Improving Videophone Subjective Quality Using Audio Information

A. Vahedian, J. Arnold, M. Frater, M. Cavenor, L. Godara

School of Electrical Engineering, University of New South Wales, Australian Defence Force Academy, Canberra, ACT 2600, Australia

ABSTRACT: This article presents a new technique which uses audio information to achieve more efficient video coding for videophone and videoconferencing applications. The direction of arrival of the audio signal at an array of microphones is used to estimate the position of the speaker's lips such that the quality of the video reconstruction can be enhanced in this crucial area. Once this estimation is performed, then a two- or three-stage quantization strategy is applied to the video information which results in the compression of the subjectively more important parts, i.e., the lips and the face of a speaker, with lower distortion. Algorithms for audio source location using the speech signals received at an array of microphones form an important part of our approach. The proposed new technique is compatible with all existing video compression standards and is much easier to implement than previously proposed techniques. © 1999 John Wiley & Sons, Inc. Int J Imaging Syst Technol, 10, 86–95, 1999

I. INTRODUCTION

Real-time conversational video telecommunications services such as videophone or videoconferencing are becoming ever more important as a substitute for face-to-face meetings. While recent innovations in video technology have been aimed at reducing the relatively high data rate associated with these services, it is recognized that such reductions cannot be made at the expense of reduced picture quality. In particular, our day-to-day experience suggests that the correspondence between the spoken word and the movements of the speaker's lips is an important issue, as we feel uncomfortable and lose concentration when, for example, a film soundtrack slips out of synchronism with the picture. In the specific case of videoconferencing, this sound–lip movement synchronization becomes vital to the success of any interaction between the conference participants. This observation is supported by the results of many psychological studies in the area of visually augmented speech recognition (Brook and Summerfield, 1983; Dodd and Campbell, 1978). However, the available data rate used for videoconferencing is generally insufficient to transmit the service without any visible coding artifacts. As a result, one of the perceived weaknesses of existing services is the poor picture quality achieved, especially around the face of a speaker.

From an information communication point of view, not all areas of a videophone or videoconferencing picture are equally important to a viewer. In a videoconference, for example, the quality with which the speaker's face is reproduced is much more important than the quality of the stationary background.

Correspondence to: A. Vahedian

Given the fact that the only part of the coding process which introduces distortion is the quantization stage, the strategy employed to quantize the different areas of the image plays a crucial part in determining the subjective quality of the received video information. Provided any new quantization strategy complies with existing international standards and does not require additional network resources, significant increases in subjective quality should be possible—for example, by allocating a larger part of the total bit rate for coding the lips and face of the speaker, while using a smaller part for the rest of the image. The problem to be solved in adopting this strategy is one of locating the significant region of the image that contains the face and, most important, the lips.

The article is organized as follows. Section II provides an overview of the problems associated with locating a face within an image and introduces a new approach that involves the use of audio information to locate the speaker's lips. In Section III this lip position estimation technique is described in detail along with the array of microphones needed for its implementation. Experiments carried out to evaluate the technique are described in Section IV, while Section V contains the results obtained. Discussion on the overall performance of the scheme together with suggestions for further work are described in Section VI, while conclusions are presented in Section VII.

II. OVERVIEW OF FACE RECOGNITION TECHNIQUES

Automatic recognition and identification of human faces and facial features in video images as well as tracking these features over time are of interest in many applications. These include object-based image coding, model and shape coding, security control, expression recognition, and intelligent human–machine interaction. While the task of locating and identifying human faces is very easy for humans, no automatic system can currently carry out this task reliably. There are issues yet to be solved, such as variable orientation, size, and partial occlusions of the face. Although the problem has been addressed by many researchers who have attempted to use properties of the picture to identify the location of a face, the issue of reliable face detection and tracking in complex scenes, and especially when there is more than one person in the picture, still remains a challenge. Several good overviews of the problem can be found in the references Chellappa et al. (1995), Samal and Iyengar (1992), and Bala et al. (1997). Other studies (Kampman and Osterman, 1997; Wollborn et al., 1997; Mech and Gerken, 1997; Aizawa and Huang, 1995; Aizawa et al., 1989) have attempted to perform an automatic adaptation of a face model to carry out model or content-based coding which aims to achieve transmission of the face and facial

