AUDIO HEAD POSE ESTIMATION USING THE DIRECT TO REVERBERANT SPEECH RATIO

Mark Barnard, Wenwu Wang and Josef Kittler

The 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 used visual information to model and recognise a subject's head in different configurations. These approaches have a number of limitations such as, 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 highly reverberant meeting 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 hypothesis is confirmed by experiments conducted on the publicly available AV16.3 database.

Index Terms— 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] or as a cue for modelling interactions among multiple speakers. Most current work on head pose estimation has focused on the use of visual information [2]. These approaches however have a number of limitations, such as inability to cope with occlusions, low resolution and differing head/facial appearances. 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 recognition system.

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 [3]. 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 ratio of direct speech to reverberant speech (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 [3, 4]. Recent research [5] 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.

As opposed to existing pose esitmation methods whose performance is prone to reverberation degradation, such as [6, 7], 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 DRR should provide a useful measure of head pose/orientation. It can be seen from the previous paragraph that this ratio is dependent on four factors: the size of the room, reflective nature of the room surface, distance from the speaker and direction of the audio source. Given that the size of the room and the materials of the room surfaces are constant, if we normalise over the distance from the speaker to the microphone, then the DRR comes to depend solely on the orientation of the source.

The aim of this paper is demonstrate that using audio information alone, the DRR, we can obtain an indication of the subject's head pose.

2. DRR ESTIMATION

In order to estimate the DRR we use the method proposed by Jeub et el. [8] to estimate the coherent to diffuse energy ratio. A noise field can be defined as a mixture of coherent

This research was supported by the Engineering and Physical Sciences Research Council of the UK (grant no. EP/H050000/1).

or diffuse noise fields. If we have two audio signals $x_1(k)$ and $x_2(k)$ with discrete time index k, the complex noise field coherence in the frequency domain is defined as [8, 9]:

$$\Gamma_{x_1x_2}(\Omega) = \frac{\Phi_{x_1x_2}(\Omega)}{\sqrt{\Phi_{x_1x_1}(\Omega) \cdot \Phi_{x_2x_2}(\Omega)}}, \quad (2)$$

where $\Phi_{x_1x_1}(\Omega)$ and $\Phi_{x_2x_2}(\Omega)$ are the auto-power spectral density (PSD) of the microphone signals $x_1(k)$ and $x_2(k)$ and $\Phi_{x_1x_2}(\Omega)$ 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

$$\Gamma^{(mix)}_{x_1x_2}(\Omega) = \frac{\sum_{n=1}^{N} \Phi^{(n)}_{x_1x_2}(\Omega)}{\sqrt{\sum_{n=1}^{N} \Phi^{(n)}_{x_1x_1}(\Omega) \cdot \sum_{n=1}^{N} \Phi^{(n)}_{x_2x_2}(\Omega)}}, \quad (3)$$

where N 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. For this case Jeub et al. [8] proposed the following heuristically motivated equation

$$DRR(\Omega) = \frac{|sinc(\Omega f_s d_{mic}/c)|^2 - |\Gamma^{(mix)}_{x_1x_2}(\Omega)|^2}{|\Gamma^{(mix)}_{x_1x_2}(\Omega)|^2 - 1}, \quad (4)$$

where d_{mic} is the distance between the two microphones in metres (in our case 0.8 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 $f > f_0$

- $\Gamma_{x_1x_2}(\Omega) < 0.1 \rightarrow$ diffuse noise
- $\Gamma_{x_1x_2}(\Omega) > 0.9 \rightarrow$ coherent noise

In practice the input signals x_1 and x_2 are segmented in windows of length $L = 320$ samples with an overlap of 75%. These windows are then transformed with FFT of length $M = 512$. For each discrete frequency bin $u = 0, \ldots U$, where $U = 256$ the value of $DRR(\Omega)$ is estimated and these are then averaged over the frequency bins,

$$DRR = \frac{1}{U} \sum_{u=0}^{U} DRR(u), \quad (5)$$

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 [8].

3. DISTANCE EFFECT AND NORMALISATION

As mentioned in the introduction in order to generalise our DRR measure we need to normalise over the distance from the microphone. This will allow us to compare DRR measurements for different locations in the room. The level of the direct signal is dependent on the distance from the speaker to the listener or microphone. This is characterised by the so called "6 dB rule" where the average level of direct speech falls by 6 dB for every doubling of distance from the lips [10].

