KNOWING WHO TO LISTEN TO IN SPEECH RECOGNITION: VISUALLY GUIDED BEAMFORMING

Udo Bub[1] *Martin Hunke*[1,2] *Alex Waibel*[1,2]

Interactive Systems Laboratories
[1]Carnegie Mellon University, Pittsburgh PA, USA
[2]University of Karlsruhe, Karlsruhe, Germany

ABSTRACT

With speech recognition systems steadily improving in performance, freedom from head–sets and push–buttons to activate the recognizer is one of the most important issues to achieve user acceptance. Microphone arrays and beamforming can deliver signals that suppress undesired jamming signals but rely on knowledge where the signal is in space. This knowledge is usually derived by identifying the loudest signal source. Knowing who is speaking to whom and where should however not depend on loudness, but on the communication purpose. In this paper, we present acoustic AND visual modules that use tracking of the face of a speaker of interest for sound source localization and beamforming for signal extraction. It is shown that in noisy environments a more accurate localization in space can be delivered visually than acoustically. Given a reliable location finder, beamforming substantially improves recognition accuracy.

1. INTRODUCTION

An essential requirement for natural and efficient human–computer interaction is the possibility of hands–free information exchange. For applications such as video conferencing it is desirable to allow partipiciants to move freely in a room while a system of sensors keeps track of the person of interest and enhances speech and other information modalities of this individual. This person shall not be physically encumbered by carrying any sensors like a closetalking microphone. Similar to inter–human communication the focus of attention should not be distracted by jamming noise.

In many studies microphone arrays have proved to be adequate for effective sound separation by means of various techniques of beamforming using spatial information (e. g. [1, 2]). Knowledge about the sound location is gained acoustically by detecting the source of the loudest speaker (e. g. [4]). Problems occur while tracking moving talkers in real–life communication situations including pausing or jamming speakers, so that these systems do not meet the above requirements.

Considering visual aspects to locate the speaker's position overcomes these limitations. In this paper we propose a combination of a visual tracking technique with a multi–microphone array in order to achieve robust speech recognition in noisy environments involving a moving speaker. A face locating and tracking system constantly determines the speaker's position by extracting visual features with a camera to guide the beam of the microphone array. We compare a purely acoustic set–up with the above described multimodal scenario and show that the latter one outperforms conventional techniques in noisy surroundings.

2. LOCATING AND TRACKING

Two different techniques for locating and tracking of a person of interest are presented. First we use a cross–correlation based acoustic speech tracker which exploits phase delay information with a microphone array. Then a visual face tracking system is introduced using features like color, shape, and movement.

2.1. Acoustic Localization

We use a linear microphone array consisting of 15 sensors. This set–up allows for signal recordings in the half plane in front of the array, i. e. the height of talkers is neglected. Localization of sound sources in general can be done by experimental measurement of differences in sound arrival times between several sensors [5]. Using these delays one can triangulate the coordinates of the source.

Following are two sequences $u_i[k]$ and $u_j[k]$ that are delayed replicas of one another. This delay equals the location k_{ij} of the maximum of the cross–correlation function

$$\phi_{ij}[k] = \sum_{m=-\infty}^{+\infty} u_i[k+m]u_j[m].$$

In our case $u_i[k]$ and $u_j[k]$ are the resulting sequences from an analog to digital conversion of two microphone signals at a sampling period T. Thus $k_{ij}T$ is an approximation of the time continuous propagation delay of the sound wave between the two sensors in low–noise environments.

In the following application the signals are sampled at a rate of 16 kHz and thus the delay can be only determined in multiples of $\frac{1}{16}$ ms which causes a too high sampling error for the desired estimations. To overcome this problem [4] proposes an interpolation of the cross–correlation function. We achieve this by 8 times zero–padding followed by an appropriate FIR lowpass filtering. Now we can more accurately determine the approximation of the theoretical phase delay in multiples of $\frac{1}{128}$ ms.

0-7803-2431-5/95 $4.00 © 1995 IEEE

Prior to computing the cross–correlations, the signals have to be windowed. Since speech signals are approximately stationary over a time period of around 20 ms, a significantly longer interval must be considered to avoid periodical secondary maxima in the correlations. A rectangular window of a length corresponding to 250 ms was found suitable.