CCC 0899-9457/99/010086-10

features at a higher quality than the rest of the picture. A recent model-based coding approach presented by Aizawa et al. (1989, 1995) proposed using a three-dimensional (3D) model that is fitted to the data field for each input video frame analysed in the encoder. Because of its high complexity, this method of automatic fitting of models, such as the wireframe models, is still far from being considered useful for real-time implementation. Approaches of this kind also suffer from a lack of flexibility, as they make assumptions about the content of an image. Another interesting approach (Eleftheriadis and Jacquin, 1995) has been to only partially model the data, i.e., model the location of specific objects known a priori to be present in the scene, and integrate this partial model to a classical video coding system. This approach is based on 2D modeling obtained from the segmentation of the input image into objects. This proposed model-assisted video coding (Eleftheriadis and Jacquin, 1995) is less complex and keeps full decoder compatibility with MPEG and H.261 coding standards. However, it has been considered a priori in their work that in videoconferencing scenes, one person is shown from the waist up in front of a still background. Therefore, it has not been considered how the location modeling approach is able to handle the analysis of the scene if there is more than one person in the conference. In any case, the available bit rate is insufficient to allow a higher bit rate for all the faces identified in the picture, even if there are no difficulties in locating the faces of all the conference participants.

On the other hand, it is possible to use an external source of information—namely, the audio information—to identify the location from which a sound emanates (i.e., the lips of a speaker among other participants in the conference). The audio-augmented technique presented in this article is based on the idea of estimating the direction of the sound source using an array of microphones. Hence, by combining the audio and video information, it is possible to achieve a significant increase in service quality with a relatively small increase in cost. This goal may be achieved by more accurate quantization giving finer spatial resolution for the area of the lips and face of the speaker, compared to the rest of the scene. The approach described in this article uses an array of microphones to locate the lips of the speaker using cross-correlation measurements and combines audio and video information to enhance the transmission quality. The technique is compatible with all existing video compression standards and has the potential to be implemented more easily than other proposed techniques.

III. ESTIMATION OF LIP POSITION USING A MICROPHONE ARRAY

Microphone arrays have attracted a lot of interest of late, mainly in two closely related areas. Reducing the noise and reverberance in sound pickup that might be present in a single microphone has been one major area and still continues to be an active area of research (Flanagan et al., 1985; Flanagan, 1991; Goodwin and Elko, 1993; Chu, 1995). Work in this area suggests solutions using multiple microphones in an array. The other application has been the location of the speaker in a conference room or auditorium to turn the camera automatically toward the speaker. In the first area, audibility is the major concern as the microphone array should be able to steer electronically toward the speaker to pickup the best signal. In the second application, however, the audibility task may not be the prime concern as a camera is required to turn to cover the speaker in the captured picture (Fischell and Coker, 1984). This is based on acoustic location, again performed on the speech signal available from the speaker, a technique now being used in videoconferencing. The technique presented in this article, however, is quite different, as it aims to achieve more accuracy and higher resolution within a part of the image. Instead of moving a camera toward the speaker, the proposed method identifies the precise location of the speaker's lips in the video frame captured by the camera. The spatial resolution required here is therefore only of the order of a macroblock (16*16 pixels) of a picture.

Estimating the position of a speaker's lips using an array of microphones works on the premise that the sound generated by the speaker arrives at various microphones at different times depending upon the spatial positions of the microphones relative to the sound source. Various direction-of-arrival (DOA) estimation techniques using an array of sensors exploit this differential delay to estimate the direction of the source with respect to the array of sensors. A review of DOA estimation methods, including their performance, sensitivity, and limitations, as well as general beam-forming principles is presented in Godara (1997). Much of the literature on DOA estimation techniques is for narrow-band point sources; extension to broadband sources, such as the case under consideration in this work, can be found in Su and Morf (1983), Wang and Karch (1985), Swingler and Krolik (1989), Ottersten and Kailath (1990), Cadzow (1990), Duron et al. (1993), Buckley and Griffiths (1988), Hung and Kaveh (1990), and Schultheiss and Messer (1993). The implementation discussed in the current article is based on an estimation of the peaks in the cross-correlation function between the signal received at various microphones located in the plane of the videocamera.

Within the general class of source-location methods based on cross-correlation measurements, there are two distinct subclasses (Gardener and Chen, 1992). In one case, a closely spaced array of sensors is used on a single platform such that the sensor spacing is typically less than half a wavelength for all the received signals. Phase alignment methods for beam steering and null steering are normally employed in this case. In the second case, two or three widely spaced sensors are employed with time-difference measurements used to obtain the location information. The former methods exploit sensor to sensor phase difference for relatively narrow-band signals whereas the latter techniques exploit relative time differences between sensors for wide-band signals to estimate the source location.