As we have the position of the speaker's head from a visual tracking system [11] we can calculate the three dimensional Euclidean distance d from the speaker to the microphone array. This gives us a simple normalisation factor of

$$N_d = \frac{1}{d}. \quad (6)$$

The normalised DRR measure is given by

$$DRR_N = N_d \cdot DRR. \quad (7)$$

We used three sequences from the AV 16.3 dataset, described in Section 4.1, to measure the effect of distance on the value of the DRR. Sequences 1, 2 and 3 of the dataset feature a subject moving to different locations in the room, adopting a single head pose and reciting the digits "one" to "ten". We selected poses directly facing the microphone array i.e. from -5 to +5 degrees and measured the DRR at each location. This should give an indication of the effect of distance alone on the DRR. The result can be seen in Figure 1 and confirms that $1/d$ would be a reasonable initial approximation for normalisation (see the fitted curve).

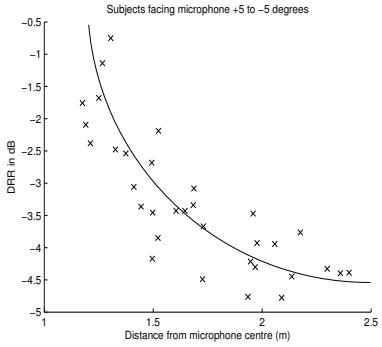

Fig. 1. Plot of DRR for different distances. In all cases the subject is directly facing the microphone array. These measurements were made on sequences 1, 2 and 3 of the AV16.3 dataset.

4. HEAD POSE ESTIMATION

4.1. Dataset and Pose Definition

Experiments were conducted on sequences from the AV16.3 dataset [12]. The data was collected in a meeting room en-

vironment with two circular eight element microphone arrays as shown in Figure 2. Visual data was also collected from three calibrated cameras mounted at the three top corners of the room. The room dimensions are 8.2 m long, 3.6 m wide and 2.4 m high with an approximate reverberation time of $RT_{60} = 0.5$ seconds at 1000 Hz.

To measure the effect of changes in head pose to the DRR we use sequences 5 and 6 from the data set, as these are the only sequences annotated for head pose. In these two sequences the subjects stand in eight locations in the room as shown in Figure 2. In each location the subjects adopt a number of head poses, for most positions there are five orientations from "North" to "South" as shown in Figure 3 and recite the numbers "one" to "ten". The range of head poses can be seen in Figure 3. The subjects only move their head in the horizontal plane and there is no tilting of the head. Examples of different head positions can be seen in Figure 5.

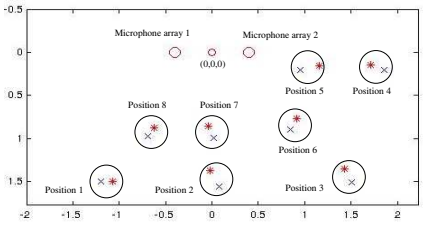

Fig. 2. Speaker positions in relation to the position of the microphone arrays. Distances of the speakers from the microphone array centre vary from approximately 1.25 m to 2.10 m. Red stars show positions for Sequence 5 and blue crosses the positions for Sequence 6.

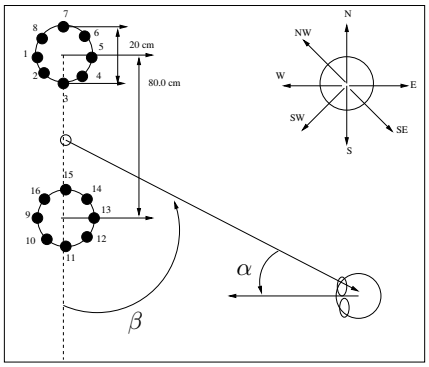

Fig. 3. Microphone array configuration showing estimation of the pose angle α and the range of head poses adopted in the data.

The direction of the speaker's head in relation to the centre of the microphone arrays is calculated as shown in Figure 3. In order to do this the position of the speaker must be accurately estimated. We did this by using the visual head tracking system described in [11] to produce a three dimensional position at each visual frame. With the various positions and the different head poses at each position this gives us a range of speaker directions relative to the microphone arrays.