Theoretically as few as three microphones and thus two delays would be sufficient to determine a talker's position. Since we have a much larger number at our disposal we use the redundant information to eliminate the effects of noise. The 14 delays between an arbitrary reference microphone and the remaining other sensors are determined. From each of the 91 resulting independent microphone combinations the location of the loudest sound source is triangulated. The median value of the resulting series of localizations is used as the final value. This algorithm is repeated every 250 ms which allows the tracking of a speaker moving in a natural way.

2.2. Visual Localization

To determinine a speaker's position by considering visual information sources we developed a face tracker for localization of human faces using a standard camcorder mounted on a pan tilt unit (PTU) allowing vertical and horizontal movement. In addition the system controls the zoom lens to maintain a constant resolution of the speaker's face in the camera image enabling feature extraction independent of the distance of the speaker to the camera.
The system operates in two modes, both considering color, movement, and shape as basic features. In the *locating mode* the system locates arbitrary faces. Among all located faces the the biggest one is assumed to be the speaker's face and is selected for further tracking in the *tracking mode*.

While tracking, the system learns specific features of the observed face and adjusts rapidly to changing lighting situations. A detailed description of the entire system is given in [6]. The system's output consists of the cartesian coordinates of the speaker's position which are used to adjust the PTU and zoom lens and to guide the beamforming.

Though skin–colors seemingly cover a considerable range, they mainly differ in brightness of the reflected colors and less in color itself. Using normalized colors by dividing the rgb–values by their sum greatly reduces the influence of brightness. The colors occuring most frequently in a sample color distribution of faces of 30 persons with all skin–types are considered as face–colors. This Face Color Classification (FCC) gives a first approximation of areas possibly containing faces. If detected, motion gives additional clues about an object's shape and helps distinguishing between the observed object and the background.

Motion gives additional clues about the object's shape and helps distinguishing between the observed object and the background. To be capable of tracking non–moving speakers, motion is not considered as a necessary feature. Figure 1 shows an original image and the largest located moving object with skin–color.

After locating an object the FCC is adjusted to the colors actually occuring in the detected face by computing a new color distribution. By repeating this procedure after each frame the system rapidly adjusts to changing illumination. To avoid misclassifications of arms, hands or other skin–colored objects as faces, the shape of all found objects is considered by neural networks.

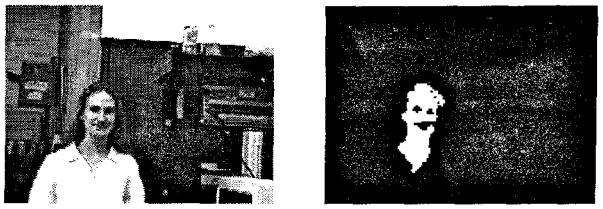

Figure 1. Camera image and extracted largest skin–colored object.

The area around the last location of a face is fed into the input retina of two neural networks. The first network estimates the position of the largest skin–colored object with face–like shape. The area around this position is passed to a second network, estimating the size of the now centered object. Using two networks for this task limits the size estimation to centered objects and leads to better results. The output of the network controls the PTU and zoom lens. The position and size of the face within the image are merged with the PTU position and zoom lense adjustment to determine the Cartesian coordinates of the speaker's position guiding the beamforming.

By using the FCC as network input the network performance could be strongly enhanced compared to gray scale or normalized color input images. Additionally retraining of the network after changing camera, framegrabber, or illumination, is not necessary because all color dependencies are bound to the FCC.

The networks were trained with back–propagation, using a training set of 5000 example images containg 24 persons with different sex, age, hair style, skin–color, etc. The faces were artificially scaled, shifted, and superimposed with different backgrounds, so that the position and size of a face in each image were known and could be used as desired output of the network during training. 20% of the images did not contain faces, enabling the network to recognize a face disappearing out of the camera image.

3. BEAMFORMING

The function of a beamformer is to enhance sound originating from a particular location. We make use of the simple but very effective method of the *delay and sum* beamformer (e. g. [1, 2, 3, 7]).

This spatial filter needs knowledge about the phase delays that can be observed between the reference microphone and the other remaining sensors for sound coming from the desired location. To steer the beam toward this location these delays are compensated at each sensor before summing up the signals. In doing so, the aligned signals orignating from the desired point are enhanced while sound coming from other locations is not in phase and thus even deterio-

rated by interference.

Both the techniques described in section 2 deliver the coordinates of a point of interest. The characteristic delays for such point are determined mathematically assuming a spherical form of speechwaves[1]. We used a computationally favorable implementation for digital processing described in [7].

