IMPROVING HANDS-FREE SPEECH RECOGNITION IN A CAR THROUGH AUDIO-VISUAL VOICE ACTIVITY DETECTION

Friedrich Faubel[1], *Munir Georges*[1], *Kenichi Kumatani*[1,2], *Andrés Bruhn*[1], *Dietrich Klakow*[1]

[1] Saarland University, Saarbrücken, Germany
[2] Disney Research, Pittsburgh, USA

ABSTRACT

In this work, we show how the speech recognition performance in a noisy car environment can be improved by combining audio-visual voice activity detection (VAD) with microphone array processing techniques. That is accomplished by enhancing the multi-channel audio signal in the speaker localization step, through per channel power spectral subtraction whose noise estimates are obtained from the non-speech segments identified by VAD. This noise reduction step improves the accuracy of the estimated speaker positions and thereby the quality of the beamformed signal of the consecutive array processing step. Audio-visual voice activity detection has the advantage of being more robust in acoustically demanding environments. This claim is substantiated through speech recognition experiments on the AVICAR corpus, where the proposed localization framework gave a WER of 7.1% in combination with delay-and-sum beamforming. This compares to a WER of 8.9% for speaker localizing with audio-only VAD and 11.6% without VAD and 15.6 for a single distant channel.

Index Terms— microphone arrays, audio-visual systems, acoustic signal detection, time of arrival estimation, automatic speech recognition

1. INTRODUCTION

Distant speech recognition (DSR) systems are of great interest in automotive environments as hands-free operation is the best way to avoid distraction of the driver [1, §1.1.3]. A DSR system typically consist of three components: a speaker localization module, a spatial filter (beamformer) and a postfilter. But all these components are directly or indirectly dependent on voice activity detection (VAD). This starts with the fact that the location of the speaker should only be determined when he or she is actually speaking. Unfortunately, the performance of audio-only VAD systems tends to drop in acoustically challenging environments. Hence, in this work, we consider using visual information in order to design a noise-robust VAD system that, in contrast to traditional audio-visual speech recognition (AV-ASR) approaches, does not require accurately clipped high resolution images of the mouth taken under artificial lighting conditions. It might be noteworthy that similar AV-VAD approaches have been proposed by Almajai and Milner [2] as well as Yoshida et al. [3]. Our work differs from these in that (1) we propose a novel Gaussian mixture filter based face tracking system that accurately identifies the mouth region; (2) we use a slightly different visual feature extraction chain that includes several normalization steps; and (3) we show how the performance of a DSR system can be improved by incorporating VAD into the speaker localization framework. The block diagram shown in Figure 1 gives an overview of the proposed system. It can essentially be described as a two stage system, in the first of which audio-visual voice activity detection is performed; and in the second of which speaker localization is improved through spectral subtraction, as described in more detail in the following.

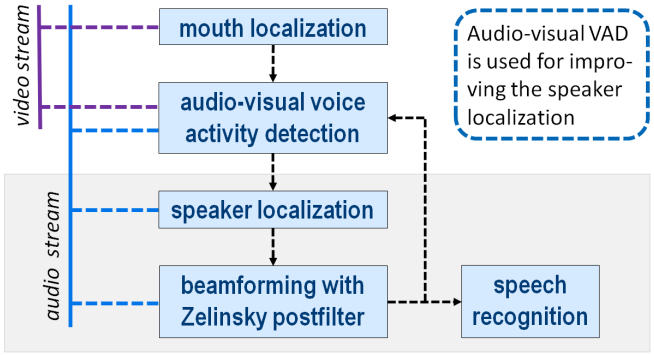


Fig. 1. *Overview of the proposed distant speech recognition system with audio-visual voice activity detection.*

Voice Activity Detection

For voice activity detection, the relative position of the speaker to the microphone array is estimated based on the multichannel cross correlation coefficient (MCCC) [4]. At the same time, the visible mouth region is identified by the face tracking system. After these localization steps, the signal of the desired speaker is extracted with a subband domain

This work was supported by the Federal Republic of Germany, through the Cluster of Excellence for Multimodal Computing and Interaction; and by the German Research Council (DFG), under the research training network IRTG 715 “Language Technology and Cognitive Systems”.

delay-and-sum beamformer [1, §13] steered towards the position of the speaker. This is followed by a postfilter to further reduce the noise. Subsequently, audio and video features are extracted and the speech and non-speech segments are determined by integrating the resulting features streams in the framework of a multi-stream hidden Markov model (HMM) decoder [5]. This provides a robust audio-visual VAD system.

