Audio Head Pose Estimation using the Direct to Reverberant Speech Ratio

Mark Barnard*, Wenwu Wang

Centre for Vision, Speech and Signal Processing, University of Surrey, GU2 7XH, UK.

Abstract

Head pose is an important cue in many applications such as, speech recognition and face recognition. Most approaches to head pose estimation to date have focussed on the use of visual information of a subject's head. These visual approaches have a number of limitations such as, an inability to cope with occlusions, changes in the appearance of the head, and low resolution images. We present here a novel method for determining coarse head pose orientation purely from audio information, exploiting the direct to reverberant speech energy ratio (DRR) within a reverberant room environment. Our hypothesis is that a speaker facing towards a microphone will have a higher DRR and a speaker facing away from the microphone will have a lower DRR. This method has the advantage of actually exploiting the reverberations within a room rather than trying to suppress them. This also has the practical advantage that most enclosed living spaces, such as meeting rooms or offices are highly reverberant environments. In order to test this hypothesis we also present a new data set featuring 56 subjects recorded in three different rooms, with different acoustic properties, adopting 8 different head poses in 4 different room positions captured with a 16 element microphone array. As far as the authors are aware this data set is unique and will make a significant contribution to further work in the area of audio head pose estimation. Using this data set we demonstrate that our proposed method of using the DRR for audio head pose estimation provides a significant improvement over previous methods.

Keywords: Audio Head Pose, Direct to Reverberant Speech Ratio

1. Introduction

The head pose or orientation of a speaker's head is an important cue for many applications. These include providing a constraint for face detection/recognition, improving speech recognition performance [1], as a cue for modelling interactions among multiple speakers [2]. Head pose is also an important cue for the delivery of 3D audio [3], an important and emerging area in audio processing. Most current work on head pose estimation has focused on

*Corresponding author

Email addresses: mark.barnard@surrey.ac.uk (Mark Barnard), w.wang@surrey.ac.uk (Wenwu Wang)

the use of visual information [4]. Indeed when used in ideal conditions visual head pose estimation method can produce very accurate results for all three degrees of freedom of the human head, *roll, pitch* and *yaw*.

The visual modality however has a number of inherent limitations. These include camera distortion, low camera resolution, multiple and variable sources of light, occlusions of subjects due to camera angle and field of view. There are also inherent differences in the visual appearance of subjects, with a great deal of variation in facial appearance and also external factors such as people wearing hats or eye glasses. However, the audio signal produced by a speaker is unaffected by these limitations and so would seem to have a role in head pose estimation either independently or complementing a visual head pose estimation system.

There are two important properties of human speech that are very pertinent to the use of audio information for head pose estimation. One property is that sound produced by the mouth has a directionality caused by the shape of the human head [5, 6]. This property means that the energy of the speech signal will be greater when the subject is directly facing the microphone than when the subject is facing away from the microphone. This effect was measured in more detail using multiple microphone arrays by Meuse and Silverman [7]. Another property of human speech is that the directivity of the signal is also dependent on its frequency [8]. One of the first comprehensive studies of this phenomenon was conducted by Chu and Warnock [9] which clearly shows the more directional nature of higher frequencies.

Audio information has been used in our previous work to aid in systems that only track the subject's head without estimating head pose [10, 11]. These approaches show that audio information is complementary to the video and can be used to overcome visual occlusions when tracking multiple people. More specifically, the direction of arrival (DOA), or azimuth angle, of a speaker, derived from the audio signal, has been proposed to cope with occlusions. The introduction of audio for combined audio-visual tracking has provided robustness to occlusions in a number of applications [12, 13, 14, 15]. Wang and Brandstein [16] propose an audio-visual tracking system using four element microphone arrays and a single steerable camera. Many audio-visual tracking methods [2, 17, 18] combine audio and visual location likelihoods within the framework of particle filters (PF) [19], a powerful stochastically based tracking method.

Fallon *et al.* [20] combine audio and visual information to jointly estimate location and head orientation using the directionality of the energy of the speech around the human head. Brutti and Lanz [21] use seven microphone arrays distributed around the walls of a room to characterize the distribution of the audio signal and to estimate the likelihood of a particular head pose, this is then combined with the likelihood produced from a visual head pose detector. Segura *et al.* [22, 23] present a method of combining audio and visual information, in this case the audio pose estimate comes from exploiting the high to low band ratio (HLBR). This measure exploits the difference in the directivity of low and high frequencies, as previously discussed.

There have also been a number of recent approaches to estimate head pose using only audio information. Matching the head pose to the distinctive radiation pattern around the head has been used for audio head pose estimation [24, 25]. These approaches require a large-aperture microphone array, more than 400 microphones are used in each case, covering the

walls of the room. Brutti *et al.* [26, 27] proposed using a modification of the global coherence field (GCF) and the orientated global coherence field (OGCF) for head pose estimation using a network of seven distributed microphone arrays. This is further extended to the multipath OGCF [28] using a single 64 element linear microphone array. Abad *et al.* [29] also exploits the difference in directivity at different frequencies, described in the previous paragraph, the HLBR, as a method of head pose estimation using only audio information. This method relies on four microphone arrays, each containing four elements, distributed around the room. Mungamuru and Aarabi [30] also use the directivity of the audio signal around the head, but exploit the change in relative attenuation of the signal rather than the directivity at different frequencies. While this algorithm is principally used for source localisation, using speaker and microphone directivity it can also give an estimation of speaker head orientation. The methods described depend on accurately measuring the energy pattern around the head. This becomes more challenging as the acoustic environment becomes more reverberant due to increased noise. Whereas other methods suffer in reverberant environments, our proposed method actually takes advantage of a reverberant environment to estimate head pose.

As opposed to existing pose estimation methods whose performance is prone to reverberation degradation, such as [25, 23], we propose instead to exploit the highly reverberant nature of many room environments, such as meeting rooms or offices. Our hypothesis is that the direct speech energy received at a far-field microphone will be greater if the speaker is facing the microphone rather than facing away, while the reverberant energy remains relatively constant. Hence the direct to reverberant speech ratio (DRR) should provide a useful measure of head pose/orientation.

In an enclosed space listeners receive speech both directly and via reverberation. Direct speech is that part of the speech that would be received in an open space without any reverberation. The level of direct speech energy is dependent on the initial sound energy, the distance to the listener or microphone and the orientation of the source [31]. Reverberation is the persistence of sound in a room due to multiple and repeated reflections from the surfaces forming the enclosed space. The level of the reverberant speech signal depends on the initial energy, the size of the room, and the nature of the surfaces in the room. The DRR is given by

$$DRR = \frac{E_d}{E_r}, \tag{1}$$

where E_d is the energy of the direct speech signal and E_r is the energy of the reverberant speech signal.

This ratio is known to be used by humans for estimating the distance to a sound source in reverberant environments [31, 32]. Recent research [33] has shown that the DRR can also be used in machine audition to estimate the distance to a sound source. In this case an equalisation-cancellation operation on a binaural reverberant signal is used to estimate the DRR. In a previous work [34] we demonstrated simple head pose detection using two sequences from the AV16.3 dataset [35] using the DRR, however this data featured only two subjects and a limited range of head poses. Our purpose in this paper is to demonstrate that the DRR measure can be used to robustly estimate head pose using only audio data. In order to achieve this we collected a comprehensive new audio head pose dataset, featuring

39 subjects, which is described in Section 4.1 and publicly available [1]. We believe that the complementary nature of audio signal, for example, it is unaffected by visual occlusions or low resolution images, will provide either a useful stand alone head pose estimation technique or fused with visual information to provide more robust head pose estimation.