Because of the nature of the signal being used in this application, i.e., broadband speech signals, the scheme uses the time difference of arrival (TDOA) estimate between the source, i.e., the lips of a speaker, and a spaced array of microphones forming a planar array. The directivity of the microphone array is used to process the sound emitted by the desired source while at the same time suppressing noise and reverberation arriving from other directions. Using the measurements from the estimates of TDOA, the position of the sound source is located by considering different points on a grid as shown in Figure 1. Starting at the outer boundary of the area under consideration, the delay for each point is measured using the optimal delay estimation method.

A. Description of TDOA Method. Let the coordinates of two microphones, i and j be denoted as (X_i, Y_i, Z_i) and (X_j, Y_j, Z_j), respectively, and let $R_{i,j}$ denote the range difference between the source and the two microphones. Assuming $(x, y, z,)$ to be coordinates of the sound source, the range difference is given by

$$R_{i,j} = \sqrt{(X_i - x)^2 + (Y_i - y)^2 + (Z_i - z)^2} - \sqrt{(X_j - x)^2 + (Y_j - y)^2 + (Z_j - z)^2} \quad (1)$$

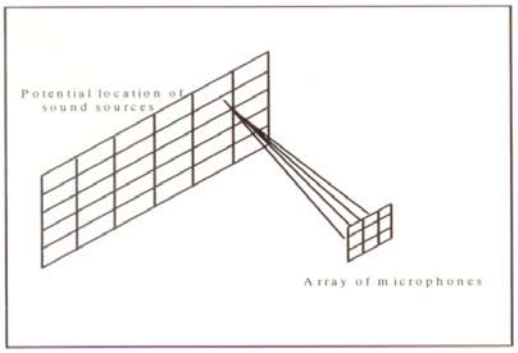

Figure 1. Array of microphones used to cover the area of interest.

which is an equation of a hyperboloid with foci at the microphones. Figure 2 shows plots of two hyperboloids depicting constant range differences $(R_2 - R_1)$ and $(R_3 - R_2)$ between two sets of microphones. One observes from the figure that the sound source lies at the intersection of two hyperboloids, and thus to locate the source on a two-dimensional (2D) surface one requires three microphones. Similarly, four microphones are required to locate a source uniquely within a 3D region.

Let v be the velocity of sound, assumed to be constant, inside the conference room where the system is placed and T_i and T_j represent the time taken by the signal to travel from the source to microphones i and j, respectively; then the range difference is related to the time difference by

$$R_{i,j} = v*(T_i - T_j) \tag{2}$$

and thus the time difference may be used in place of the range difference. As discussed above, at least four independent measurements of these differential delays need to be made to locate a 3D position of a sound source.

In the experiment conducted for the study reported in this article, the sound source (a human speaker) was placed in front of a planar array. As the distance between the plane of the array and the source was known from the experimental design, the two hyperboloids were formed from TDOA measurements of three fixed microphones to provide an intersecting point to locate the position of the sound source, and thus only three microphones were used. The experiment was implemented in two stages. First, estimates of TDOA values were obtained. This was followed by the processing of these values to determine a location estimate. The process is briefly described below.

1. TDOA Estimates. The method used in this work for obtaining these estimates is the generalized correlation method. Assume that the signals $x_1(t)$ and $x_2(t)$ are received by two spatially separated microphones owing to a speech signal $s(t)$ emanating from a distant source, namely, the lips of a speaker in a videoconference room. In the presence of the white noise, $n(t)$, which is uncorrelated with the speech signal, the microphone signals can be modeled as low-pass stationary processes and can be expressed as:

$$x_1(t) = s_1(t) + n_1(t) \tag{3}$$

$$x_2(t) = \alpha s_1(t - D) + n_2(t) \tag{4}$$

where a real scalar α denotes an amplitude scaling factor and D denotes the differential delay between the arrival of the signal at each of the microphones. A common method of determining the time delay D is to compute the cross-correlation function and estimate the argument which maximizes this function. The cross-correlation $\mathcal{R}$ between $x_1(t)$ and $x_2(t)$ is defined as

$$\mathcal{R}_{x1x2}(\tau) = E[x_1(t)x_2(t - \tau)], \tag{5}$$

where $E[.]$ denotes the expectation operator. The argument τ that maximizes Equation (5) provides an estimate of the time delay. In practice, $\mathcal{R}(\tau)$ is not available and is estimated from $x_1(t)$ and $x_2(t)$ recorded over a finite time. To improve the accuracy of the cross-correlation function, and hence the estimate of delay, $\hat{D}$, it is necessary to pass $x_1(t)$ and $x_2(t)$ through a filter prior to computing the cross-correlation function to accentuate the signal passed to the correlator at frequencies for which the signal-to-noise ratio (SNR) is highest, and to simultaneously suppress the noise power. The schematic diagram for this arrangement is shown in Figure 3.

