Speaker segmentation is an important task in multi-party conversations. Overlapping speech poses a serious problem in segmenting audio into speaker turns. We propose an audio-visual speech separation system consisting of an array microphone with eight sensors and an omnidirectional color camera. Multiple concurrent speeches are segmented by fusing the two heterogeneous sensors. Each segmented speech is further enhanced by a linearly constrained minimum variance beamformer. Regardless of co-existing wide-band sound sources and pictures of human in a reverberant environment the proposed system effectively separates multiple target speeches.

1. Introduction

Speaker segmentation is an important task not only in conversations, meetings, and task-oriented dialogues but also in many speech processing applications such as a large vocabulary continuous speech recognition system, a dialog system, and a dictation system. Overlapping speech occupies a central position in segmenting audio into speaker turns [1]. Results on segmentation of overlapping speeches with an array microphone are reported by using binaural blind signal separation [2], dual-speaker hidden Markov models [3] and speech/silence ratio incorporating Gaussian distributions to model speaker locations with time delay estimates [4]. Speaker tracking using a panoramic image from five video stream input and an array microphone is reported in [5, 6]. These are the two extremes of concurrent speaker segmentation; one approach depends solely on audio information while the other depends mostly on video.

As for audio part all the works in [2–6] are based on time-delay estimates (TDE). However, even though there are a weighting function from a maximum likelihood approach and a phase transform to cope with ambient noises and reverberations, TDE-based techniques are vulnerable to contamination from explicit directional noises [7]. On the other hand, signal subspace methods have an advantage of adopting multiple-source scenarios. In addition, they are relatively simple and clear, and also provide high resolution and asymptotically unbiased estimates of the angles for wide-band signals [8,9]. Thus, we adopt a subspace approach as an audio sensor. As for video part, we have an omni-directional color camera with a 360° field of view so that all human are viewed simultaneously. To find multiple human, two features are used: skin color and human shape. Skin regions have nearly uniform colors, so that the face and hand regions can be easily distinguished using color segmentation. To decide whether the candidates from skin-colored blob detection is a human or not, three shape models of human upper-body are incorporated with it.

Because the subspace based audio sensor is as accurate in its performance as the video sensor, the proposed method takes a happy medium of the two extremes. Therefore, the proposed sensor fusion helps one sensor compensate for the other sensor’s weakness. In addition to segmenting multiple speeches into speaker turns, separating each speech is possible using spatial information of the target and temporal characteristics of interferences and noises. The rest of this paper is organized as follows. Section 2 describes a subspace approach to localize multiple sound sources on real streaming audio data. Section 3 delineates detection of multiple human based on skin color and human shape models. The audio-visual system to fuse two heterogeneous sensors is explained in section 4. Segmentation and separation of multiple speakers are deployed in section 5. Experiments, their evaluations, and discussion about the proposed method are given in section 6.

Figure 1: An array microphone with eight sensors and an omni-directional color camera.
2. Wide-band sound source localization

The array microphone depicted in Fig. 1 is an equipment of speakers’ angle estimator because it is isotropic in azimuth and can find the angles of sound sources of all directions. The subspace approach is based upon a spatial covariance matrix from the observed signals via ensemble average over an interval assuming that their estimation parameters i.e. the angles between an array microphone and each speaker are fixed.

The observed data is given by an $m$-dimensional vector ($m$ sensors) in frequency domain as

$$z_k(t) = A(\theta, f_k)s_k(t) + n_k(t),$$

where $z_k(t)$, $s_k(t)$, and $n_k(t)$ are an observation ($m \times 1$), a source ($d \times 1$), and a measurement noise ($m \times 1$) vector in the $k$th frequency bin at discrete time $t$, respectively, and $A(\theta, f_k)$ is a transfer function matrix consisting of steering vectors, $a_k(\theta)$, which represent attenuation and delay reflecting the propagation of the signal source at direction $\theta$ to the array in the $k$th frequency bin.

A spatial covariance matrix for observations is obtained for every consecutive frames by $R_k = E\{z_kz_k^H\}$, where $\cdot^H$ denotes the Hermitian transpose. A spatial covariance matrix $N_k$ was pre-calculated when there were no explicit directional sound sources. Therefore, solving the generalized eigenvalue problem, $R_k \cdot E = N_k \cdot E \cdot \Lambda$ results in a generalized eigenvalue matrix, $\Lambda$ and its corresponding eigenvector matrix, $E = [E_S | E_N]$, where $E_S = [e_1^S, \cdots e_d^S]$ and $E_N = [e_1^N, \cdots e_m^N]$ are matrices of eigenvectors which span a signal subspace and a noise subspace, respectively. A spatial spectrum for the $k$th bin is then

$$P_k(\theta) = \frac{a_k^H(\theta) a_k(\theta)\Lambda}{a_k^H(\theta)E_N E_k^H E_k(\theta)},$$

where $a_k(\theta)$ is the steering vector. Spatial spectrum, $P(\theta)$ indicating the possibility of a sound source at direction $\theta$ is obtained by averaging the $P_k(\theta)$’s over frequency bins.