The remainder of this paper is organised as follows: in the next section we outline our proposed method of audio head pose estimation along with the high to low band ratio (HLBR) method of audio head pose estimation. In section 3 we describe the process of audio head pose estimation. In Section 4 we introduce a new database for audio head pose estimation and present experiments on this new database. Finally in Section 5 we present our conclusions.

2. Pose Dependent Features

In this section we first give details of the energy and HLBR features for audio head pose estimation introduced by [22]. We then present our proposed method of audio head pose estimation exploiting the DRR produced in a reverberant room environment.

2.1. Energy and HLBR Features

Given an array of K microphones audio energy for a window of N samples at the k^{th} microphone, x_e^k, is calculated by:

$$x_e^k = \frac{\sum_{n=0}^{N} w_k^2(n)}{N}, \tag{2}$$

where w is the amplitude of the k^{th} microphone signal.

As discussed in Section 1 this phenomenon has been exploited previously by researchers for estimating the head pose. In this paper, we define the high frequency audio band to be between 3500 and 4500 Hz and the low audio band as 200 to 400 Hz, as given by Abad *et al.* [29].

So for the k^{th} microphone in the array we measure the high to low band ratio (HLBR), x_r^k according to

$$x_r^k = \frac{e_k^L}{e_k^H} \tag{3}$$

where e_k^L and e_k^H are the low band and high band energy respectively calculated according to Equation (2). This is illustrated in Figure 1.

2.2. DRR Feature

In order to estimate the DRR we use the method proposed by Jeub et el. [36] to estimate the coherent to diffuse energy ratio using a microphone pair $\vec{k} = \{m_k, m_{k+1}\}$. A noise field can be defined as a mixture of coherent or diffuse noise fields. If we have two audio signals w_1

[1] http://www.cvssp.org/avbss/dataset/

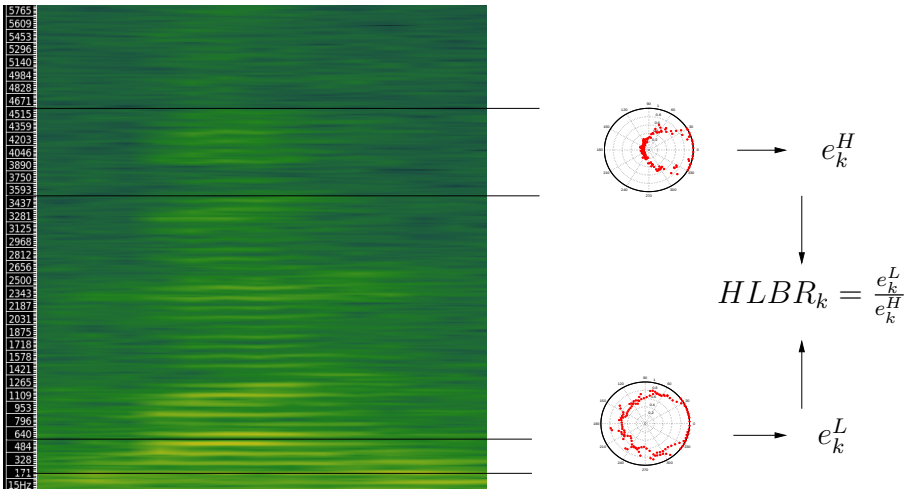


Figure 1: An example to illustrate how the HLBR is calculated from the low and high frequency bands of the audio signal. The left plot is the spectrogram of the word 'one' uttered by a subject, where the high and low frequency bands used for the estimation of the HLBR are indicated with two pairs of parallel solid lines. The polar plots corresponding to the high and low frequency bands are also shown. More details about such polar plots can be found in Figure 5. It can be seen that the energy information in higher frequencies is more directional as compared with that in the lower frequencies.

and w_2, received at microphones m_k and m_{k+1} respectively, the complex noise field coherence in the frequency domain is defined as [36, 8]:

$$\Gamma_{w_1w_2}(\Omega) = \frac{\Phi_{w_1w_2}(\Omega)}{\sqrt{\Phi_{w_1w_1}(\Omega) \cdot \Phi_{w_2w_2}(\Omega)}}, \tag{4}$$

where $\Phi_{w_1w_1}(\Omega)$ and $\Phi_{w_2w_2}(\Omega)$ are the auto-power spectral density (PSD) of the microphone signals $w_1(t)$ and $w_2(t)$ and $\Phi_{w_1w_2}(\Omega)$ is the cross-PSD. The normalised radian frequency is given by $\Omega = 2\pi f/f_s$, where f_s is the sampling frequency, in our case 16 kHz, and f is the frequency of the signals in Hz.

A mixed noise field can be considered as a superposition of the diffuse and coherent noise fields of all the sound sources. Assuming all noise sources are uncorrelated the general complex noise coherence function is

$$\Gamma^{(mix)}_{w_1w_2}(\Omega) = \frac{\sum_{s=1}^{S} \Phi^{(s)}_{w_1w_2}(\Omega)}{\sqrt{\sum_{s=1}^{S} \Phi^{(s)}_{w_1w_1}(\Omega) \cdot \sum_{s=1}^{S} \Phi^{(s)}_{w_2w_2}(\Omega)}}, \tag{5}$$

where S is the number of sound sources and the superscript mix denotes the noise field mixture. In an enclosed reverberant space we can consider the coherent noise field to be the direct speech signal and the diffuse noise field to be the reverberant speech signal. Therefore, the following heuristically motivated equation can be used to calulate the DRR [36]

$$DRR(\Omega) = \frac{|sinc(\Omega f_s d_{mic}/c)|^2 - |\Gamma^{(mix)}_{w_1w_2}(\Omega)|^2}{|\Gamma^{(mix)}_{w_1w_2}(\Omega)|^2 - 1}, \tag{6}$$

where d_{mic} is the distance between the two microphones in metres (in our case 0.1 m) and c is the speed of sound which is set to a constant 340 m/s in all following experiments. The first zero crossing point of the sinc function is given by $f_0 = c/(2d_{mic})$ in our case $f_0 = 200$ Hz. A threshold is imposed on the coherence of $\Gamma_{max} = 0.99$. The following thresholds are used for determining the coherent and diffuse noise for $f > f_0$

- $\Gamma_{w_1 w_2}(\Omega) < 0.1 \rightarrow$ diffuse noise
- $\Gamma_{w_1 w_2}(\Omega) > 0.9 \rightarrow$ coherent noise

This gives us the final DRR measurement for each microphone pair $\vec{k}$

$$x_d^k = \frac{1}{U} \sum_{u=0}^{U} DRR(u), \tag{7}$$

where the discrete frequency bin u relates to Ω via $u = \frac{U\Omega}{2\pi}$. The full details of the implementation can be found in [36].