Improved Speaker Localization

In order to improve the speaker localization, the non-speech segments identified by audio-visual VAD are used to estimate average per-channel noise spectra, which are then used to individually enhance the signal of each channel with power spectral subtraction. After that, the speaker positions are reestimated from the noise-subtracted array signal. This gives more accurate position estimates as the application of spectral subtraction reduces the influence of noise on the cross correlation. After reestimation of the speaker location, beamforming and postfiltering are reperformed and the resulting, enhanced audio signal is fed to the speech recognizer.

Evaluation

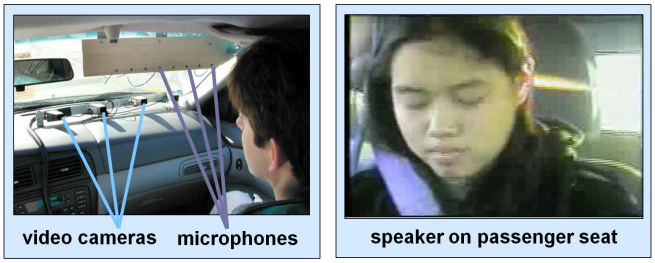
The effectiveness of the proposed DSR system is demonstrated through speech recognition experiments on the audio-visual car (AVICAR) [6] corpus. It should be mentioned that this is a very challenging corpus, which stands out in that it was recorded under real conditions – that is, in different cars driving with up to 55 miles per hour. Apart from adverse acoustic conditions with SNRs as low as -10dB, there are serious challenges on the visual side. Light conditions can change abruptly. The cameras are shaking. There are compression artifacts and the upper part of the face is occasionally occluded due to hair flying in the wind.

Balance

The remainder of this paper is organized as follows. In Section 2 we describe audio-visual feature extraction along with the face tracking system used for identifying the mouth position. This is followed by descriptions of the audio-visual VAD and robust speaker localization systems, in Sections 3 and 4. Experimental results are finally presented in Section 5.

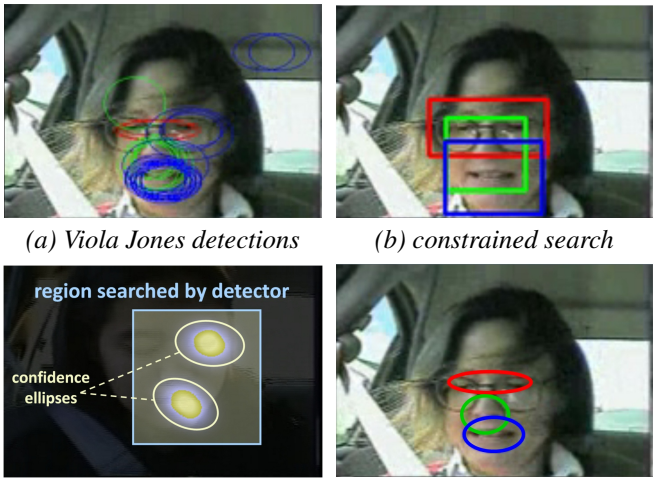
2. AUDIO-VISUAL FEATURE EXTRACTION

As mentioned before, audio-visual voice activity detection on the AVICAR [6] corpus poses a serious challenge. Next to adverse acoustic conditions, there is a variety of problems on the visual side, which caused many of the standard face and mouth detection algorithms to fail. But accurate detection of the mouth region is imperative for the visual front end. Hence, we developed our own robust face tracking system, a rough sketch of which is given in Section 2.1. After detection of the mouth region, visual features are extracted as described in Section 2.2. Audio features are extracted as explained in Section 2.3.



