-6

[6] Sonka, M., Hlavac, V., Boyle, R.: Image Processing, Analysis, and Machina Vision, In PWS Publishing, 1998, ISBN 0-534-95393-X

[7] Maison, B., Neti, Ch., Senior, A., "Audio-visual speaker recognition for video broadcast news: some fusion techniques", IEEE Multimedia Signal Processing (MMSP99), Denmark, Sept, 1999

[8] Chaloupka, J., "Automatic Lips Reading for Audio-Visual Speech Processing and Recognition", In: Proc. of ICSLP 2004, Korea, 2004, pp. 2505-2508, ISSN 1225-441x

[9] Nouza, J., Zdansky, J., David, P., Cerva, P.,Kolorenc, J., Nejedlova, D.: Fully Automated System for Czech Spoken Broadcast Transcription with Very Large (300K+) Lexicon. In: Interspeech 2005, September, 2005, Lisboa, Portugal, pp. 1681-1684, ISSN 1018-4074

[10] Nejedlova, D.,"Creation of Lexicons and Language Models for Automatic Broadcast News Transcription", Dissertation Thesis, Technical University Liberec, Czech Rep., 2006

[11] Cerva, P., David, P., Nouza, J.: Acoustic Modeling Based on Speaker Recognition and Adaptation for Improved Transcription of Broadcast Programs. In: Specom 2005, October, 2005, Patras, Greece, pp. 183-186, ISBN 5-7452-0110-x

[12] Vandecatseye A. et al, “The COST278 pan-European Broadcast News Database”, Proc. of LREC 2004, Lisbon, Portugal, May 2004