# Online Speaker Adaptation of an Acoustic Model using Face Recognition

Pavel Campr[1], Aleš Pražák[2], Josef V. Psutka[2], and Josef Psutka[2]

[1] Center for Machine Perception, Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, Technická 2, 166 27 Prague 6, Czech Republic
cmp.felk.cvut.cz
camprpav@cmp.felk.cvut.cz
[2] Department of Cybernetics, Faculty of Applied Sciences, University of West Bohemia in Pilsen, Univerzitní 8, 306 14 Pilsen, Czech Republic
www.kky.zcu.cz
{aprazak,psutka_j,psutka}@kky.zcu.cz

**Abstract.** We have proposed and evaluated a novel approach for online speaker adaptation of an acoustic model based on face recognition. Instead of traditionally used audio-based speaker identification we investigated the video modality for the task of speaker detection. A simulated on-line transcription created by a Large-Vocabulary Continuous Speech Recognition (LVCSR) system for online subtitling is evaluated utilizing speaker independent acoustic models, gender dependent models and models of particular speakers. In the experiment, the speaker dependent acoustic models were trained offline, and are switched online based on the decision of a face recognizer, which reduced Word Error Rate (WER) by 12% relatively compared to speaker independent baseline system.

**Keywords:** acoustic model, speaker adaptation, face recognition, multimodal processing, automatic speech recognition

## 1 Introduction

Automatic speech recognition systems are used in many real world applications. An unpleasant problem is the frequent and sometimes very rapid change of speakers. This disallows to use an online speaker adaptation technique, which requires relatively long part of speech for adaptation. Special focus is given to real-time systems, such as automatic subtitling systems, where this problem is more visible. One solution is to enhance online speaker adaptation techniques, but here we study a different multimodal approach that uses video modality for rapid speaker change detection.

The proposed system uses a face recognizer to identify the speaker's identity and gender, in times lower than 100 ms. Based on the results, the pre-trained acoustic models in the LVCSR system can be switched, which leads to decrease in WER compared to the baseline system using only a speaker independent model. Advantages and disadvantages of such an approach, compared to traditional audio-only approach, are discussed. The experiment is carried out on TV broadcast of Czech parliamentary meetings.

The paper is organized as follows. Section 2 describes LVCSR system used in experiments. Section 3 describes face recognition system used for gender and identity

estimation. Section 4 presents proposed system as a whole, with focus on an acoustic model selection from video stream. Section 5 presents the experiment and results. Final sections shortly discuss open problems for future work and conclude the paper.

## 2 Real-time Automatic Speech Recognition System

One of the key applications for real-time LVCSR systems is the automatic online subtitling of live TV broadcasts. To increase the recognition accuracy, some speaker adaptation techniques with suitable fast online speaker change detection can be used [1]. Traditionally, the speaker change detections are based only on audio track analysis [2].

A common metric of the performance of a LVCSR system is Word Error Rate (WER), which is valuable for comparing different system and for evaluating improvements within one system. It was used in this paper for the evaluation.

In the following, the LVCSR system used for the evaluation of proposed system is described. The baseline LVCSR system uses speaker independent (SI) acoustic models. Gender dependent acoustic models are used if the gender of the speaker is known. Finally, speaker adaptation of an acoustic model is used to fine-tune the models for particular speakers. All models are trained offline before they are used in the online experiments. In a real-world application, the models could be trained and added to the bank of acoustic models in the course of time.

### 2.1 Language Model Details

The language model was trained on about 52M words of normalized Czech Parliament meeting transcriptions from different electoral periods. To allow captioning of an arbitrary (including future) electoral period, five classes for representative names in all grammatical cases were created and filled by current representatives. See [3] for details. To reduce out-of-vocabulary (OOV) word rate, additional texts (transcriptions of TV and radio broadcasting, newspaper articles etc.) were added. The final trigram language model with vocabulary size of 588923 words was trained using Kneser-Ney smoothing.

### 2.2 Acoustic Model Details

The acoustic model was trained on 585 hours of parliamentary speech recordings using automatic transcriptions. Since we trained the previous acoustic model on a much smaller amount of speech records (90 hours) using manual transcriptions, we tried to upgrade that model using parliamentary recordings collected during the real captioning [4]. These records were automatically recognized and reliable parts of automatic transcriptions were used for acoustic model training. Only words, which had confidence greater than 99% and their neighboring words had confidence greater than 99% too, were selected. Hence, so the real amount of training data was 440 hours. For details on the confidence measure see [5]. Such an enhanced acoustic model reduced WER by 20% relatively.