(a) recording setup *(b) image from one camera*

Fig. 2. *The AVICAR corpus. The image to the left shows the recording setup with four cameras on the dashboard and eight microphones on the sun visor. The image to the right shows one of the speakers from the perspective of the leftmost camera, including compression artifacts and lighting effects.*



(a) Viola Jones detections *(b) constrained search*

(c) determined search region *(d) output of the filter*

Fig. 3. *Mouth Localization with the Face Tracking System. The red, green and blue circles indicate eye, nose and mouth detections, respectively. The rectangles are search regions.*

2.1. Mouth Localization

The mouth localization system used in this work simultaneously tracks the eyes, nose and mouth of the speaker. As shown in Figure 3-(a), Viola-Jones based detectors [7] provide potential positions for each of the facial features, which are then integrated in a Bayesian filtering framework. That is achieved by tracking each facial feature with a bank of Kalman filters as sketched in Figure 4. Multiple observations (object detections) are treated by splitting filters [8], which, in contrast to the data association approaches taken in [9] and [10], truly allows the filter to simultaneously consider multiple concurrent hypotheses. As the splitting approach creates the problem of an exponentially growing number of filters, we use a merging step in order to reduce the number of filters, similar as originally proposed in [11] for multi-target tracking with a radar.

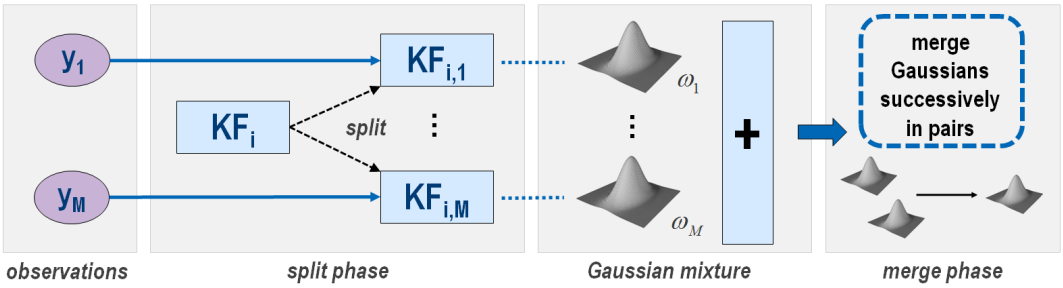


Fig. 4. *Tracking with multiple observations using the split and merge Gaussian mixture filter [8]. In the split phase, multiple observations (object detections) are handled by splitting the Kalman filters [8] and then assigning each of the resulting filters to one of the observations. In the merge phase, the computational complexity is reduced by merging the Kalman filters successively in pairs, until a predetermined number of filters is reached.*

In this work, merging is repeated until 10 filters remain. The Viola-Jones detectors are tuned to provide a large number of hypotheses with the aim of reducing detection failures. As the resulting increase in the number of detections, however, comes at the expense of a higher number of false alarms, gating [12] is used in order to efficiently suppress wrong hypotheses by discarding detections that lie outside the confidence ellipses predicted by the Kalman filters. In contrast to the original gating technique, we determine a rectangular region that includes all the confidence ellipses stemming from different filters of the same facial feature and then restrict the search area of the corresponding Viola-Jones detector to this rectangle, as portrayed in Figure 3-(b), (c).

The filters for different facial features interact by using a probabilistic scale and rotationally invariant face geometry model for calculating confidence scores of the detected feature positions. The same face geometry model is employed for inferring the positions of missing facial features as well as for further restricting the search space of the detectors, where the latter is achieved through intersection with the area in which the feature was predicted by the face geometry model. This helps in situations where features cannot be detected due to occlusion or adverse lighting conditions. As a last resort fail-safe mechanism, the tracking algorithm is reinitialized if the face confidence score is too low or if two facial features are missing in five successive frames.