4. EXPERIMENTS

We distinguish between two noise situations in our laboratory for the experiments: ***background noise*** and *competing noise*. In the first case the audio signal is randomly deteriorated by various jamming low level signals (humming fans, etc.), but the main speaker is clearly dominant. In the case of competing noise a second sound source which competes with the speaker has been installed, emanating music from a radio at a high output level. Speech registered with a single array element is hardly intelligible even for human listeners. The lab has blank concrete walls and no acoustic favorable arrangements have been taken on the room. Altogether these facts ensure that the following evaluations have been made based on real data in realistic environments.

The array spans a length of 112 cm and the spacing of the 15 sensors is non–uniform, ranging from 3.5 cm to 14 cm. Facing the array, the outermost left microphone is chosen as the origin of a reference coordinate system. Its x–axis is parallel to the array (pointing towards the center of the array) and the y–axis is perpendicular. The competing sound source is located at $x = -20$ cm, $y = 130$ cm and the recording room allows talker movements between 100 cm to 350 cm from the array.

We used the recoginzer of the JANUS-System [9] to determine recognition rates on continuous speech. Since this recognizer is trained on speech recorded with a closetalking microphone, the acoustic models have to be adjusted to the acoustic characteristics of the microphone array. This can be done by either retraining the recognizer or – as in our case – by mapping the mel–scale data of the microphone array to those of the closetalking microphone with the aid of neural networks [8].

The Resource Management is chosen as database and a speaker uttered 45 sentences for each of the evaluations of recognition rates.

4.1. Localization

The evaluation of the accuracy of acoustic localization is based on a series of 230 measurements for talkers sitting on the same static chair. Table 1 shows the results, where σ is the series's standard deviation from the given real location.

	background noise		competing noise	
	x [cm]	y [cm]	x [cm]	y [cm]
Real location	160.0	150.0	160.0	150.0
Deviation σ	12.6	16.0	112.7	56.0

Table 1. Accuracy of acoustic localization.

[1]These delays are unique for a spot in the area of interest. Therefore it might be more appropriate to talk about *spotforming*.

As can be seen, the proposed localization algorithm works very accurately with background noise because the situation offers a dominant signal. Interestingly the localization parallel to the array works better than perpendicular to it.

The accuracy deteriorates considerably with competing noise because the localization algorithm is now distracted by the jamming sound source. Since the ratio of loudness between talker and jammer varies over time the location finder jumps back and forth or (as in most cases) chooses a location somewhere between both sound sources. The standard deviation for competing noise in table 1 is higher on the x–axis than on the y–axis because the geometric distance between talker and jammer is bigger in x–direction (180 cm) than in y–direction (20 cm).

The accuracy and reliability of the system for visual localization were evaluated on test sequences of over 2000 images of 7 persons with different skin–type in front of different backgrounds. The persons were asked to perform arbitrary movements, to stand up and sit down frequently, so that the PTU was forced to movement. The face could be located dependent on the sequence in 96% to 100% of all images. The average difference of the actual position of the face and the system's output was less than 10% of the face size. The system is working in real time and delivers approximatly 20 positions per second. It is obvious that this technique is not susceptible to any kind of acoustic noise.

4.2. Beamforming

Table 2 shows the improvement of recognition rates using beamforming seperately without automatic sound source localization in background noise. The talker was not allowed to move and the beam was controlled manually.

The previously mentioned neural network is used for channel adaptation to avoid a retraining of the recognizer and causes already a remarkable boost in recognition accuracy. In comparison with the single element located in middle of the array, beamforming increases the recognition rate by 22.5%.

Single microphone without mapping	15.5%
Beam without mapping	21.2%
Single microphone with mapping	58.3%
Beam with mapping	79.8%
Closetalking microphone	88.0%

Table 2. Word accuracy with background noise and static talker.

The effectiveness of beamforming for moving talkers depends largely on the location of the sound sources relative to each other. However, the improvement provided by beamforming in our application is clearly audible in all situations.

4.3. Acoustically vs. visually guided beamforming

Figure 2 illustrates the set–up of the entire system. The beamforming is guided either by acoustic or by visual localization of the speaker's position. Then the extracted time domain speech samples are transformed into mel–scale coefficients and after neural mapping transmitted to the speech recognition system.