3. Pose Classification

We treat the problem of audio head pose estimation as a classification problem. In order to do this we train a linear Support Vector Machine (SVM) classifier for each pose in a one against the rest training scheme. We treat all the measurements from the microphone array as a feature vector, so instead of using the results from an individual microphone or microphone pair we treat the measurements from the entire array as a single feature. So for each measurement method we have following feature vectors: $\vec{x}_e = [x_e^1, x_e^2, \dots x_e^{K_e}]^T, \vec{x}_r = [x_r^1, x_r^2, \dots x_r^{K_r}]^T$ and $\vec{x}_d = [x_d^1, x_d^2, \dots x_d^{K_d}]$, where $\vec{x}_e$, $\vec{x}_r$ and $\vec{x}_d$ are the energy, HLBR and DRR feature vectors respectively and K_e, K_r and K_d are the length of the energy, HBLR and DRR feature vectors respectively. The elements of the audio energy, HBLR and DRR feature vectors are produced according to Equations (2), (3) and (7) respectively. As the DRR measure is taken from each microphone pair the length of the feature vector for the DRR measure, K_d is shorter than the length of the feature vectors for the energy, K_e and HLBR, K_r.

Using these feature vectors a set of standard linear SVM classifiers are trained for each head pose using the three different features described in the previous paragraph. So for each feature we have a set of classifiers $\mathbf{C} = \{c_1, c_2 \dots, c_H\}$, where H is the number of classes we wish to classify, in our case head poses. For each feature vector in the test set we produce a set of likelihoods for each set of classifiers $\mathbf{G} = \{g_1, g_2 \dots, g_H\}$. We select the head pose P according to

$$P = \arg\min_{h \in \mathcal{H}}(g_h) \tag{8}$$

where $\mathcal{H} = \{1, 2, \dots H\}$

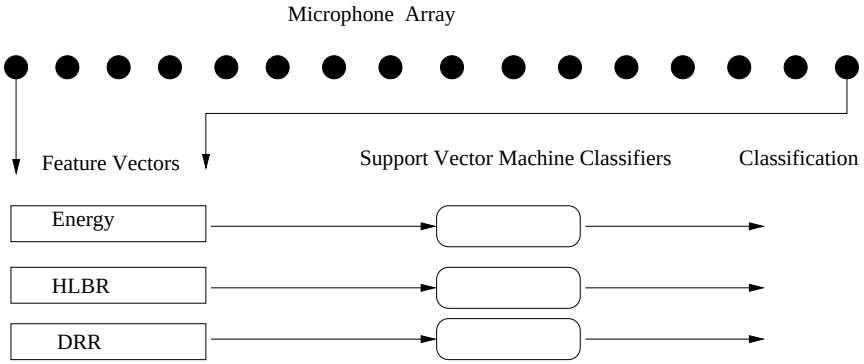


Figure 2: Feature vectors are formed from the measurements from each microphone (or microphone pair in the case of the DRR). These are then used to train SVM classifiers for each pose.

4. Experiments and Results

4.1. Dataset and setup

In order to investigate the subject of audio head pose estimation we have collected a new data set specifically designed for researching the problem of audio head pose estimation. Recordings were made in three different room environments at the University of Surrey. The first room (room 1) was a dedicated recording studio, the configuration of this studio is shown in Figure 4, unfortunately this room is not symmetrical. The dimensions of the room are 3.75 m x 4.35 m x 4.30 m x 4.88 m and has a ceiling height of 3.0 m. This room was used to train the models used to classify the data from all rooms. The second room (room 2) is a large seminar/meeting room containing office furniture, bare walls and carpet on the floor. The room is rectangular with dimensions 8.4 m x 6.6 m and a ceiling height of 3.5 m. A third room (room 3) was selected to give a different acoustic environment to the first two rooms. This room was furnished like a domestic living room, with soft furnishings, bookcases and small tables. This room was again rectangular with dimensions 4.0 m x 5.0 m and ceiling at 3.0 m. Photographs of room 2 and 3 can be seen in Figure 3. The reverberation times (RT60) of the rooms were calculated as the average of the RT60s evaluated for the subbands between 500Hz and 2000Hz of the room impulse responses recorded in these rooms. It was found that they are 300 ms for room 1, 450 ms for room 2 and 250 ms for room 3. In all these rooms, the data was collected using a sixteen-element linear microphone array which was positioned 0.90 m from one wall of the room and 1.50 m above the floor. The microphone array was 1.70 m long with a 0.10 m spacing between each microphone.

A total of 56 subjects were recorded (31 male and 25 female) in total. Training data was collected from 19 subjects (10 male and 9 female) in room 1, none of the these subjects feature in the subsequent testing sets. Using the remaining 37 subjects we then collected three test sets featuring 20 subjects in room 1, 19 subjects in room 2 and 20 subjects in room 3. All recordings were made in English, although the subjects speak a variety of native languages.

Additionally to test the effect of visual occlusions on the performance of our approach we took a subset of 5 subjects from the groups recorded in rooms 2 and 3 and recorded them

(a) Room 2 (b) Room 3

Figure 3: The photos of rooms 2 and 3, where room 3 can be seen with the microphone array in place and the screen used for visual occlusions placed in position P1.

with a plastic sheet measuring 0.8 m x 2.5 m placed in the centre of the array at position P1, seen on Figure 4 and in Figure 3 at a distance of 0.5 metres. The subject then repeated recording procedure described below at postions P2, P3 and P4.

Pose number	Direction from microphone array centre
Pose 1	0 degrees
Pose 2	45 degrees
Pose 3	90 degrees
Pose 4	135 degrees
Pose 5	180 degrees
Pose 6	225 degrees
Pose 7	270 degrees
Pose 8	315 degrees

Table 1: A summary of the pose direction in relation to the microphone array.

Each element of the microphone array consists of a Behringer ECM8000 omni-directional condensing microphone. The audio signal from the microphones was sampled at a frequency of $f_s = 44.1$ kHz. The configuration of the microphone array can be seen in Figure 4.

Each subject commences the recording procedure by standing at position 1 (P1), as shown in Figure 4. During the recording each subject adopts a series of eight head poses in relation to the microphone array, these eight different pose angles are shown in Figure 4. After adopting each head pose the subject pronounces the first ten digits in English, the words 'one' to 'ten'. The subject then turns anti-clockwise through 45 degrees and repeats the procedure, so adopting eight head poses. The subject then moves to position P2 and

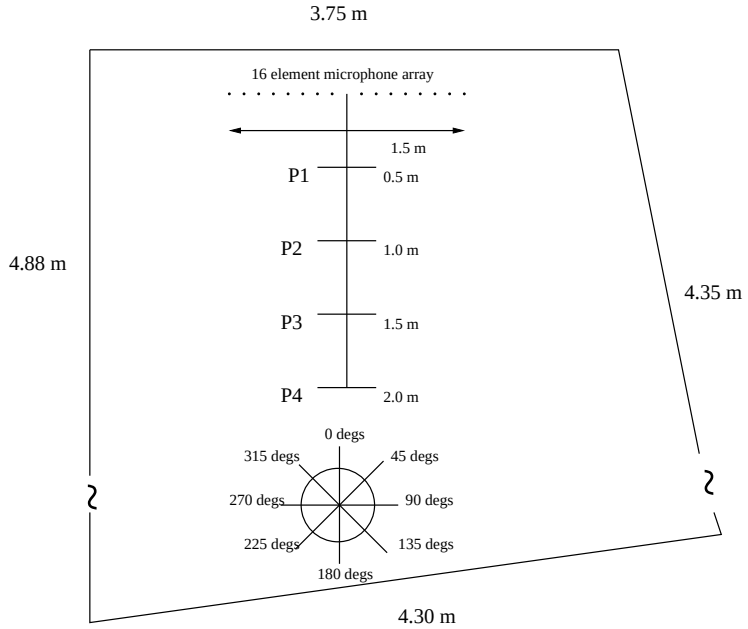


Figure 4: Layout of the room 1 used for recordings the data used to train all classification models. The positions of the speakers denoted by P1, P2, P3 and P4 and at each position recordings for eight different head poses were made for each subject.