2.2. Visual Feature Extraction

In order to explain the visual feature chain, let $\mathcal{I}_{\text{ROI}}(k,i,j)$ denote the intensity value of the (i,j)-th pixel of the 100×80 region of interest (ROI) around the detected mouth of the k-th video frame. Then, feature extraction starts by reducing illumination effects of the ROI through use of logarithmic intensity values and subsequent mean and variance normalization:

$$\mathcal{I}_{\text{log-norm}}(k,i,j) = \frac{\log \mathcal{I}_{\text{ROI}}(k,i,j) - \mu}{\sigma}. \qquad (1)$$

In this equation, μ and σ are the mean and variance of the log pixel intensities calculated over the entire utterance. The motivation for taking the logarithm is based on the fact that it converts multiplicative illumination effects into an additive bias term which can easily be removed with mean normalization. The result is shown in Figure 5.

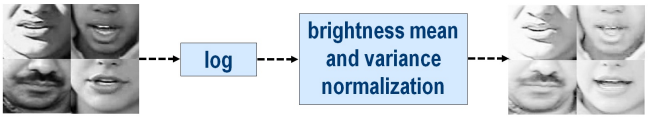
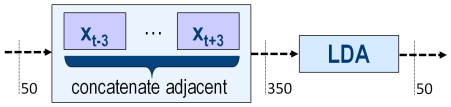


Fig. 5. *Mean and variance normalization on logarithmic pixel intensities reduces illumination and skin color effects.*

(a) static feature extraction reduces dimensionality to 50



(b) dynamic feature extraction captures motion dynamics

Fig. 6. *Visual Feature Extraction. Static features reduce the dimensionality of gray-value images while normalizing for inter-utterance variations. Dynamic features account for the dynamics of speech by stacking adjacent features and performing a second LDA on top of that.*

After constructing a vector from these normalized log pixel intensities, principle component analysis (PCA) is applied in order to reduce the dimension from 100×80 to 200. This is followed by a second mean normalization stage in which now every PCA coefficient is normalized independently. After PCA with mean normalization, linear discriminant analysis (LDA) is performed in order to further reduce the dimension to 50 and to improve the discriminability of speech and non-speech classes. The reason for using PCA as a pre-processing step to LDA is that it gives more stable results in large dimensions.

As the above features are incapable of capturing mouth movements, we concatenate the feature vectors of 7 adjacent frames and perform a second LDA on top of that, as proposed in [13]. This processing chain was found to give good results in a comparative study of different visual features [14]. The frame rate of the visual feature stream was finally adjusted to that of the audio feature stream by simply repeating the same feature vector.

2.3. Audio Feature Extraction

After localization of the speaker according to section 4.2, a delay-and-sum beamformer [1, §13] is constructed so as to steer a beam towards the estimated speaker position. This is followed by Zelinski post-filtering [15] in order to further reduce the noise. These speech enhancement steps are performed in the subband domain, after analysis with a DFT modulated filter bank, which is designed to minimize the in-band and residual aliasing terms individually rather than keeping the perfect reconstruction property [1, §11.7]. After enhancement, 13-dimensional MFCC features are extracted for each frame of speech; Cepstral mean and variance normalization are applied; and the final 39-dimensional feature vector is obtained by concatenating the normalized MFCC features with their first and second order derivatives (Δ and $\Delta\Delta$ features).

3. AUDIO-VISUAL VOICE ACTIVITY DETECTION

For VAD, the audio and visual feature streams from the front-end were combined within the framework of a multi-stream HMM [5]. In order to explain this in more detail, let $\mathbf{a}$ and $\mathbf{v}$ denote the audio and video features, respectively. Then, the likelihood of observing an audio-visual feature vector $\mathbf{av} = [\mathbf{a}^T, \mathbf{v}^T]^T$ at a hidden state j can be expressed as

$$b_j(\mathbf{av}) = b_{a,j}(\mathbf{a})^{\lambda_a} \times b_{v,j}(\mathbf{v})^{\lambda_v}, \qquad (2)$$

where λ_a and λ_v are stream exponential weights, and where the $b_{a,j}(.)$ and $b_{v,j}(.)$ are the observation likelihoods of the audio and visual models, respectively. Based on the results of preliminary experiments, we set $\lambda_a = 0.9$ and $\lambda_v = 0.1$. For the audio stream, we used a monophone acoustic model with subphonemes. That means, each monophone consisted of a left-to-right HMM with three hidden states. For the visual stream, we built Gaussian mixture models (GMM) for the speech and non-speech classes only, due to poor discriminability between different visemes (visual phonemes).