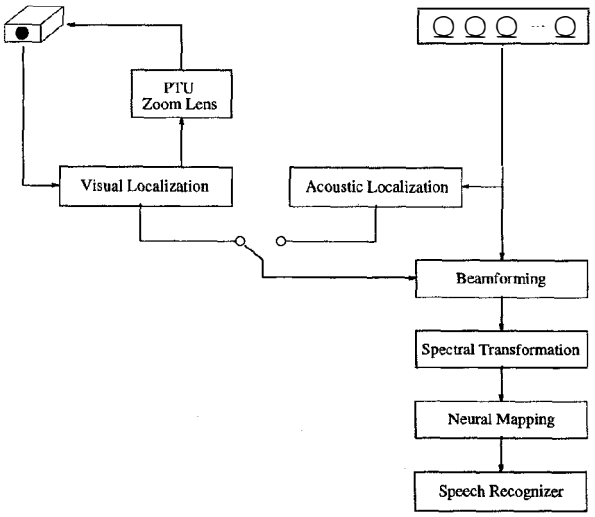

Figure 2. Acoustical and visually guided beamforming.

Table 3 presents the improvement of recognition rates by visually guided beamforming. The evaluations include neural mapping (not necessary for the closetalking microphone). It shows that visually and acoutically guided beamforming lead approximately to the same recognition rate in background noise because both localization mechanisms work accurately. In comparison with table 2 it can be seen also that the performance of the system decreases due to the fact of a moving speaker instead of a static one.

Competing noise impairs the recognition performance based on a single microphone severely. Acoustically guided beamforming improves the accuracy, but suffers from the distraction of the array's focus by the second sound source.

	backg. noise	comp. noise
Single microphone	59.8%	14.5%
Acoustically guided beam	69.5%	43.4%
Visually guided beam	68.9%	54.6%
Closetalking microphone	88.1%	88.4%

Table 3. Word accuracy with moving speakers.

Only visual localization works accurately enough in competing noise and thus visually guided beamforming shows a remarkable improvement in recognition accuracy compared to the single microphone and the acoustically guided beam.

5. CONCLUSION

For future applications like videoconferencing it is desirable to have hands-free systems offering robust speech recognition in noisy environments.

We have demonstrated an improvement of speech recognition by means of multimodal clues allowing for moving speakers in realistic surroundings without usage of a closetalking microphone.

6. ACKNOWLEDGEMENTS

This research was sponsored by the Department of the Navy, Office of Naval Research under Grant No. N00014–93–1–0806. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U. S. Government. The authors would like to thank especially Thomas Sullivan from CMU for much helpful advice about microphone arrays.

REFERENCES

[1] van Compernolle D., Ma W., Xie F., van Diest M.; *Speech Recognition in Noisy Environments with the Aid of Microphone Arrays*; Speech Communication; vol. 9, no. 5-6; 1990

[2] Flanagan J. L., Johnston J. D., Zahn R., Elko G. W.; *Computer Steered Microphone Arrays for Sound Transduction in Large Rooms*; Journal of the Acoustical Society of America; vol. 78, no. 5; 1985

[3] Flanagan J.L., Berkley D.A., Elko G.W., West J.E., Sondhi M.M.; *Autodirective Microphone Systems*; Acustica 73; 1991

[4] Silverman H. F., Kirtman S. E.; *A Two Stage Algorithm for Determining Talker Location from Linear Microphone Array Data*; Computer Speech and Language; vol. 6, no. 2; 1992

[5] Monzingo R. A., Miller T. W.; *Introduction to Adaptive Arrays*; Wiley and Sons; 1980

[6] Hunke H. M.; *Locating and Tracking of Human Faces with Neural Networks*; Technical Report CMU-CS-94-155, CMU, Pittsburgh USA; 1994

[7] Pridham R. G., Mucci R.A.; *Digital Interpolation Beamforming for Lowpass and Bandpass Signals*; Proceedings of the IEEE; vol. 67, no. 6; 1979

[8] Che C., Lin Q., Pearson J., de Vries B., Flanagan J. L.; *Microphone Arrays and Neural Networks for Robust Speech Recognition*; ARPA Human Language Technology Workshop; 1994

[9] Woszczyna M. et al.; *JANUS 93: Towards Spontaneous Speech Translation*; IEEE Intern. Conf. on Acoustics, Speech and Signal Processing; 1994