An estimated cross-spectral density function can also be computed in the frequency domain, and then the estimated cross-correlation function is obtained via an inverse Fourier transform. Using the latter approach, therefore, the role of the two filters providing y_1 and y_2 is as a weighting factor for the cross-spectral density of the two input signals. This approach has been adopted, as the frequency domain processing lends itself well to filtering of the signal prior to the computation of the cross-correlation function.

The cross-correlation function $\mathcal{R}$ and the cross-power spectral density G between the two signals are related by

$$\mathcal{R}x_1x_2(\tau) = \int_{-\infty}^{+\infty} Gx_1x_2(f)e^{j2\pi f\tau}\,df \tag{6}$$

When $x_1(t)$ and $x_2(t)$ are filtered, then the cross-correlation between them is

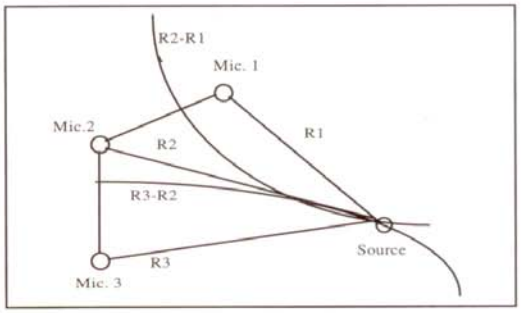

Figure 2. Hyperbolic position location with three microphones for a 2D solution.

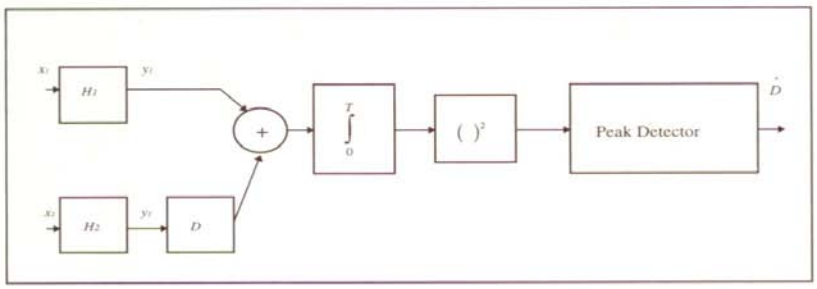

Figure 3. Recorded signal filtered, delayed multiplied, and integrated for a range of delays until a peak in output.

$$\mathcal{R}y_1y_2(\tau) = \int_{-\infty}^{+\infty} \psi(f) G x_1 x_2(f) e^{j2\pi f\tau}\, df \tag{7}$$

where $\psi(f) = H_1(f) H_2^*(f)$, and denotes the frequency weighting.

Again, owing to the finite length of the recorded signals, only an estimate of cross-power spectral density function of the two signals, $\hat{G}x_1x_2(f)$, can be obtained. Consequently, one needs to evaluate the integral

$$\mathcal{R}y_1y_2(\tau) = \int_{-\infty}^{+\infty} \psi(f) \hat{G} x_1 x_2(f) e^{j2\pi f\tau}\, df \tag{8}$$

for estimating delay.

Among the several delay estimation processors (Knapp and Carter, 1976), the phase transform (PHAT) has been selected for the study presented in this article, as it leads to a simple processor implementation and is also less costly in terms of the processing time required to perform the calculations (Carter et al., 1972). This estimator uses the weighting such that

$$\psi(f) = H_1(f) \cdot H_2(f) = \frac{1}{|Gx_1x_2(f)|} \tag{9}$$

and therefore

$$\mathcal{R}y_1y_2(\tau) = \int_{-\infty}^{+\infty} \frac{\hat{G}x_1x_2(f)}{|Gx_1x_2(f)|} e^{j2\pi f\tau}\, df \tag{10}$$

For uncorrelated noise, the cross-power spectral density between the two noise components $Gn_1n_2(f) = 0$. Thus, it follows that

$$|Gx_1x_2(f)| = \alpha G s_1 s_2(f) \tag{11}$$

Ideally, when our estimate of the cross-power spectral density is a close approximation to the actual spectral density, i.e., when $\hat{G}x_1x_2(f) = Gx_1x_2(f)$,

$$\frac{\hat{G}x_1x_2(f)}{|Gx_1x_2(f)|} = e^{j\theta(f)} = e^{j2\pi fD} \tag{12}$$

which has unit magnitude and therefore

$$\mathcal{R}y_1y_2(\tau) = \delta(t - D) \tag{13}$$

This indicates that under ideal conditions, the correlation function is an impulse at the differential delay D and is therefore easily determined.