A peculiar feature of this subspace approach is in the steering vector set. Though the $a_k(\theta)$’s can be calculated geometrically according to the array configuration, we estimated them every $5^\circ$ spacing. Two main advantages arising from the use of the estimated steering vectors are an increased spectrum resolution and an automatic array calibration. The spatial spectrum using the estimated steering vectors has higher resolution and is peakier than that using geometrically calculated ones. Furthermore, the direct estimation cancels the mismatch between the ideal and the real array arrangements.

3. Multiple human detection

An input color image is converted into two images: a color-transformed gray image and an edge image (see Fig. 2). The first is generated by a color normalization, $r = \frac{R}{R+G+B}$, $g = \frac{G}{R+G+B}$, and $r + g + b = 1$, followed by a color transform. The color transform is expressed as a 2D Gaussian function, $N(m_r, \sigma_r; m_g, \sigma_g)$, where $(m_r, \sigma_r)$ and $(m_g, \sigma_g)$ are mean and standard deviation of red and green component, respectively. The normalized color reduces the effect of the brightness which significantly affects a color perception processing, while remaining the color components. A transformed pixel has a high intensity when its value gets close to skin color. The second is the average of three edge images, red, green, and blue. Based on the size and the center-of-gravity of each skin-colored blob in the color-transformed gray image we obtain size-normalized candidates for human upper-body in the edge image.

For human shape matching, we made three shape model images of human upper-body in accordance with human poses consisting of a front, a left-side, and a right-side view. To calculate the similarity between a shape model image and the candidate edge image we measure the Hausdorff distance between them. The Hausdorff distance defines a measure of how much the sets are similar to each other [10]. It is composed of two asymmetric distances. Given two sets of points, $A = \{a_1, \cdots, a_n\}$, the shape model image and $B = \{b_1, \cdots, b_n\}$, the candidate edge image, the Hausdorff distance is defined as

$$H(A, B) = \max (h(A, B), h(B, A)),$$

where $h(A, B) = \max_{a \in A} \min_{b \in B} \|a-b\|$. The function $h(A, B)$ is called the directed Hausdorff distance from $A$ to $B$. It identifies the point that is farthest from any point of $B$, and measures the distance from $a$ to its nearest neighbor in $B$. In other words, one of the directed distance is small when every point in $A$ is close to some point in $B$. When both are small then the candidate edge image and the shape model image look like each other. The triangle inequality of the Hausdorff distance is particularly important when multiple stored shape model images are compared to an edge image. Thereby, we can detect the human upper-body and his/her pose as well. Hence, the proposed method detects multiple human in cluttered environments that have skin-color noise, illumination change, and complex background (see Fig. 2).

![Image](image_url)

Figure 2: Skin color and shape models allow the detection of multiple humans. Circles on the input image show the detection results.
4. The proposed sensor fusion strategy

An observation data set $\mathbf{Z}(t)$ consists of multi-channel audio stream, $\mathbf{z}_a(t)$ with its element $z_k(t)$, $k = 1, \ldots, m$, observed by the $k$th microphone in time domain and an omni-directional vision data, $\mathbf{z}_v(t) = I(\rho, \theta; t)$.

$$
\mathbf{Z}(t) = \{\mathbf{z}_a(t), \mathbf{z}_v(t)\} \quad (4)
$$

Because the two sensors are assumed to be independent, the proposed posterior probability of $\theta$ given the observation data is

$$
p(\theta|\mathbf{Z}(t)) \propto p_a(\theta|\mathbf{z}_a(t))p_v(\theta|\mathbf{z}_v(t)), \quad (5)
$$

We propose that a normalized version of the spatial spectrum, $P(\theta)$ qualifies itself for a probability density function (pdf) of the audio sensor,

$$
p_a(\theta|\mathbf{z}_a(t)) = P(\theta) / \sum_{l=1}^{L} P(\theta_l). \quad (6)
$$

On the other hand, a pdf of the video sensor is proposed as a Gaussian mixture model of 1D Gaussian functions each centered at the center-of-gravity $\theta_i$ of detected human $i$. Its variance $\sigma_i^2$ is an increasing function of the angular range of the detected human,

$$
p_v(\theta|\mathbf{z}_v(t)) = \sum_i \alpha_i N(\theta_i, \sigma_i^2), \quad (7)
$$

where $\alpha_i$ is a mixture weight such that $\sum_i \alpha_i = 1$. Because a small value of the distance means that the candidate matches one of the shape model images well, the mixture weight is a decreasing function of the Hausdorff distance.