The audio and visual models were trained independently. But, in order to ensure synchronicity, the visual GMM training was bootstrapped from audio labels. During VAD, we used an audio-visual HMM with the same topology as the audio HMM. The observation likelihoods of the visual model were evaluated by mapping each phoneme to the speech GMM and by mapping silence to the non-speech GMM, as shown in Figure 7. Then, the observation probability of the audio-visual feature vectors were evaluated according to (2) and the alignment of the audio-visual features to speech and non-speech states was found with the Viterbi algorithm.

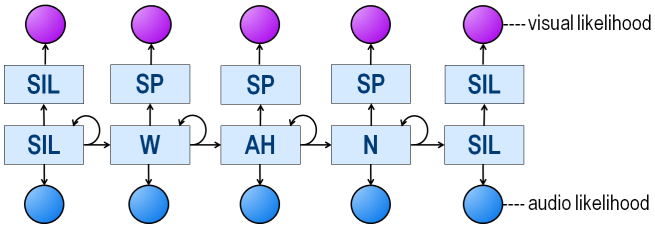


Fig. 7. *Part of the Multistream Hidden Markov Model used for audio-visual VAD. The graphical model shows the state sequence for the word "one". Acoustic states are mapped to speech (SP) and non-speech (SIL) visual states. Subphonemes are not shown for reasons of simplicicity.*

4. ROBUST SPEAKER LOCALIZATION

For acoustic localization of the speaker, we used the multi-channel cross correlation coefficient (MCCC) [4, 16]. The MCCC can be viewed as a generalization of the cross-correlation coefficient to the multichannel case. We considered using it as its robustness for time delay estimation in adverse environments has been demonstrated in [16]. Section 4.2 gives a more detailed description of the MCCC-based localization algorithm. It is used in both stages of the DSR system from Figure 6, that is audio-visual VAD and speech recognition. In the second stage, however, localization is performed after noise reduction of the individual channels, as described in Section 4.1.

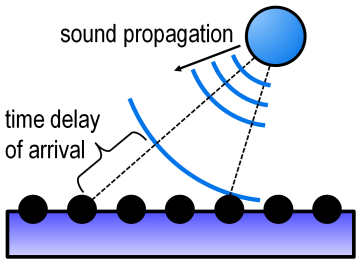


Fig. 8. *The arrival of sound waves at the microphone array introduces microphone-dependent time delays.*

4.1. Power Spectral Subtraction

Due to its simplicity, spectral subtraction (SS) [17] is one of the most widely used techniques for noise suppression. In this work, we use it in its power spectral form, as a preprocessing step to speaker localization. Denoting the desired speech and noise signals captured with the m-th microphone at frame k

and subband frequency bin f by $X_m(k,f)$ and $V_m(k,f)$, respectively, the observed spectrum in the power spectral domain can be approximated as

$$|Y_m(k,f)|^2 \approx |X_m(k,f)|^2 + |V_m(k,f)|^2. \quad (3)$$

Hence, given a noise power spectrum $\hat{V}_m(k,f)$ estimated from the non-speech segments identified by the audio-visual VAD system, the power spectrum of the desired signal can be obtained by subtraction:

$$|\hat{X}_m(k,f)|^2 = \max\left\{|Y_m(k,f)|^2 - \alpha|\hat{V}_m(k,f)|^2, \beta\right\} \quad (4)$$

In this equation, α is the overestimation factor [17] and β is the spectral floor [17]. Based on results of preliminary experiments, we set $\alpha = 4.0$ and $\beta = 0.01$. After subtraction, the clean speech spectrum was reconstructed by using the magnitude of (4) and the phase of the original signal. The estimated clean speech signal was obtained by transforming the spectrum back into the time domain. The relatively large overestimation factor led to aggressive noise removal, up to the complete elimination of heavily noise corrupted speech regions. This was found to improve the performance of speaker localization, although the phase of the noisy speech signal was used for reconstruction.