2. *Estimation of Lip Position.* Once reliable estimates of TDOA have been made, these may be substituted for the corresponding range difference estimates, into the hyperbolic equations to solve for the position of the speaker's lips.

Three microphones are used for the process. Figure 4 shows the position of the sound source with respect to the microphones which are arranged in a plane. Let these three microphones be referred to as a, b, and c, respectively. Microphone a is placed at the origin of the coordinate system. Microphone b is placed on the y axis and microphone c is placed on the x axis with their coordinates being (0, b, 0) and (c, 0, 0), respectively. Denoting $T_{ab} = T_a - T_b$ and $T_{ac} = T_a - T_c$ as the TDOA estimates between the two pairs of microphones of interest, namely a–b and a–c, one obtains (Fang, 1990)

$$\sqrt{x^2 + y^2 + z^2} - \sqrt{(x - b)^2 + y^2 + z^2} = V * T_{ab} = R_{ab} \tag{14}$$

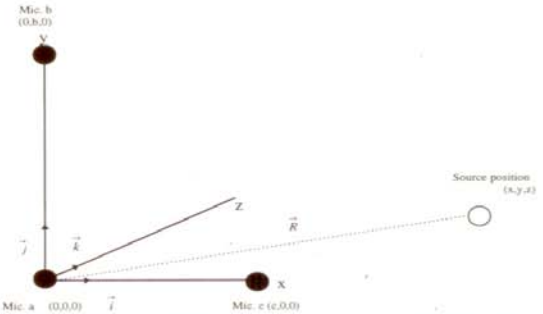

Figure 4. Source position referenced to local coordinates defined by microphone plane.

Figure 8. (a–c) Original images in which the position of the lips (+) has been estimated from the acoustical signal.

tions, the two sound sources should be located with the active source (active speaker) position tracked while the interfering sound source is suppressed. If the interference results in a spurious peak and a false source location, its effect on the video will not be perceptible if it lasts for only a short time, as the algorithm will be quickly updated by the real source location. The case of two persistent sources can be treated as a conversation between two participants in the same conference room, and therefore it is entirely up to the video coding controller to apply a suitable coding approach for such a scenario. For example, if speakers are talking among themselves, it is highly likely that the quality at this site is less of an issue and so it could be determined that no specific areas of the image should be improved in quality. Alternatively, it might be possible to locate and improve both areas. Figure 9 indicates how interference caused by the speech of a second speaker at the left of the active speaker has resulted in another delay estimate pointing to the second position in the scene approximately 45 samples to the left. In our implementation the search algorithm is constrained to a limited central area to eliminate reflections and sound sources other than those within the field of view. Improved algorithms based on beam/null steering and

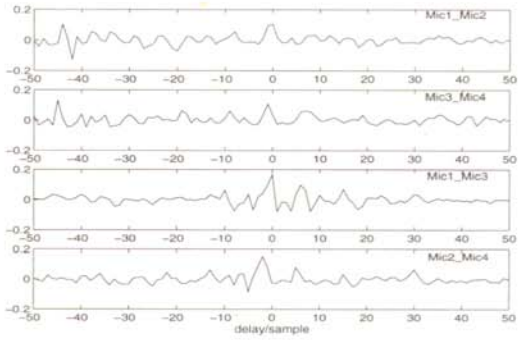

Figure 9. An interference recognized from left.

Table I. Statistical results from coding of sequences.

Format	No. of Frames	Quantizer Background	Quantizer Face	Bits per Frame (Excluding Frame 1)	Bits Required for Frame 1	PSNR (d
CIF	100	10	10	1577	31,677	35.5
CIF	100	14	5	1588	29,534	37.7
QCIF	100	8	8	652	2325	36.2
QCIF	100	9	4	660	2288	37.5

array processing are planned as future work to address this important issue with an even greater ability to discriminate against out-of-scene sound sources.