Two major categories exist in sensor fusion strategies. One is that the sensors are fused in a feature level. For example, we have two logical sensors in the human detection, a skin color detector and a shape matcher. The former has a color discriminant which expresses the probability that a given pixel is skin-colored. The latter has a distance measure to evaluate how closely a given image patch resembles one of the shape model images. Basically feature level sensors are processed in parallel. However, to fast up the speed of calculation, the latter is used as a verifier in this case. The other is that the sensors are fused in a decision level. By fusing the feature level sensors, a pdf of a decision level sensor comes out. For example, the pdf of the video in (7) falls in this category. Note that segmenting speakers mainly relies upon the azimuth of them, the pdf of the video sensor is reduced to a 1D Gaussian mixture model in this case. The pdf for the audio in (6) is another example. From the $m$ sensor data the probability of existing unknown sound sources at each direction is obtained. Fusing the audio and video sensors as in (5) we can classify the active speakers by detecting the azimuths that produce posterior probabilities exceeding a threshold. Hence, the proposed audio-visual speaker localization algorithm effectively finds multiple active speakers regardless of co-existing wide-band sound sources and human pictures.

![Figure 3: Individual sensor pdfs and the posterior probability are shown when there are three speakers (of which two speakers are actually speaking), a loudspeaker playing a music, and a picture of human.](image)

5. Multiple speaker separation

A linearly constrained minimum variance beamformer (LCMVBF) is used to separate each target speech from the segmented multiple concurrent speeches. A use of the beamformer poses a serious problem of cancelling the target speech due to mismatch between actual and presumed steering vectors. In general, neither the actual steering vector nor the target-free covariance matrix is tractable in our system either. Instead, we focus on obtaining the target-free covariance matrix precisely. The proposed audio-visual fusion system described in section 2 through 4 notices the beamformer very accurately the ef

\[
\mathbf{W}_k = \frac{(\mathbf{R}_k + \lambda \mathbf{I})^{-1} \mathbf{a}_k(\theta_0)}{\mathbf{a}_k^T(\theta_0)(\mathbf{R}_k + \lambda \mathbf{I})^{-1} \mathbf{a}_k(\theta_0)}, \quad (8)
\]

where $\theta_0$ is a target direction and $\lambda$ is a diagonal loading factor. $\mathbf{R}_k$ is a covariance matrix in the $k$th frequency bin for target-free intervals. Note that the diagonal loading factor, $\lambda I$ further mitigates the cancellation of target signal due to slight mismatch of steering vectors.
6. Experiments and Discussion

A real-time speaker localization module based on the proposed sensor fusion has been implemented on a standard Pentium IV 2.5GHz Desktop computer. It has further been integrated with an LCMVBF. Our goal is to separate each speaker’s utterance from the mixture of multiple concurrent speeches by the co-work of the proposed speaker localization and the LCMVBF. The target free covariance matrix plays an important role in the beamformer. The proposed audio-visual speaker localization algorithm helps the beamformer accurately segment the target free intervals on multiple active speeches.

Experiments were done in a reverberant and cluttered environment. As shown in Fig. 2(a) there are three people, a picture of human, and a loudspeaker. Audio input streams are segmented into frames of size 500ms. The current frame is overlapped with the previous frame by 250ms. Thus, the spatial spectrum \( P(\theta) \) is calculated every 250ms. From the image sequence via the omnidirectional video camera the proposed detector finds multiple human at an average rate of 10 frames per second. Fusing the two, we can identify which speakers are active by rejecting a noisy sound source of the loudspeaker and a picture of human that is not a real human. An example of the separation result is depicted in Fig. 4.

Figure 4: Multiple speaker separation result. Three concurrent speeches and a music are mixed at the array microphone input on the top row. Each speech separated by the linearly constrained minimum variance beamformer is shown on three bottom rows. Demo sounds are available via http://myhome.naver.com/flyers/publications.html

The proposed method has several strong points over other works reported so far. The first is that it is robust against noises because a subspace method with elaborately measured steering vectors is incorporated into the whole system. The second comes from three shape models for human upper-body. Human upper-body is more adequate than the whole human body because human trunk and legs are often occluded by other objects in a cluttered environment. Moreover, pose estimation is possible because we also adopt profiles as human shape models. When we improve its ability to human tracking, the pose information is especially useful for particle filtering. The robustness against steering vector mismatch deserves the third merit. Since the actual steering vectors are unavailable in practice, the problem of cancelling target speech can be overcome by target-free covariance matrix with diagonal loading method, which, in turn, is possible by the accurate segmentation. Intuitive and simple sensor fusion strategy is the fourth advantage. By the audio-visual sensor fusion, we can effectively keep a loudspeaker and a picture of human from active speakers. Finally, it is directly applicable to a large vocabulary continuous speech recognition system or a dictator for distant talk to make an automatic meeting records. As for the speech recognition system, the proposed method serves as not only a speech enhancer but also an end point detector.

7. References