4.2. Localization based on the MCCC

For localizing the speaker with a linear equi-spaced array – as it is used in the AVICAR corpus [6] – the multichannel signal in direction of θ can be defined as

$$\mathbf{x}_M[n,\theta] = \begin{bmatrix} x_1[n] \\ x_2[n + d\sin(\theta)/c] \\ \vdots \\ x_M[n + (M-1)d\sin(\theta)/c] \end{bmatrix}$$

if the far-field assumption is made. In this equation, d is the distance between the microphones and c is the speed of sound. In order to calculate the MCCC, we need to compute a spatial correlation (covariance) matrix of observations over the entire utterance. The spatial correlation matrix can be expressed as

$$\mathbf{R}_M[\theta] = E\left\{\mathbf{x}_M[n,\theta]\mathbf{x}_M^T[n,\theta]\right\}, \quad (5)$$

where $E\{\cdot\}$ denotes the expectation operator. With this, the MCCC can be computed as

$$\rho_M^2[\theta] = 1 - \frac{\det\left(\mathbf{R}_M[\theta]\right)}{\Pi_{i=1}^M \sigma_i^2}, \quad (6)$$

where $\det(\cdot)$ stands for the determinant and where σ_i^2 is the i-th diagonal component of the spatial correlation matrix $\mathbf{R}_M[\theta]$ [4]. For the case $M = 2$, it can be readily confirmed that the MCCC is equivalent to the cross-correlation coefficient. Further calculating $\rho_M^2[\theta]$ for all possible directions of arrival θ, the angle of the speaker is obtained as the maximum of $\rho_M^2[\theta]$:

$$\hat{\theta}_M = \arg\max_\theta \ \rho_M^2[\theta]. \quad (7)$$

5. EXPERIMENTS

In order to evaluate the proposed system under realistic conditions, we performed a set of experiments on the phone number task of the AVICAR corpus [6]. This corpus stands out in that it was recorded in real cars, under five different conditions. In the IDL condition, the car is standing still with the engine running (idle). In the 35D, 35U, 55D and 55U conditions, the car is driving at 35 and 55 miles per hour, respectively, with the windows up (U) or down (D). The signal to noise ratio (SNR) varies between 15 and -10 dB, due to engine noise, wind, road noise from the tires as well as from other cars passing by. All of the speech data was recorded in English. Multichannel audio data is available for 87 speakers out of which about 60 are native speakers of American English [6]. Video data is available for 86 subjects.

5.1. ASR System and Setup

In the speech recognition experiments, the feature extraction of the ASR system was based on 13-dimensional mean and variance normalized MFCC features plus delta and delta-delta features. Cepstral mean and variance normalization were performed on the speech frames identified by standard energy-based voice activity detection. Speech recognition was performed with a word trace decoder, as described in [1, §7.1]. The state network used for decoding consisted of a precompiled weighted finite-state transducer, which was optimized as described in [1, §7.2]. For phone number recognition, we trained a monophone acoustic model with up to 64 Gaussians per state on the IDL digits and phone number tasks of the AVICAR corpus. The selected training material comprised 1282 single digits as well as 1298 10-digit phone numbers. In addition to single channel data, a certain amount of delay-and-sum beamformed data was added to improve the performance of the DSR system. As a test set, we used the the phone number task recorded in the 35D condition in which the car is driving at 35 miles per hour with the windows down. This set consisted of 1273 phone numbers with a total recording length of 108 minutes.