Resolution is another important issue. The need for high resolution arises primarily when relatively small movements of the head gives rise to multiple nonseparable signals. Differences in arrivals from a very small movements in the source that are separated by less than the width of their cross-correlation peaks usually cannot be resolved by this method. In fact, the short length of the audio samples implies that

$$\hat{G}x_1x_2(f) \neq Gx_1x_2 \tag{27}$$

and hence

$$\mathcal{R}(\tau) \neq \delta(t - D) \tag{28}$$

Increasing both the distance between the microphones and the sampling rate of the audio signals can result in better resolution. Practical limitations in this kind of application, however, dictate the overall performance from a resolution point of view. With the given setup, as stated earlier, a resolution of ± 3.3 cm/sample is obtainable. This by itself provides a sufficiently high resolution within the physical dimensions of the conference scene. When this resolution is projected into the video picture, then the resolution is determined solely by the dimensions of the scene captured by the video camera and the zoom factor. For example, if a scene of 1.5×2 m is captured in the CIF format image, consisting of 352*288 pixels, the ± 3.3-cm resolution results in a ± 5.5-pixel resolution. Since the reduced quantizer step size is applied to a basic rectangular template of 64 vertical pixels by 48 horizontal pixels, the above level of resolution is always sufficient to ensure that the main facial features have been enhanced.

The effect of the depth of the room, i.e., the parameter z in Equations (21) and (25), is highly dependent on the time-delay magnitude. For instance, if the speaker is located in the middle of the picture, the TDOA estimator yields essentially small or zero time-delay components. In this situation, a variation in the depth has no significant effect on the overall performance of the algorithm in locating the precise position of the lips of the speaker. While the sensitivity of the expression in Equation (21) to the z parameter can be mathematically traced, an experimental approach was taken to examine the sensitivity of the algorithm to changing the depth of the room to 3.3 m $\pm$ 10%. This represents, in effect, a movement of the speaker ± 30 cm forward and backward along the camera axis. The resultant x, y values varied by only ± 1 cm about their original estimates. This is far less than the minimum detectable resolution in terms of the image properties. However, if the location of the speaker is significantly displaced from the center, giving rise to large time-delay magnitudes, then the effect of the variation in depth i more pronounced.

The length of the data record also needs to be considered in view of the fact that for real-time implementation the length of the audio segment needs to be kept as short as possible. On the other hand, to achieve the desired signal-to-noise ratio and to accommodate pauses within the speech signal, the length of the segment has to be sufficiently long. This not only results in higher processing time than is desirable for real-time implementation, but also may result in the acquisition of audio data generated during rapid motion of the speaker which will introduce the issue of moving sources. It will imply, therefore, that the recorded speech signal is essentially from two or more separate sources (owing to the significant movement of the speaker). A longer data record also slows the frame updating rate and this is undesirable as the result of position estimation needs to be applied to the video frames from which the audio information had been obtained. With a sampling rate of audio signals used in this study—namely, 32 kilosamples/s—an audio segment length of 4000–5000 samples has been found to result in reliable position estimates.

B. Video Quality. The sequences coded using the two-stage quantizer provide significantly better quality in regions of the face where the correspondence between the speech and movement of the lips is subjectively important. The static background in the latter case has been coded more coarsely (QSS = 14 compared to QSS = 10 in the default coding scheme) without significant deterioration in the way in which the background has been reproduced.

The other aspect of the coding process which affects the subjective quality of the received video data is the forced updating rate. In both the H.261 and H.263 codecs, this function is achieved by forcing the use of INTRA mode in the coding algorithm. While the update pattern is not defined, the standards require that every macroblock be updated in INTRA mode at least once in every 132 times it is transmitted. Using the proposed technique, it may well be possible to use audio information to force a shorter update rate for those macroblocks coded at a smaller quantization step size, while to compensate for the increased data rate which results a larger update rate could be used when there is no speech within the scene.

Table I provides some statistical results from coding the 100-frame sequence both in QCIF and CIF formats. While the overall bit rate for each sequence remains the same, there is a 2.2-dB improvement in the peak signal-to-noise ratio (PSNR) of the CIF sequence calculated for the face area, i.e., measured on 12 macroblocks coded differently from the rest of the image. Table I indicates a 1.3-dB increase in PSNR when the QCIF sequence was coded using a QSS = 4 for the four macroblocks on the lip area. Two reconstructed frames in CIF format are shown in Figure 10(a,b). Although the full impact of this scheme can only be fully appreciated when the entire sequence is played back on a good-quality monitor, the two

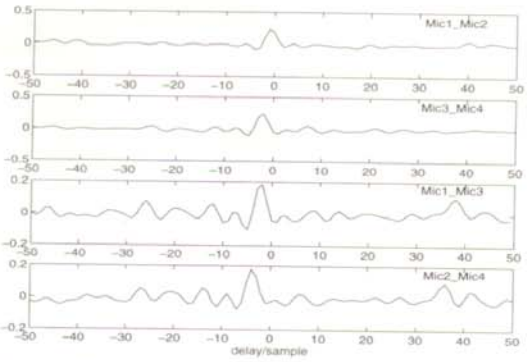

Figure 6. TDOA estimation when speaker is located at the center.