We used three-state HMMs and Gaussian mixture models with 44 of multivariate Gaussians for each state. The total number of 236940 Gaussians is used for the SI model. In addition, discriminative training techniques were used [6]. The analogue input speech signal is digitized at 44.1 kHz sampling rate and 16-bit resolution format. We use PLP feature extraction with 19 filters and 12 PLP cepstral coefficients, both delta and delta-delta sub-features were added. Feature vectors are computed at the rate of 100 frames per second (fps).

Gender dependent acoustic models were trained using an automatic speaker clustering method [5]. The initial split of the data was based on male/female markers obtained from the manual transcriptions utilizing previous acoustic model. The resulting acoustic models contained 29 Gaussians per state for men and 14 Gaussians for females. Discriminative training techniques were used for both acoustic models as well.

Acoustic model adaptation of specific speakers was carried out using unsupervised fMLLR adaptation [7]. Reliable parts of automatic transcriptions were chosen in the same way as for acoustic model training. Only one transformation matrix and shift vector was trained on available speech data (from 40 seconds to 150 minutes) for each speaker using gender dependent acoustic model.

## 3 Face Recognition

As shown in Figure 2, the goal of face recognition module is to detect and track faces in the image sequence, and to estimate their gender and identities, all in real time.

The task of real-time face detection and identification was widely studied and existing solutions are capable to solve this task with high accuracies [8] [9].

In this paper, we use a detector of facial landmarks based on Deformable Part Models, proposed by Uřičář [8]. In addition to the required position of the face in the image, this real-time face detector provides a set of facial landmarks like nose, mouth and canthi corners. Such landmark positions are used for normalized face image construction, which is used in the next recognition step.

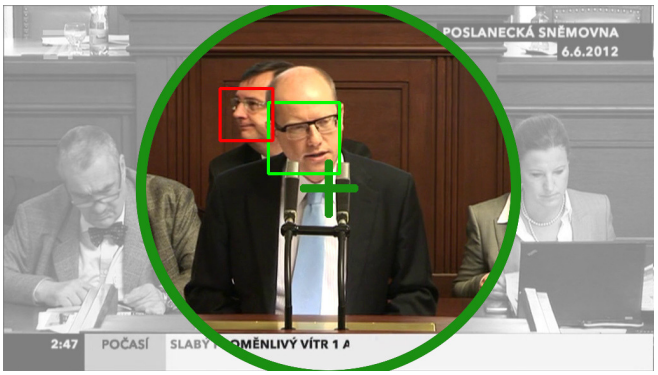


**Fig. 1.** Example of a processed video frame. The circle denotes a face detection zone. The speaker's face is marked by the green rectangle, non-speaker face by the red rectangle.

The problem of gender and identity estimation are classification problems, for which we use multi-class Support Vector Machine (SVM) classifier, more precisely the implementation done by Franc [10]. The classifier uses image features, computed from normalized face images, based on Local Binary Patterns (LBP). For the gender classification task we used SVM learned on 5 million examples that was presented by Sonnenburg [9]. The results of classificators, which process single images, are accumulated for whole face tracks in the video. The decisions are accumulated and are provided as a global decision when the confidence is high enough.

Example depicted in Figure 1 presents results of the face detector, that is limited to process only highlighted circular area, where the speaker occurrence is expected. The results of gender and identity estimation are presented later in Experiment & Results section.

## 4 Proposed Multimodal LVCSR System

Figure 2 presents the schema of the proposed real-time automatic speech recognition system that uses face recognition to detect speaker changes. The main question is how to associate the face(s) coming from video modality with the speaker(s) coming from audio modality. This opened topic, sometimes referred as ”audiovisual speaker diarization”, is discussed for example in [11] or [12]. To our knowledge, no other work presented the use of speaker change detection from video modality for the use in real-time LVCSR system.

Our system is based on a speech recognizer as described in Section 2, on face recognizer as described in Section 3 and on acoustic model selector, that identifies the speaker based on the results from face recognizer, i.e. it answers the question ”who is the speaker” based only on the video stream.

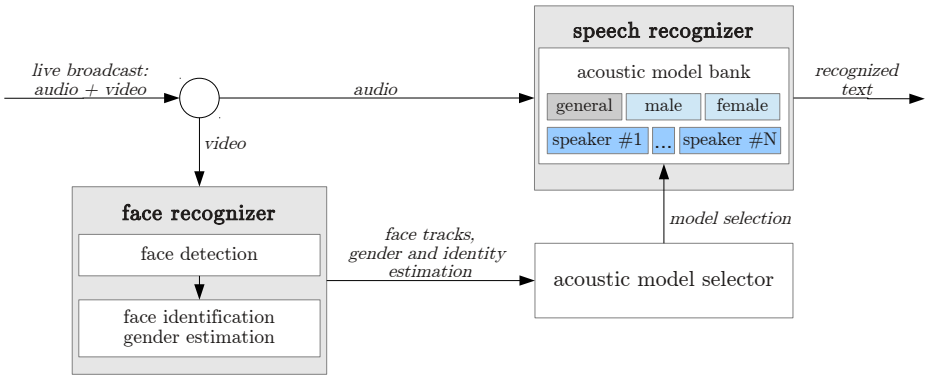


**Fig. 2.** Schema of an online automatic speech recognition system using multiple acoustic models, which are activated based on identity and gender estimations obtained from the face recognition module.