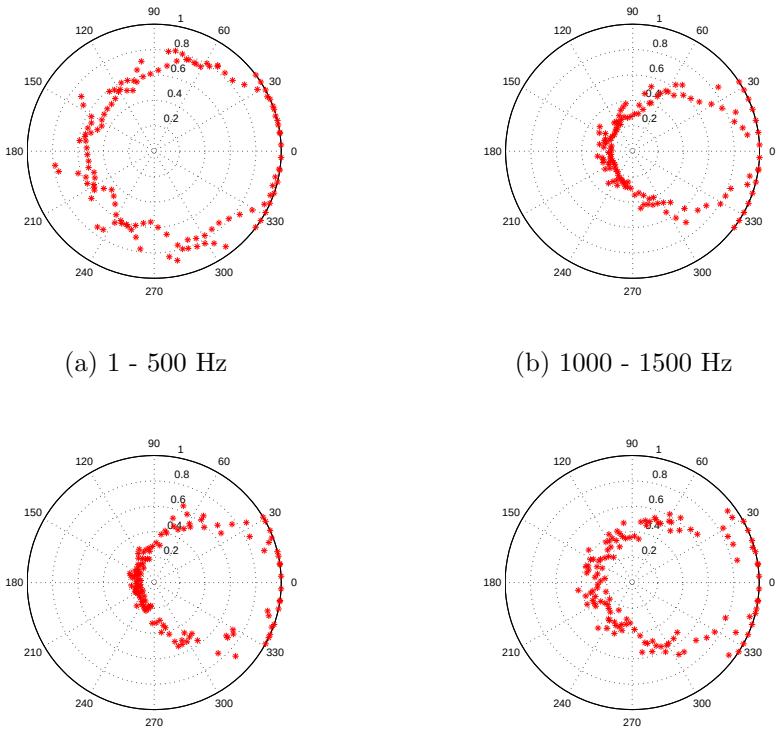
repeats the procedure, and again at positions P3 and P4. Positions P1, P2, P3, and P4 are 0.5, 1.0, 1.5 and 2.0 metres respectively from the centre of the microphone array.

The recorded sequences were then manually annotated in terms of head pose and distance from the microphone array. The length of the longest and shortest sequences were 8.5 s and 2.4 s respectively with a mean length over all sequences being 4.6 s. Unfortunately three subjects missed the final pose at P1. This gives a total of 2493 sequences, 312 examples of each head pose except pose eight which has 309 examples.

To esitmate the DRR, as described in Section 2.2, the input signals w_1 and w_2 are segmented in windows of length $L = 320$ samples with an overlap of 75% between the neighboring windows. These windows are then transformed with FFT of length $M = 512$. For each discrete frequency bin $u = 0, \dots U$, where $U = 256$, the value of $DRR(\Omega)$ is estimated and these are then averaged over the frequency bins. For the energy based measurement we select the band 3500 to 4000 Hz as this is the area in the frequency spectrum that showed the most directivity. Also as mentioned for the HLBR measure, following on from the work of Abad *et al.* [29], we use 200 to 400 Hz as the low band and 3500 to 4500 Hz as the high band.

4.2. Pose estimation

In Section 1 we described the directional nature of the audio energy distribution around a subject's head, our initial measurements confirm this effect. The plots shown in Figure 5 show the measure to the average audio energy received at individual microphones for all subjects. An angle of 0 degrees indicates that the speaker is directly facing the microphone



(a) 1 - 500 Hz

(b) 1000 - 1500 Hz

(c) 4000 - 4500 Hz

(d) 5500 - 6000 Hz

Figure 5: Directivity pattern of audio energy around the human head for different frequency bands. The polar plots show the average audio energy for all 39 subjects for each pose and for each microphone in the array.

and an angle of 180 degrees indicates they are facing away from the microphone. In Figure 5, the difference of the radiation pattern between Figure 5a and Figures 5b and 5c can be seen with the higher frequencies being far more directional. It can be seen in Figure 5 that the directivity of the speech signal also depends on the frequency of the audio signal. Figure 5a shows far less directivity in the frequency band up to 500 Hz as compared with the high frequency bands such as those shown in Figures 5b, 5c and 5d.

In our initial set of experiments we examine the performance of the three audio measures, energy, HLBR and DRR, described in Sections 2.1 and 2.2. We randomly divided the data into training and testing sets based on the subject, so we have 19 subjects in the training set and 20 subjects in the testing set. These experiments show results for training and testing in the same room, in this case room 1. This provides us with an indication of the robustness to changes in speaker as there are no sequences of the same speaker in the training and the test set. Initially we also included all positions in the training and test set, in later experiments we will test the robustness to changes in position and also the effect of training in one room and testing in different room environments.

We use two of the measures proposed for the CLEAR 2006 evaluation: *Pan Correct*

Classification (PCC) and *Pan Correct Classification within a Range* (PCCR). The first one is the standard classification rate defined as $PCC = \frac{N_{corr}}{N_{all}}$ where N_{corr} is the number of correctly recognised poses and N_{all} is the number of poses in the test set. The second one is a modification of the PCC allowing for an error of ± 1 with the adjacent classes, so for example if the actual pose is pose 3 both poses 2 and 4 will be counted as correct. Table 2 shows the results of DRR based audio head pose stimation as compared with energy and HLBR based methods. From this table, we can see that using the DRR measure clearly provides better performance over the two other methods of audio based head pose estimation, with one based on audio energy [24, 25] and the other on high to low band ratio (HLBR) [22].

Additionally, it can be seen in Table 3 that the DRR is robust to changes in the testing location. The energy based and HLBR results show a great deal of variability when tested at different distances from the microphone array, while the results for the DRR measure remain consistently high for all positions. The effect of the subject's location in relation to the microphone array in regards to both training and test is further examined in Section 4.3.

Overall results	PCC	PCCR
Audio Energy	0.60	0.85
HLBR	0.43	0.66
DRR	0.86	1.00

Table 2: Head pose average recognition results when classifiers are trained on the data from all the positions and tested on data from all the positions in room 1. The results are given in recognition rate for each position and also for all positions overall. Results of the classification rates are shown for both *Pan Correct Classification* (PCC) and *Pan Correct Classification within a Range* (PCCR).

Distance from array	0.5m	1.0m	1.5m	2.0m
Audio Energy	0.68	0.74	0.55	0.43
HLBR	0.33	0.50	0.43	0.46
DRR	0.89	0.82	0.89	0.86

Table 3: Head pose recognition results using training data from all positions and testing on different positions. The classifiers are trained on the data from all the positions and tested on the data from each position in room 1 separately. The results are given in terms of the *Pan Correct Classification* (PCC) classification rate.

Tables 4, 5 and 6 show the confusion matrices between the eight head poses for all three methods. Following the description shown in Figure 4, pose angle 0 is directly facing the centre of the microphone array whilst in pose angle 180 the subject is facing away from the microphone array.