Processing on a pair of audio segments from two channels results in a delay estimation, calculated in terms of a number of samples, between the two microphones. With a common microphone as the origin of the array coordinate system (0, 0, 0), delay estimates from each of two sets of signals provides, therefore, an indication of the delay between each microphone and the common one. At a sampling rate of 32 kilosamples/s, a time delay equivalent to one sample then represents a range difference of approximately 1 cm between the two microphones, which in turn represents a spatial difference of approximately 6.6 cm (i.e., ±3.3-cm resolution) across the scene. Given the typical dimensions of a conference room, a ±3.3-cm spatial resolution is sufficient for successfully finding the location of the face of the speaker. It has been assumed that the active speaker is not walking around in the scene, as this is not usual in videoconferencing applications.

V. RESULTS

A. Speaker Localization. In the audio-processing section, performing the phase transform PHAT has resulted in reliable TDOA estimation result. As an example, Figure 6 shows the delay estimate when the speaker is seated almost at the center of the scene. The range in which a delay is expected has been set to ±50 samples, which covers a range of ±1.65 m from the center of the scene, thus eliminating any undesired delay estimations due to the reflections from the side walls. Figure 7 illustrates the histogram of distribution of error in estimating the correct TDOA. To obtain this histogram, the algorithm performed 100 TDOA estimation operations on audio data taken from the first 20 s of the recorded video sequence, during which time the speaker had no significant movement. It was anticipated, therefore, that the algorithm would consistently yield the same TDOA estimates and that the position location algorithm would point to the same location. According to this histogram, 75 estimations (i.e., 75%) resulted in no error in yielding the TDOA, and the estimation was indeed the expected value. In 25 estimations, however, TDOA measurements resulted in an error in estimation of up to ±2 sample delays, which in terms of spatial resolution represents a maximum of ±6.6 cm in the plane of the speaker. This histogram clearly indicates that the algorithm is well capable of performing its task within the defined frame of room dimensions and required resolution. Figure 8(a–c) represents the result when the position estimation was transformed into the picture attributes to identify the position of the lips. A small cross has been superimposed on each original image at a position estimated to be the location of the source of sound. For each of the three quite different scenarios, the technique has correctly identified the lip position to within an acceptable margin of error. However, on occasions, reflections and other multipath arrivals result in inconsistent estimations—for example, when the speaker faces away from the camera as in Figure 8(c). Signal power is examined by the proposed technique to discriminate speech signals from background noise prior to carrying out the estimation process. This is a key measure for both identifying the real speech signal and hence allowing the audio samples to be processed for delay estimation.

B. Video Coding. Two sequences have been coded using an H.261 coder. The first was in QCIF format of the sequence recorded in the conference room. With the four macroblocks surrounding the estimated position of the lips coded at a quantizer step size (QSS) of 4, the rest of the picture was coded at QSS = 9. This produced the same total bit rate as when the entire sequence was coded at QSS = 8. The second set was in CIF format of the same sequence, with a section of 12 macroblocks covering the estimated position of the lips and face. A rectangular template with an aspect ratio of 4×3 was used to fit on the macroblock marked as the location of the lips. With a default height of 4 macroblocks and a width of 3, this template can enlarge in accordance with the zoom factor. The face area was coded at QSS = 5 and the remaining macroblocks coded using QSS = 14. This produced the same total bit rate as when the entire sequence was coded at QSS = 10. A comparison between the reconstructions obtained with these quantiser step sizes may be carried out by inspecting Figure 10(a,b).