4.2. DRR versus Head Pose

For each head pose we used the two microphone arrays to create a set of eight binaural sources. Each binaural source is composed of a microphone from each array so as to keep the baseline between the microphones at a constant angle. For example pairs would be (1, 9) or (3, 11). The average of each DRR measurement from these microphone pairs is taken as the DRR for that position and head pose. The results for two room locations from both speakers can be seen in Figures 4 and 6. It can be seen that the value of the DRR for both speakers is clearly higher as the angle of the speaker's mouth to the microphone arrays approaches zero. It can also be seen that the DRR for each speaker at each position is similar, and this confirms that the DRR is robust to differing energy levels of different speakers. However it can be seen that the DRR decreases with the distance from the microphone, this shows that distance while not influencing the distribution of the DRR over various head poses does influence the absolute value of the DRR.

We also found that the angle of the speaker to the baseline between the binaural microphone pairs, i.e. angle β in Figure 3, does not affect the distribution of the DRR. In the results shown in Figure 4 and Figure 6, the angle β for each position is approximately 0 and 45 degrees respectively. This confirms the results reported in [8], despite the fact that we have a much larger distance between microphones, 0.8 m as opposed to 0.15 m in [8].

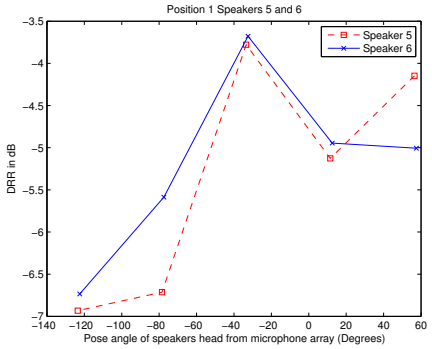

Fig. 4. DRR measurements for both subjects for position 1 at a range of head pose angles. Distance of the speakers to the microphone array centre is approximately 2.05 m.

4.3. An Estimation/Recognition Example

To recognise head pose we must normalise for the distance of the speaker from the microphones. For this we apply the nor-

Fig. 5. Video frames showing some of the head poses in the data set. The subject is standing in position 3.

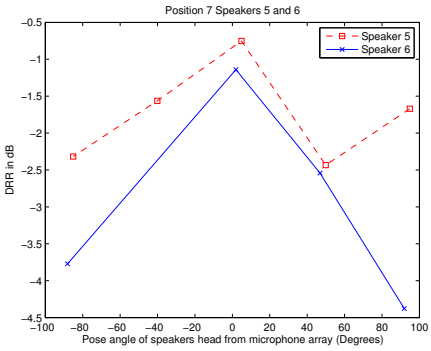

Fig. 6. DRR measurements for both subjects for position 7 at a range of head pose angles. Distance of the speakers to the microphone array centre is approximately 1.3 m. The DRR measure for subject 6 at -40 degrees is omitted because it was above 1 and so considered unreliable [8].

malisation factor N_d from equation (6). In this set of experiments we define one sequence as the training sequence and the other sequence as the test sequence, then reverse these two. We define the angle between the direction of the lips and the direction of arrival (DOA) of the speaker to the centre of the microphone arrays as α, as shown in Figure 3. We quantise the pose angle into four pose classes defined by: $-30 < \alpha < 30$, $-30 > \alpha > -60$ & $30 < \alpha < 60$, $-60 > \alpha > -90$ & $60 < \alpha < 90$ and $-90 > \alpha$ & $\alpha > 90$ where all angles are in degrees.

Using the training set we calculate the mean of the normalised ratio DRR_N for each of the head pose classes defined in the previous paragraph. If we take the mean of each class as the centre of the distribution of the DRR_N for each class we can then define the boundary between classes as the mid-point between these centres. Using this simple thresholding we conducted a small series of pose recognition experiments. The metric used in these experiments is the *Pan Correct Classification* (PCC) defined as the percentage recognition rate within the 4 classes each spanning 30 degrees [13]. Unfortunately due to the nature of the DRR our method cannot be directly compared with other methods [7, 13]. Our method currently gives an offset from zero degrees (directly facing the microphones) instead of a single pose angle, so 60 degrees and -60 degrees are the same class. The results of training on one sequence and testing on the other are shown in Table 1. We show the recognition results using normalised and unnormalised DRR values for training and testing on each sequence in turn. While not directly comparable, our proposed method shows similar results for PCC to those in [7, 13]. The advantage of normalising for the distance from the microphone can clearly be seen.