It can also be seen in the confusion matrices in Tables 4, 5 and 6 that when errors do occur using the DRR measure the confusion is always with an adjoining pose, as opposed

Pose Angle	0	45	90	135	180	225	270	315
0	65	0	2	0	0	2	2	9
45	6	63	9	2	0	0	0	0
90	2	22	47	8	1	0	0	0
135	4	4	18	30	19	3	0	2
180	7	4	9	8	37	9	2	4
225	9	0	3	9	11	25	15	8
270	3	0	1	2	2	11	43	18
315	6	0	0	2	0	2	14	56

Table 4: Confusion Matrix for audio energy in the band 3500 Hz to 4000 Hz. The results are obtained over all the subjects and positions in the room 1 test set, with the subjects standing directly in front of the 16 element microphone array.

Pose Angle	0	45	90	135	180	225	270	315
0	32	11	2	3	3	9	7	13
45	6	34	21	3	5	0	10	1
90	2	13	32	8	5	7	11	2
135	3	9	10	26	5	16	6	5
180	4	3	10	6	28	13	9	7
225	4	2	4	9	8	32	17	4
270	6	5	4	5	4	8	42	6
315	12	1	3	2	10	6	11	35

Table 5: Confusion Matrix for HLBR. The results are obtained over all the subjects and positions in the room 1 test set, with the subjects standing directly in front of the 16 element microphone array.

Pose Angle	0	45	90	135	180	225	270	315
0	78	1	0	0	0	0	0	1
45	0	77	3	0	0	0	0	0
90	0	8	68	4	0	0	0	0
135	0	0	1	72	7	0	0	0
180	0	0	0	5	60	15	0	0
225	0	0	0	0	7	68	5	0
270	0	0	0	0	0	6	67	7
315	6	0	0	0	0	0	12	62

Table 6: Confusion Matrix for DRR. The results are calculated over all subjects and positions in the room 1 test set, with the subjects standing directly in front of the 16 element microphone array.

to the energy and HLBR measures. Interestingly the confusion matrices also show that for energy and HLBR most of the large errors in pose classification occur when the subject is facing away from the array such as angles 135, 180 and 235. This may be due to the fact that both methods rely on the direct speech signal to the microphones which is greatly reduced as the subject turns away from the microphone array, on the other hand the DRR also includes the reverberant part of the speech as part of the measurement and so is more robust to the subject turning away from the array.

4.3. The effect of distance/room location on pose estimation

In order to further examine the effect of distance from the microphone array on pose estimation we performed experiments of training on data from one position and the testing on all positions. These experiments were conducted using models trained and test in room 1. The positions are the subject standing in front of the centre of the 16 element microphone array at distances of 0.5, 1, 1.5 and 2 metres, these positions labelled as P1, P2, P3 and P4 respectively can be seen in Figure 4. This will demonstrate the robustness of using the training data from only one location in the room to train our classifier model.

	Testing position			
Training position	P1	P2	P3	P4
P1	0.71	0.45	0.23	0.26
P2	0.52	0.66	0.26	0.29
P3	0.39	0.44	0.43	0.27
P4	0.24	0.36	0.38	0.48

Table 7: Percentage recognition rates for audio energy training on one position and testing on the other positions.

The results in Table 7 show that using energy is not robust to training the classifiers in one location and testing in a different location within the room. It can be seen that while

results along the diagonal of Table 7 are reasonable, they fall off sharply when the testing location is different from the training location. This can be seen even more dramatically in the case of the HLBR measure shown in Table 8 where the results for training and testing in different room locations becomes essentially random.

	Testing position			
Training position	P1	P2	P3	P4
P1	0.42	0.23	0.20	0.16
P2	0.25	0.48	0.16	0.19
P3	0.15	0.20	0.48	0.19
P4	0.19	0.21	0.18	0.55

Table 8: Recognition rates for HLBR training on one position and testing on the other positions.

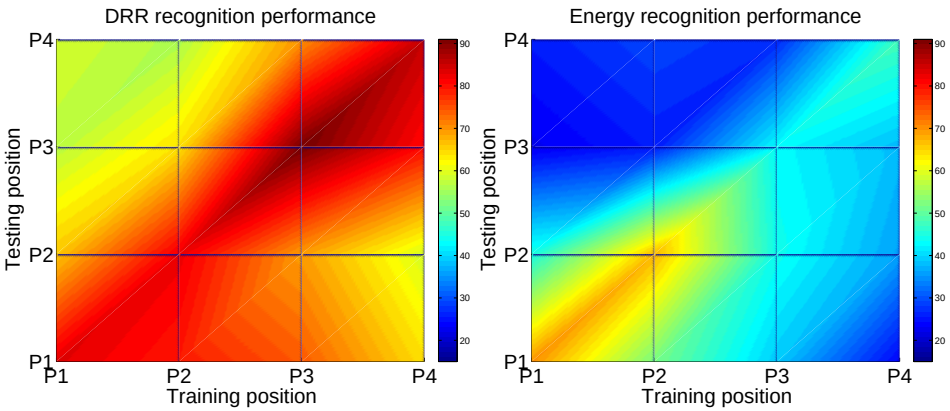
As opposed to the results for energy and HLBR shown in Table 7 and 8 the results for using the DRR in Table 9 show much more robustness to training and testing in different locations. To illustrate this visually we have interpolated and plotted the results of each method as shown in Figure 6. In these plots red colours indicate a higher recognition rate while blue colours indicate a lower recognition rate, ideally we would like to see an even red colour over the entire plot indicating that the method is robust to changes in location between the training and test data. The plots in Figure 7 show that the value of the DRR for a single subject is reasonably consistent over all positions.

	Testing position			
Training position	P1	P2	P3	P4
P1	0.83	0.66	0.57	0.59
P2	0.79	0.82	0.64	0.55
P3	0.74	0.69	0.91	0.70
P4	0.64	0.59	0.80	0.85

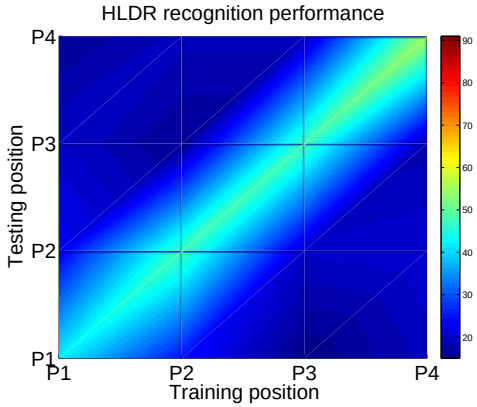
Table 9: Recognition rates for DRR training on one position and testing on the other positions.

4.4. Effect of lateral displacement from the microphone array

In previous experiments the position of the subject was in the centre of the microphone array. We now examine the effect of lateral movement of the subject with relation to the microphone array. Along with distance from the array this gives us a variety of locations within room 1. This is in order to test the performance of the different methods when the subject is not directly in front of the microphone array. All experiments in this section were conducted using models trained with data from room 1 and tested on data collected from room 1.