VI. DISCUSSION AND PROPOSALS FOR FURTHER WORK

A. Audio DSP. One of the main issues in exploiting audio signals for speaker location is the interference inside the scene when, for example, sources other than the speaker are active. In such situa-

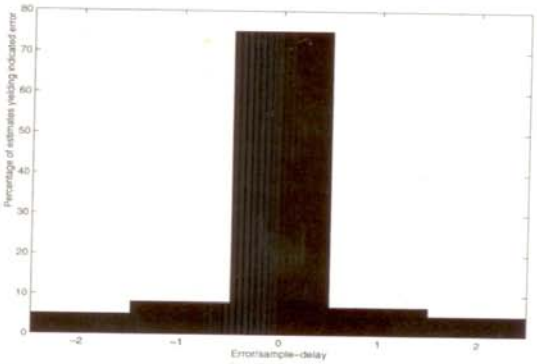

Figure 7. Histogram of error distribution in performing the location estimation.

B. Dodd and R. Campbell, Hearing by eye: The psychology of lip-reading, Erlbaum, London, 1987.

M.A. Doron, A.J. Weiss, and H. Messer, Maximum-likelihood direction finding of wide-band sources, IEEE Trans Signal Process 41 (1993), 411–414.

A. Eleftheriadis and A. Jacquin, Automatic face location detection and tracking for model-assisted coding of video teleconferencing sequences at low bit-rates, Signal Process Image Commun 7 (1995), 231–248.

B.T. Fang, Simple solutions for hyperbolic and related position fixes, IEEE Trans Aerospace Electronic Systems 26 (1990).

D. Fischell and C.H. Coker, A speech direction finder, ICASSP-84, pp. 19.8.1–19.8.4.

J.L. Flanagan, Autodirective microphone system, Acoustica 73 (1991), 58–71.

J.L. Flanagan, J.D. Johnston, R. Zahn, and G.W. Elko, Computer-steered microphones arrays for sound transduction in large rooms, J Acoust Soc Am 78 (1985), 1508–1518.

W.A. Gardener and C. Chen, Signal-selective time-difference-of-arrival estimation for passive location of man-made signal sources in highly corruptive environment: Part I: Theory and method, IEEE Trans Signal Process 40 (1992).

L.C. Godara, Application of antenna arrays to mobile communications: Part II: Beam-forming and DOA considerations, Proc IEEE 85 (1997), 1195–1247.

M.M. Goodwin and G.W. Elko, Constant beamwidth beam-forming, IC-ASSP-93, pp. I-169–I-172.

H. Hung and M. Kaveh, Coherent wide-band ESPRIT method for direction-of-arrival estimation of multiple wide-band sources, IEEE Trans Acoust Speech Signal Process ASSP-38 (1990), 354–356.

M. Kampmann and J. Ostermann, Automatic adaptation of a face model in a layered coder with an object-based analysis-synthesis layer and a knowledge-based layer, Signal Process Image Commun 9 (1997), 201–220.

M. Kampmann and J. Ostermann, Automatic adaptation of a face model in a layered coder with an object-based analysis-synthesis layer and a knowledge-based layer, Signal Process Image Commun 9 (1997), 201–220.

C.H. Knapp and G.C. Carter, The generalized correlation method for estimation of time delay, IEEE Trans Acoust Speech Signal Process ASSP-24 (1976).

R. Mech and P. Gerken, Automatic segmentation of moving objects, Doc ISO/IEC JTC1/SC29/WG11 MPEG97/1949, Bristol, UK, April 1997.

B. Ottersten and T. Kailath, Direction-of-arrival estimation for wide-band signals using the ESPRIT algorithm, IEEE Trans Acoust Speech Signal Process ASSP-38 (1990), 317–327.

A. Samal and P.A. Iyengar, Automatic recognition and analysis of human faces and facial expressions: A survey, Pattern Recog 25 (1992), 65–77.

P.M. Schultheiss and H. Messer, Optimal and sub-optimal broad-band source location estimation, IEEE Trans Signal Process 41 (1993), 2752–2763.

G. Su and M. Morf, The signal subspace approach for multiple wide-band emitter location, IEEE Trans Acoust Speech Signal Process ASSP-31 (1983), 1502–1522.

D.N. Swingler and J. Krolik, Source location bias in the coherently focussed high-resolution broad-band beam former, IEEE Trans Acoust Speech Signal Process ASSP-37 (1989), 143–145.

H. Wang and M. Kaveh, Coherent signal-subspace processing for the detection and estimation of the angle of arrival of multiple wide-band sources, IEEE Trans Acoust Speech Signal Process ASSP-33 (1985), 823–831.

M. Wollborn, M. Kampmann, and R. Mech, Content-based coding of videophone sequences using automatic face detection, Picture Coding Symp 1997, Berlin, Germany, pp. 547–551.