### 4.1 Acoustic Model Selection

This module (see Figure 2) should identify the speaker, in real time, from the results of the face recognition module. Generally, the video stream can contain several or none faces, similarly to audio stream where the speakers can overlap or be quiet. Additionally, it is not expected that a speaker is always visible, and, vice versa, the face always speaks. A discussion about all the cases and possible solutions are presented in [11], but only for offline processing.

Here, for the real time processing, we impose some rules for the input broadcast that facilitate the speaker identification from the video. The rules are based on the type of input data, in our case for the parliamentary broadcasts. Sample image is shown in Figure 1, where the speaker is visible in the middle of the frame. Sometimes, the camera is switched and the frame contains a view of the whole room. Such knowledge about the broadcast allows to build a simple and real-time mapping between the faces and the speaker:

1. ignore small faces (we ignore faces smaller than 75 pixels)
2. the face whose position is closest to the center of the screen is denoted as current speaker
3. if the identity of the speaker's face is recognized, the acoustic model of this particular speaker is activated; otherwise go to step 4
4. if the identity is not recognized but the gender is, the gender dependent acoustic model is activated; otherwise go to step 5
5. if we have no knowledge about the speaker's face at all, the general acoustic model is activated

Such rule-based system is not general, but is sufficient for the first experiment and evaluation in this area, and it can be enhanced and generalized later, taking inspiration from offline variant of this task [11].

Additional decision must be done when all the faces are lost. We examined two strategies. The first is to immediately activate the SI model. The second is to keep the current acoustic model until the next face is found. In our experiment, the second strategy performed slightly better and is presented in the following section. This can be caused by occasional camera switches, that are present in the broadcast and show some graphics or the whole parliamentary room even if the speaker is still speaking.

## 5 Experiments & Results

Several experiments based on gender or speaker change detection were performed on one test record (200 minutes, 22286 words, 35 speakers, 105 speaker changes, 0.2% OOV words, perplexity of 315) simulating on-line transcriptions. The camera captured the speaker in 70.6% of the time, in the rest of the time the camera captured the whole parliamentary room or some graphics.

The face recognition module was able to identify the gender immediately in one frame, the median number of frames required to identity the face was 3, corresponding to 75 ms (with fps 25). The module identified 74 faces fully and in 34 cases only

the gender was identified. The fails of identity recognition were mostly caused by the side orientation of the face or overlapping of the face with the microphone. The face classifier was trained on a closed set of speakers that were present in the test record.

The evaluation of the entire system follows. Firstly, only gender dependent acoustic models were switched during recognition based on speaker gender changes that were manually annotated. Relative WER reduction of 11% was achieved over baseline system with the SI acoustic model. Furthermore, by switching off-line prepared speaker adapted acoustic models, we reduced WER by 15% relatively over the baseline system (see Table 1). This represents a maximal improvement achievable by the proposed approach.

Next, the same experiments were performed based on gender and speaker changes detected automatically as described above. The WER reduction was slightly worse than in the case of manual annotation. To summarize, the baseline WER, our system WER and maximal achievable WER are 9.68%, 8.96% and 8.62%. for gender recognition only. For identity recognition, the results are 9.68%, 8.62% and 8.22%, with 11% relative WER reduction (which was 15% for the manual annotation).

In Table 1, the gender dependent results and results for two individual speakers are presented as well. The speaker "Schwarzenberg" is known for his speech disorder that leads to the worst WER of all speakers in the test record. The speaker "Němcová" acts as a chair of the meeting, so she often speaks during applauses or other types of speech noise, overlaps with other speakers, or is not visible in the camera when she speaks, so that an acoustic model of another speaker is used.

To conclude, the results show that decisions of a face recognizer can be used very efficiently for speaker change detection, which leads to improvement of the LVCSR system. Such approach has some advantages and disadvantages compared to audio-based detection, which is widely used nowadays. The advantage is the accurate and fast response of the face recognizer (usually less than 100 ms after the face is found). The disadvantage is the need to identify the speaker from the video, which is only possible when the speaker's face is visible.