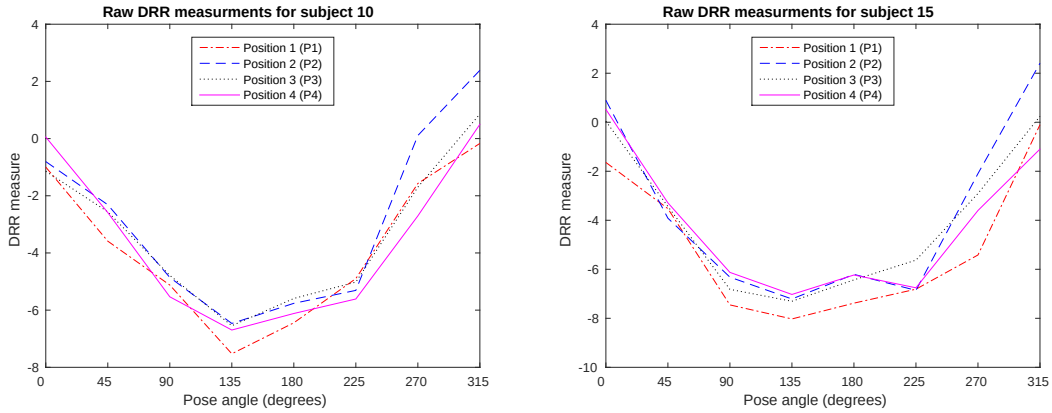


(a) DRR recognition performance (b) Energy recognition performance



(c) HLBR recognition performance

Figure 6: Performance of training on data from one location and testing on the other locations. Red colours indicate a higher recognition rate while blue colours indicate a lower recognition rate. Plots with more even red colours as opposed to strong diagonals indicate more robustness to changes in location.



(a) Subject 10 (b) Subject 15

Figure 7: The DRR values for two randomly selected subjects (10 and 15) over all positions (P1 to P4).

In order to simulate this we divide the microphone array of 16 elements into three smaller sub-arrays of eight elements each. Given that we are now using an array half of the size of those used in previous experiments, the results in this section are not directly comparable to the previous two sets of experiments.

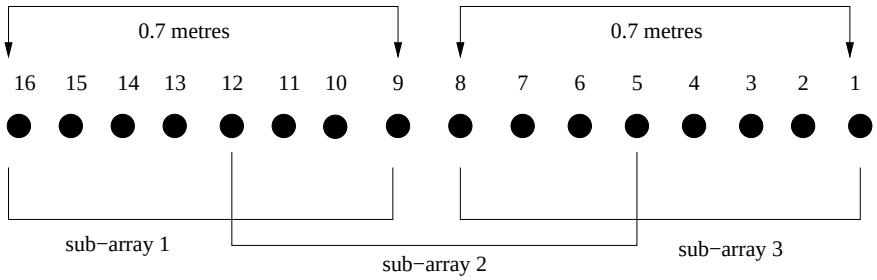


Figure 8: Configuration of microphone sub-arrays for lateral displacement experiments. Each black circle represents an individual microphone and eight microphones from the original array form the three sub-arrays.

The configuration of the three sub-arrays can be seen in Figure 8 with each being a smaller eight element array of 0.7 metres in length. Sub-array 1 simulates the subject standing at the extreme right hand side of the array, sub-array 2 simulates the subject standing in the centre of the array and sub-array 3 simulates the subject standing at the extreme left of the array. These twelve simulated subject positions can be seen in Figure 9. The same division of training and test data as defined in Section 4.2 is used in the following experiments. We train three sets of classifiers one for each sub-array of eight microphones. Therefore, the models trained on sub-array 1 simulate the scenario where the subject is in positions P1+, P2+, P3+ and P4+. For the models trained on sub-array 2, the speaker is in the centre of an eight-element array (P1 to P4). The models trained on sub-array 3 simulate the situation where the speaker stands in positions P1-, P2-, P3- and P4-.

The results shown in Table 10 clearly show that the DRR measure is reasonably robust

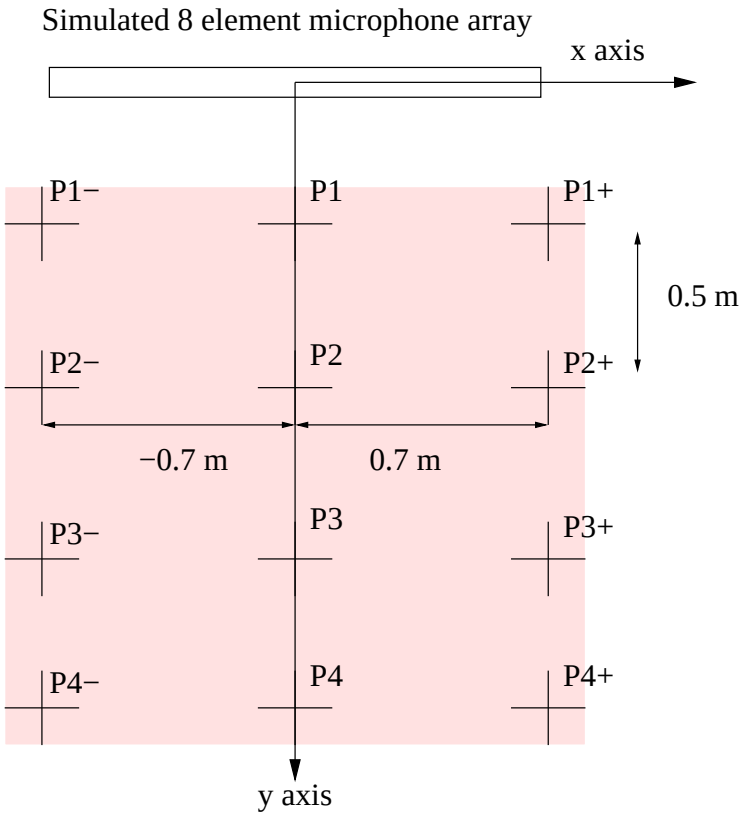


Figure 9: The twelve speaker positions simulated by sampling from the three sub-arrays shown in Figure 8. The shaded area is the region shown in the plots in Figure 10.

to both the reduction in the number of microphones and also changes in the location of the speaker relative to the microphone array. The first column is the average recognition rate of the positions P1- to P4- shown in Figure 9, the second column is the average of the positions P1 to P4 and the third column is the average of the positions P1+ to P4+. The energy and HLBR methods have particularly poor results that seem to mostly be due to the reduction in the number of microphones used. The results for positions P1 to P4 can be compared with those in Table 2 to show the effect of halving the number of elements in the array from 16 to 8. It can be observed that halving the number of microphones results in a 0.14 drop in the recognition rate based on the DRR. However, the performance drop due to the reduced number of microphones is much more significant in the energy and HLBR methods.

	Microphone sub-array		
	1	2	3
Energy	0.39	0.38	0.30
HLBR	0.28	0.30	0.22
DRR	0.56	0.72	0.65

Table 10: Percentage recognition rates for microphone sub-array average over positions P1 to P4. The configuration of the microphone sub-arrays is shown in Figure 8. The first column is the average recognition rate of positions P1- to P4- shown in Figure 9, the second column is the average of the positions P1 to P4 and the third column is the average of the positions P1+ to P4+.