Training Sequence	DRR	DRR_N
Sequence 05	0.51	0.60
Sequence 06	0.31	0.50

Table 1. Results of recognition experiments using one sequence for training and the other for testing. Showing the classification rate for both DRR and DRR_N.

5. RELATION TO PRIOR WORK

There have been a number of approaches to head pose estimation using audio information. Some of these take advantage of the fact that speech is not radiated evenly from a subject's head [14]. Matching the head pose to this uneven radiation pattern around the head has been used for audio head pose estimation [15, 6]. However these approaches require a large-aperture microphone array covering the walls of the room, while our method works with a simple portable microphone pair. Another approach is to exploit the fact that different frequencies radiate differently from a speaker, with high frequencies being much more directional than lower frequencies. In [7, 13] a ratio of high frequency energy to low frequency energy is proposed for head pose estimation. A modified version of the SRP-PHAT algorithm is proposed by Mungamuru and Aarabi [16]. While this algorithm is principally used for source localisation, using speaker and microphone directivity it can also give an estimation of speaker head orientation. Whereas these methods suffer in reverberant environments, our proposed method actually takes advantage of a reverberant environment to estimate head pose.

6. CONCLUSIONS

We have presented here initial investigations and results on a novel form of audio head pose detection based on the DRR. We show that as the pose angle of the speaker's head relative to the microphones increases, the value of the DRR decreases. We have observed this effect for all speaker positions in the data. We have shown that the value of the DRR depends upon the pose orientation of the speakers head relative to the microphones and the distance of the speaker from the microphones. A simple normalisation based on the distance of the speaker from the microphone is introduced. Using this normalisation for distance we show that reasonable pose recognition results can be obtained. We believe that these results could be improved with a more detailed analysis of the effect of room acoustics on the estimation of the DRR.

7. REFERENCES

[1] S. T. Shivappa, B. D. Rao, and M. M. Trivedi, "Role of head pose estimation in speech acquisition from distant microphones," in *ICASSP*, 2009, pp. 3557–3560.

[2] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 607–626, 2009.

[3] A. W. Bronkhorst and T. Houtgast, "Auditory distance perception in rooms," *Nature*, vol. 397, no. 6719, pp. 517–520, Feb. 1999.

[4] D. Mershon and L. King, "Intensity and reverberation as factors in the auditory perception of egocentric distance," *Attention, Perception, and Psychophysics*, vol. 18, pp. 409–415, 1975.

[5] Y.C. Lu and M. Cooke, "Binaural estimation of sound source distance via the direct-to-reverberant energy ratio for static and moving sources," *IEEE Trans. on Audio, Speech & Language Processing*, vol. 18, no. 7, pp. 1793–1805, 2010.

[6] A. Levi and H. F. Silverman, "A robust method to extract talker azimuth orientation using a large-aperture microphone array," *IEEE Trans. on Audio, Speech & Language Processing*, vol. 18, no. 2, pp. 277–285, 2010.

[7] C. Canton-Ferrer, C. Segura, J.R. Casas, M. Pardàs, and J. Hernando, "Audiovisual head orientation estimation with particle filtering in multisensor scenarios," *EURASIP J. Adv. Signal Process*, vol. 2008, Jan. 2008.

[8] M. Jeub, C. M. Nelke, C. Beaugeant, and P. Vary, "Blind estimation of the coherent-to-diffuse energy ratio from noisy speech signals," in *European Signal Processing Conference (EUSIPCO)*, 2011.

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

[10] D. Davis and C. Davis, *Sound System Engineering*, H. W. Sams, 1987.

[11] M. Barnard, W. Wang, J. Kittler, S.M.R. Naqvi, and J.A. Chambers, "A dictionary learning approach to tracking," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012.

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

[13] A. Abad, C. Segura, C. Nadeu, and J. Hernando, "Audio-based approaches to head orientation estimation in a smart-room," in *INTERSPEECH*, 2007, pp. 590–593.

[14] P. Meuse and H. F. Silverman, "Characterization of talker radiation pattern using a microphone array," in *ICASSP*, 1994.

[15] J. M. Sachar and H. F. Silverman, "A baseline algorithm for estimating talker orientation using acoustical data from a large-aperture microphone array," in *ICASSP*, 2004.

[16] B. Mungamuru and P. Aarabi, "Enhanced sound localization," *IEEE Trans. on Systems, Man, and Cybernetics, Part B*, vol. 34, no. 3, pp. 1526–1540, 2004.