**Table 1.** Results of LVCSR system: Word Error Rates (WER). Columns denoted as "gender" represent results for the case where only information about speaker's gender was used for acoustic model adaptation, columns "identity" relate to the case where both gender and identity were used. At the bottom, results for two worst performing speakers are presented: speaker "Schwarzenberg" leads to the worst absolute WER, speaker "Němcová" leads to increased WER with enabled adaptation.

| | | *without* | *with adaptation* | | | |
|---|---|---|---|---|---|---|
| | | *adaptation,* | *manual annotation* | | *face recognition* | |
| *speaker* | *words* | *baseline* | *gender* | *identity* | *gender* | *identity* |
| all | 22286 | 9.68% | 8.62% | 8.22% | 8.96% | 8.62% |
| men ♂ | 19635 | 9.91% | 9.00% | 8.51% | 9.07% | 8.71% |
| women ♀ | 2651 | 7.95% | 5.82% | 6.03% | 7.93% | 7.81% |
| Schwarzenberg | 212 | 42.16% | 35.20% | 32.10% | 35.20% | 32.88% |
| Němcová | 865 | 6.64% | 5.31% | 7.11% | 9.34% | 9.34% |

## 6 Future Work

Many enhancements can be applied to the whole proposed system or to particular modules. In the speech recognition module, the speaker dependent acoustic models can be trained online as proposed in [1]. A speaker verification could be employed to verify the decisions of the face recognizer. In the face recognition module, the face models could be improved during the time from the new data. Lip activity detector or correlation between lip movements and the audio signal could be employed for a more robust identification of the speaker among found faces.

The speaker identification could be based on both audio and video streams, where the results of traditional audio-based speaker detector and video-based speaker detector could be combined. It is expected that the multimodal processing would provide better results compared to a scenario with only one modality.

## 7 Conclusions

We have proposed a real-time speaker change detection system based on face recognition. It can be incorporated into an automatic speech recognition system to switch acoustic models, which leads to the reduction of Word Error Rate (WER). Several experiments based on gender or speaker change detection were performed on a test recording simulating on-line transcriptions. Relative WER reduction of 7% was achieved over baseline system with speaker independent acoustic model using only a gender switch detection. Furthermore, by switching off-line prepared adapted acoustic models of speakers, we reduced WER by 11% relatively.

## 8 Acknowledgments

This research[1] was supported by the Grant Agency of the Czech Republic, project No. GAČR P103/12/G084.

## References

1. Pražák, A., Zajíc, Z., Machlica, L., Psutka, J.V.: Fast Speaker Adaptation in Automatic Online Subtitling. International Conference on Signal Processing and Multimedia Applications (1) (2009) 126–130
2. Ajmera, J., McCowan, I., Bourlard, H.: Robust Speaker Change Detection. IEEE Signal Processing Letters **11**(8) (2004) 649–651
3. Pražák, A., Psutka, J.V., Hoidekr, J., Kanis, J., Müller, L., Psutka, J.: Automatic Online Subtitling of the Czech Parliament Meetings. Text, Speech and Dialogue, LNCS (2006) 501–508

[1] The access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum, provided under the programme "Projects of Large Infrastructure for Research, Development, and Innovations" (LM2010005) is highly appreciated.

4. Trmal, J., Pražák, A., Loose, Z., Psutka, J.: Online TV Captioning of Czech Parliamentary Sessions. Text, Speech and Dialogue, LNCS **6231** (2010) 416–422
5. Psutka, J.V., Vaněk, J., Psutka, J.: Speaker-clustered Acoustic Models Evaluated on GPU for on-line Subtitling of Parliament Meetings. TSD 2011 (2011) 1–7
6. Povey, D.: Discriminative Training for Large Vocabulary Speech Recognition. PhD thesis, Cambridge University, Engineering Department (2003)
7. Zajíc, Z., Machlica, L., Müller, L.: Robust Statistic Estimates for Adaptation in the Task of Speech Recognition. Text, Speech and Dialogue (2010) 464–471
8. Uřičář, M., Franc, V., Hlaváč, V.: Detector of Facial Landmarks Learned by the Structured Output SVM. In: VISAPP '12: Proceedings of the 7th International Conference on Computer Vision Theory and Applications1, Rome, Italy (2012) 547–556
9. Sonnenburg, S., Franc, V.: COFFIN: A Computational Framework for Linear SVMs. Technical Report 1, Center for Machine Perception, Czech Technical University, Prague, Czech Republic (2009)
10. Franc, V., Sonneburg, S.: Optimized Cutting Plane Algorithm for Large-Scale Risk Minimization. Journal of Machine Learning Research **10** (2009) 2157–2192
11. El Khoury, E., Sénac, C., Joly, P.: Audiovisual diarization of people in video content. Multimedia Tools and Applications (2012) 1–29
12. Bendris, M., Charlet, D., Chollet, G.: People indexing in TV-content using lip-activity and unsupervised audio-visual identity verification. In: 2011 9th International Workshop on Content-Based Multimedia Indexing (CBMI), IEEE (2011) 139–144