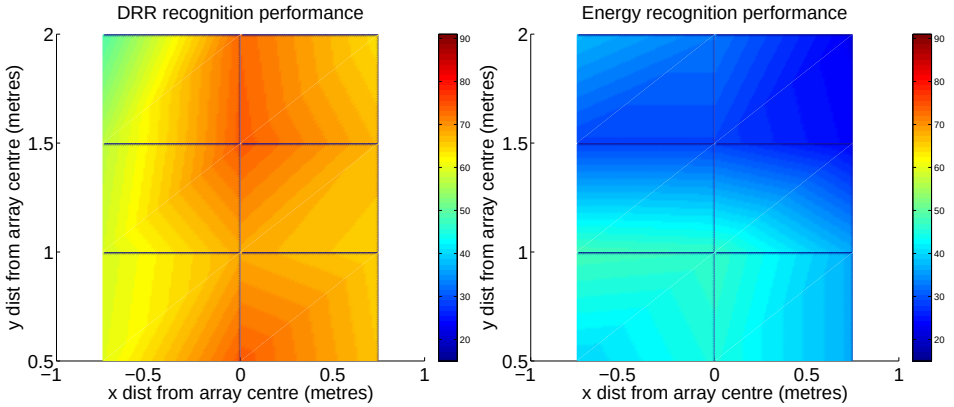
.

In Figure 10 the results for lateral displacement and also distance from the array are plotted. Red colours indicate higher recognition rates, ideally we want the colours to remain even over the plot as the position changes in the x and y directions shown in Figure 9, the shaded area is the area plotted in Figure 10. It can be seen that while the DRR results are the best when the subject is directly in front of the microphone array, i.e. positions P1 to P4, it can still produce reasonable results when the position of the subject is transposed by 0.7 m in either direction along the x axis, i.e. positions P1- to P4- and P1+ to P4+. It can also be seen that the results for both energy and HLBR are generally poor over the entire area when using the smaller eight element microphone sub-arrays.

4.5. Robustness to changes of room and visual occlusions

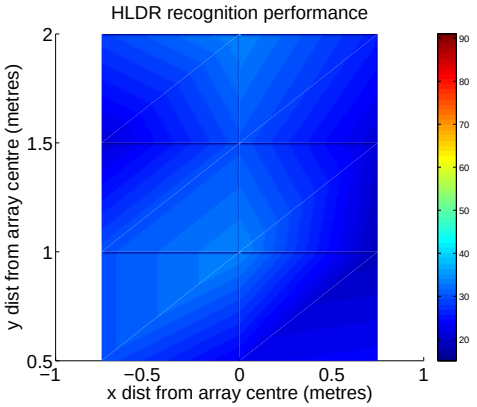
As a final test of the proposed method we recorded subjects in two additional rooms (room 2 and room 3), described in Section 4.1. As opposed to room 1 which was a dedicated recording studio, the additional rooms are ordinary multi-purpose spaces. Room 2 is a large seminar type room and room 3 is a smaller meeting room, furnished like a domestic living room. Photographs of these rooms can be seen in Figure 3.

In order to test the robustness of the methods, we only use the recordings from rooms 2 and 3 as test data, and we use the models trained on data from room 1 for pose classification. This proved quite challenging as can be seen from the initial results in Table 11 showing the recognition rates for the different methods averaged over all positions. These results show



(a) DRR recognition performance

(b) Energy recognition performance



(c) HLBR recognition performance

Figure 10: Performance using three microphone sub-arrays to simulate different room locations. The different simulated room locations were sampled according to the grid pattern shown in Figure 9. Red colours indicate a higher recognition rate while blue colours indicate a lower recognition rate. Plots with more even red colours rather than a strong response in the centre show the robustness of each method to lateral displacement of the subject's position.

quite a dramatic drop in performance for our proposed method when the classification models are trained using data a different room environment to the testing environment. While the performance is reduced, the DRR still outperforms the energy and HBLR methods. The greater drop in performance can be accounted for by the DRR method relying more on the room environment than the other methods. However, when we look at the PCCR measure for the DRR method, we can see that while the method becomes less accurate in a different room environment it does not fail completely, in fact based on the PCCR measure DRR outperforms the two other methods by a significant margin. Tables 12 and 13 show that most errors for the DRR are due to the misclassification of a very similar head pose. It can also be seen that the classification rate of the DRR is lower in the case of room 2 with an RT60 value of 450 ms, as opposed to room 3 with an RT60 of 250 ms.

Method	Room 2 (PCC)	Room 3 (PCC)	Room 2 (PCCR)	Room 3 (PCCR)
Energy	0.42	0.45	0.61	0.67
HLBR	0.42	0.37	0.59	0.55
DRR	0.51	0.56	0.86	0.89

Table 11: Classification rates for all methods in rooms 2 and 3 averaged over all positions. The PCC and PCCR measures described in Section 4.2 are shown.

In a final set of experiments we examine the effect of visual occlusions on the performance of our method. In this scenario we assume we have an audio-visual pose detection system with a camera mounted in the centre of the microphone array. To create an occlusion of this camera we placed a plastic screen between the subject and the array. The screen is placed in position P1 and the subject then performs rotations at positions P2, P3 and P4. This screen can be seen in position in the photograph in Figure 3b. We select a subset of five subjects for each room to perform this additional test. The results in Table 14 show that placing a screen between the array and the subject significantly degrades the performance of the system. One interesting result however was how the performance varied depending on

Pose Angle	0	45	90	135	180	225	270	315
0	37	34	0	0	0	0	0	6
45	1	74	1	0	0	0	0	0
90	0	32	32	11	0	0	0	0
135	0	17	12	37	10	0	0	0
180	2	29	0	13	30	0	2	0
225	1	1	0	11	30	9	9	15
270	0	0	0	0	4	4	18	48
315	0	0	0	0	0	0	1	70

Table 12: Confusion Matrix for DRR tested in room 2 using models trained with data from room 1. The results are obtained over all the subjects and positions in the test set.

Pose Angle	0	45	90	135	180	225	270	315
0	56	14	0	5	0	3	0	2
45	13	21	22	13	0	1	0	0
90	0	0	61	18	0	1	0	0
135	0	0	14	56	10	0	0	0
180	1	0	1	12	60	6	0	0
225	0	0	0	16	30	33	1	0
270	1	0	0	12	5	29	24	9
315	5	0	0	0	0	15	12	48

Table 13: Confusion Matrix for DRR tested in room 3 using models trained with data from room 1. The results are obtained over all the subjects and positions in the test set.

the distance of the subject to the occlusion. It can be seen in Table 15 that the performance improves as the subject moves away from the occlusion. This is because as the subject moves away more microphones become visible and so the speech signal from the subject has an unimpeded path to more microphones thus giving a better estimation of the DRR. Although this is a small sample size of 10 subjects over both rooms, this clearly shows that performance improves as the distance to the occlusion increases.

Method	Room 2 (PCC)	Room 3 (PCC)	Room 2 (PCCR)	Room 3 (PCCR)
Energy	0.39	0.37	0.50	0.46
HLBR	0.37	0.32	0.49	0.43
DRR	0.47	0.49	0.65	0.62

Table 14: Classification rates for all methods in rooms 2 and 3 for 5 subjects with visual occlusion in place averaged over all positions. The PCC and PCCR measures described in Section 4.2 are shown.

Position	Room 2 and Room 3 (PCC)
P2	0.31
P3	0.49
P4	0.64

Table 15: Combined classification rates for the DRR in rooms 2 and 3 for 5 subjects with visual occlusion in place shown for three positions 0.5, 1 and 1.5 metres from the occluding screen respectively. The PCC measure described in Section 4.2 are shown.