5.2. Results

Table 1 shows the word error rates (WERs) we obtained on the AVICAR corpus. The first row gives results for a single distant channel, that is, without array processing. The rows below show the results that were obtained after delay-and-sum beamforming with Zelinsky postfiltering. These rows differ only in how source localization was performed. Evidently, the best performance was achieved when audio-visual VAD controlled spectral subtraction was used as a preprocessing step to localization. In this case the WER was 7.1%, which compares to a WER of 8.9% with spectral subtraction based on audio-only VAD and to a WER of 11.6% for source localization without a prior speech enhancement step. The dif-

array processing	localization	WER
none	–	15.6
DSB + Postfilter	MCCC	11.2
DSB + Postfilter	MCCC + VAD-based SS	8.9
DSB + Postfilter	MCCC + AV-VAD-based SS	7.1
DSB + Postfilter	broadside assumption	8.5

Table 1. *Word error rates obtained on the AVICAR corpus. for a single distant microphone and after delay-and-sum (DSB) beamforming with a Zelinsky postfilter. The speaker position was either assumed to be perpendicular to the array (broadside) or it was estimated based on the multi-channel cross correlation (MCCC), with an optional spectral subtraction (SS) step. The average noise spectrum (required for SS) was estimated based on audio only (VAD) or audio-visual voice activity detection (AV-VAD), respectively.*

ference between audio-only and audio-visual VAD can be explained by the fact that our audio-visual VAD system mainly provides a lower false positive rate than the audio-only VAD system [14]. This prevents the cancellation of target speech signal components due to leakage into noise estimates. Now, comparing the above results to the fifth row of Table 1 reveals that the proposed localization method with audio-visual VAD is actually the only method that could improve over the simple assumption that the speaker is located perpendicular to the array (broadside). But even in this case we still get a relative improvement of 16.5% in WER.

6. CONCLUSIONS

We have described a new strategy for distant speech recognition, which uses audio-visual VAD in order to improve the performance of the speaker localization system. The effectiveness of the proposed approach was demonstrated through speech recognition experiments on a challenging audio-visual corpus, under realistic conditions.

7. REFERENCES

[1] Matthias Wölfel and John McDonough, *Distant Speech Recognition*, Wiley, New York, 2009.

[2] I. Almajai and B. Millner, "Using audio-visual features for robust voice activity detection in clean and noisy speech," *Proceedings of the European Signal Processing Conference*, pp. 988–993, Aug. 2008.

[3] T. Yoshida, K. Nakadai, and H. G. Okuno, "Two-layered audio-visual speech recognition for robots in noisy environments," *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 988–993, Oct. 2010.

[4] Jacob Benesty, Jingdong Chen, and Yiteng Huang, *Microphone Array Signal Processing*, Springer, 2008.

[5] S. Dupont and J. Luettin, "Using the multi-stream approach for continuous audio-visual speech recognition: Experiments on the M2VTS database," *Proceedings of the International Conference on Speech and Language Processing*, 1998.

[6] B. Lee et al., "AVICAR: Audio-visual speech corpus in a car environment," *Proceedings of Interspeech*, pp. 2489–2492, Oct. 2004.

[7] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, vol. 57, pp. 137154, Oct. 2001.

[8] F. Faubel, M. Georges, B. Fu, and D. Klakow, "Robust gaussian mixture filter based mouth tracking in a real environment," *Proceedings of the Visual Computing Research Conference (by the Intel Visual Computing Institute)*, Dec. 2009.

[9] C. Rasmussen and G. D. Hager, "Probabilistic data association methods for tracking complex visual objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, no. 10, pp. 560–576, June 2001.

[10] X. Ren, "Finding people in archive films through tracking," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2008.

[11] D. J. Salmond, "Mixture reduction algorithms for uncertain tracking," *Technical Report 88004, Royal Aerospace Establishment*, Jan. 1988.

[12] Y. Bar-Shalom and T. E. Fortmann, *Tracking and Data Association*, Academic Press, 1988.

[13] J. Luettin G. Potamianos, C. Neti and I. Matthews, "Audio-visual automatic speech recognition: An overview," in *Audio-Visual Speech Processing*, MIT Press, ISBN: 0-26-222078-4, 2006.

[14] M. Georges, "A comparative study of features for audio-visual speech recognition," M.S. thesis, Saarland University, Saarbrücken, Germany, 2010.

[15] C. Marro, Y. Mahieux, and K. U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Transactions on Speech and Audio Processing*, vol. 6, pp. 240–259, 1998.

[16] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: an overview," *EURASIP Journal on Applied Signal Processing*, pp. 1–19, 2006.

[17] P. C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, June 2007.