5. Conclusions

We have presented here investigations and results on a novel form of audio head pose detection. We compared using the directivity of the audio energy, the HLBR and the DRR

for classifying eight head poses. We show that using the DRR significantly improves the classification performance over the other two methods. We further show that using the DRR improves robustness to changes in the speaker's location in the room and also changes in their position relative to the microphone array. We then measured the performance of a system trained in one room and then tested in two different rooms. These experiments showed that while overall performance decreased our proposed DRR method still outperformed the two other baseline methods. We also tested the effect of visual occlusions of the microphone array on the detection performance, and this showed that performance decreased more significantly as the subject moved closer to the occlusion. We believe this current work could be extended to audio visual head pose estimation where the advantages of audio over video, such as robustness to lighting changes and occlusions, could be exploited to the improve the overall performance.

Acknowledgement

The authors wish to thank the associate editor and the anonymous reviewers for their very helpful contributions to improving the quality of this paper. This research was supported by the Engineering and Physical Sciences Research Council of the UK (grant no. EP/H050000/1 and EP/K014307/1). The support from EPSRC and the MOD University Defence Research Collaboration (UDRC) in Signal Processing is gratefully acknowledged

References

[1] S. T. Shivappa, B. D. Rao, M. M. Trivedi, Role of head pose estimation in speech acquisition from distant microphones, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2009.

[2] S. Ba, J. Odobez, Multiperson visual focus of attention from head pose and meeting contextual cues, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (1) (2011) 101–116.

[3] M. Ubilla, D. Mery, R. F. Cadiz, Head tracking for 3d audio using the nintendo wii remote, in: Proceedings of the International Computer Music Conference, 2010.

[4] E. Murphy-Chutorian, M. M. Trivedi, Head pose estimation in computer vision: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (4) (2009) 607–626.

[5] H. K. Dunn, D. W. Farnsworth, Exploration of pressure field around the human head during speech, The Journal of the Acoustical Society of America 10 (3) (1939) 184–199.

[6] A. Warnock, W. Chu, J.-C. Guy, Directivity of human talkers, Canadian Acoustics 30 (3).

[7] P. Meuse, H. F. Silverman, Characterization of talker radiation pattern using a microphone array, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1994.

[8] H. Kuttruff, Room Acoustics, Taylor and Francis, 2000.

[9] W. Chu, A. Warnock, Detailed directivity of sound fields around human talkers, Tech. Rep. December, Institute for Research in Construction National Research Council Canada Tech Rep (2002).

[10] M. Barnard, P. Koniusz, W. Wang, J. Kittler, S. M. Naqvi, J. Chambers, Robust multi-speaker tracking via dictionary learning and identity modeling, Multimedia, IEEE Transactions on 16 (3) (2014) 864–880.

[11] V. Kilic, M. Barnard, W. Wang, J. Kittler, Audio assisted robust visual tracking with adaptive particle filtering, Multimedia, IEEE Transactions on 17 (2) (2015) 186–200.

[12] S. M. Naqvi, M. Yu, J. A. Chambers, A multimodal approach to blind source separation of moving sources, IEEE Journal of Selected Topics in Signal Processing.

[13] D. Zotkin, R. Duraiswami, L. Davis, Joint audio-visual tracking using particle filters, EURASIP Journal of Applied Signal Processing 2002 (1) (2002) 1154–1164.

[14] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, I. McCowan, Audio-visual probabilistic tracking of multiple speakers in meetings, IEEE Transactions on Audio, Speech, and Language Processing 15 (2007) 601–616.
[15] Q. Nguyen, J. Choi, Audio-visual data fusion for tracking the direction of multiple speakers, in: Proceedings of the International Conference on Control, Automation and Systems, 2010.
[16] C. Wang, M. Brandstein, A hybrid real-time face tracking system, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1997.
[17] K. Nickel, T. Gehrig, H. K. Ekenel, J. W. McDonough, R. Stiefelhagen, An audio-visual particle filter for speaker tracking on the clear'06 evaluation dataset, in: CLEAR, 2006, pp. 69–80.
[18] B. Matteo, T. Murtaza, C. Andrea, Multi-modal particle filtering tracking using appearance, motion and audio likelihoods, in: Proceedings of the IEEE International Conference on Image Processing, 2007.
[19] M. Arulampalam, S. Maskell, N. Gordon, T. Clapp, A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking, IEEE Transactions on Signal Processing 50 (2) (2002) 174–188.
[20] M. F. Fallon, S. Godsill, A. Blake, Joint acoustic source location and orientation estimation using sequential Monte Carlo, in: Proceedings of the International Conference on Digital Audio Effects, 2006.
[21] A. Brutti, O. Lanz, A joint particle filter to track the position and head orientation of people using audio visual cues, in: Proceedings of the European Signal Processing Conference, 2010.
[22] C. Segura, C. Canton-Ferrer, A. Abad, J. Casas, J. Hernando, Multimodal head orientation towards attention tracking in smartrooms, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2007.
[23] C. Canton-Ferrer, C. Segura, J. Casas, M. Pardàs, J. Hernando, Audiovisual head orientation estimation with particle filtering in multisensor scenarios, EURASIP Journal of Advanced Signal Processing 2008.
[24] J. M. Sachar, H. F. Silverman, A baseline algorithm for estimating talker orientation using acoustical data from a large-aperture microphone array, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2004.
[25] A. Levi, H. F. Silverman, A robust method to extract talker azimuth orientation using a large-aperture microphone array, IEEE Transactions on Audio, Speech and Language Processing 18 (2) (2010) 277–285.
[26] A. Brutti, M. Omologo, P. Svaizer, Oriented global coherence field for the estimation of the head orientation in smart rooms equipped with distributed microphone arrays, in: Ninth European Conference on Speech Communication and Technology, 2005.
[27] A. Brutti, M. Omologo, P. Svaizer, C. Zieger, Classification of acoustic maps to determine speaker position and orientation from a distributed microphone network, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2007.
[28] P. Svaizer, A. Brutti, M. Omologo, Environment aware estimation of the orientation of acoustic sources using a line array, in: iProceedings of the European Signal Processing Conference, 2012.
[29] A. Abad, C. Segura, C. Nadeu, J. Hernando, Audio-based approaches to head orientation estimation in a smart rooms, in: Proceedinsg of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2007.
[30] B. Mungamuru, P. Aarabi, Enhanced sound localization, IEEE Transactions on Systems, Man, and Cybernetics, Part B 34 (3) (2004) 1526–1540.
[31] A. W. Bronkhorst, T. Houtgast, Auditory distance perception in rooms, Nature 397 (6719) (1999) 517–520.
[32] D. Mershon, L. King, Intensity and reverberation as factors in the auditory perception of egocentric distance, Attention, Perception, and Psychophysics 18 (1975) 409–415.
[33] Y. Lu, M. Cooke, Binaural estimation of sound source distance via the direct-to-reverberant energy ratio for static and moving sources, IEEE Transactions on Audio, Speech and Language Processing 18 (7) (2010) 1793–1805.
[34] M. Barnard, W. Wang, J. Kittler, Audio head pose estimation using the direct to reverberant speech

ratio, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2013.

[35] G. Lathoud, J. M. Odobez, D. Gatica-Perez, AV16.3: an audio-visual corpus for speaker localization and tracking, in: Proceedings of the MLMI Workshop, 2004.

[36] M. Jeub, C. M. Nelke, C. Beaugeant, P. Vary, Blind estimation of the coherent-to-diffuse energy ratio from noisy speech signals, in: Proceedings of the European Signal Processing Conference, 2011.