HISTOGRAM BASED DOA ESTIMATION FOR SPEAKER LOCALISATION IN REVERBERANT ENVIRONMENTS

Matthew Trinkle (1) and Ahmad Hashemi-Sakhtsari (2)

(1) Department of Electrical Electronic Engineering, The University of Adelaide, Adelaide SA 5005
(2) Language Technology and Fusion Group, National Security and ISR Division, Defence Science and Technology organization – Edinburgh – SA 5111

ABSTRACT

This paper introduces a Direction of Arrival (DOA) estimation technique for microphone arrays that has improved robustness in reverberant environments. The technique applies a separate beamformer at each frequency to estimate the DOA of the speech signal at that frequency. A histogram of the DOA estimates from all frequencies is then formed, which represents the probability of the signal coming from any particular direction. The histograms can be averaged over consecutive data blocks to improve the reliability. The peak of this angular probability distribution gives an estimate of the DOA. This technique can be used to determine the probability of the person being at any particular location in the room, by combining the angular probability distributions from multiple microphone arrays at different locations in the room.

Index Terms— Speaker Localisation, Microphone Array, DOA Estimation, Room Reverberation.

1. INTRODUCTION

Several methods have been used to track speaker locations using multiple microphones. Some are based on time difference of arrival measurements (TDOA) between microphones, [1],[2] from which the speaker position can be obtained by hyperbolic location [3]. Other techniques use multiple arrays of closely spaced microphones from which the direction of arrival (DOA) information can be combined to obtain the speaker location [4]. One of the main challenges in all of these techniques is the highly reverberant acoustic environment that exists in most rooms.

In this paper we consider obtaining the speaker location from the DOA measurements from multiple widely separated microphone arrays. Each array provides the DOA information about the speaker which is then fused to obtain the speaker location.

An effective way of fusing the DOA information from multiple microphone arrays is described in [4]. A Spatial Likelihood function (SLF) is introduced which gives the probability of the speaker being at any position in the room given the measurements from the microphone arrays. The signals from each microphone array are processed to determine the probability of the signal coming from a particular direction. This angular probability function from each array can be combined into a SLF for the entire room using simple geometry and the resulting SLF's can be combined to give a combined SLF for all the microphone arrays. In some cases the SLF's from each array are simply added, but in [4] it is suggested that not all the SLF's should be weighted equally as the SLF's from some arrays may be more reliable than others due to their distance from the speaker or physical obstructions.

A number of standard DOA estimation techniques can be applied to estimate the probability distribution of the DOA for each microphone array and hence the STF. These include MUSIC, Maximum Likelihood (ML), but a narrow-band signal is assumed. Due to the broadband nature of the audio signals, a focusing technique is typically applied to allow these algorithms to operate at a single frequency [5]. The SRP-PHAT [4] algorithm uses an alternative approach of dealing with the broadband nature of audio signals by estimating the DOA from the cross-correlation function between all pairs of microphones. It achieves improved robustness against reverberation by implementing the generalised cross-correlation function (GCC), which treats all frequency components equally by normalising them. Considering the TDOAs between all microphone pairs can be computationally intense if it needs to be done for all positions in the room and a simplified version that makes the plane wave assumption has been introduced in [6].

In this paper we introduce an algorithm that is similar to the SRP-PHAT in that it treats all frequency components of the signal equally. It also makes the plane wave assumption as in [6]. However, rather than combining the information from all frequencies before forming the TDOA and resulting DOA estimate, this technique first estimates the DOA of the signal within each frequency bin using the relative phase shifts between microphones, and then forms a histogram of these DOA estimates to obtain the probability distribution of

978-1-4799-8389-6/15/$31.00

the DOA of the speaker signal. This technique is expected to be robust against reverberation, as although some frequency components may have very large DOA errors due to reverberation, the majority of frequency components are expected to be close to the true direction. Averaging the histograms of the DOA estimates over several snapshots is also expected to improve the results significantly, as the effect of reverberation on different frequency components is expected to change significantly with relatively small movements of the speaker.

2. DOA ESTIMATION TECHNIQUE

Each microphone array is assumed to consist of a linear array of N microphones spaced by a distance d. The signal incident on each microphone is modelled as a plane wave. Let the k'th bin of the K point Fourier Transform of the n'th microphone be represented by: $X_n(k)$. The frequency of the k'th bin is given by: $f_k = (k/K)f_s$. The DOA within each frequency bin is estimated by finding the angle for which the conventional beamformer $P_{CBF}(k,\theta)$ has maximum power, where $P_{CBF}(k,\theta)$ operates on the k'th frequency bin and adds the signals from direction θ in phase.

$$P_{CBF}(k,\theta) = \underline{v}_k^H R(k) \underline{v}_k \quad (1)$$

Where the l^{th} component of the v vector for the k'th FFT bin is given by:

$$\underline{v}_k(l) e^{-j\frac{2\pi d(l-1)sin(\theta)f_k}{c}}$$

And the covariance matrix from the microphone outputs for the k'th FFT bin is given by:

$$R(k) = E\{X(k)X(k)^H\}$$

Where

$$X(k) = \begin{bmatrix} X_1(k) \\ \vdots \\ X_N(k) \end{bmatrix}$$

The expectation operator is approximated by averaging the FFT outputs from several consecutive partially overlapping data blocks.

In practice only those frequency bins with reasonable signal to noise power ratio are used to obtain a DOA estimate. Thus the algorithm initially estimates the background noise power in each FFT bin and then only finds the DOA in those frequency bins that exceed the noise power in the respective bin by a pre-determined threshold.

Thus the basic procedure is as follows:

(1) Capture a data block when there is no speech.
(2) Split this data block into blocks of length K and perform a FFT on each block.
(3) Find the largest magnitude squared for each FFT bin over all data blocks and define this as the noise floor for that particular FFT bin $N(k)$.
(4) Start capturing data blocks of length M with speech signals on each microphone.
(5) For each data block: Estimate the spatial covariance matrix $R(k)$ for each frequency bin. This is done by splitting the data into smaller data blocks of size K with P% overlap and then doing a FFT on each data block. The spatial covariance matrix for each frequency bin k is then averaged from the outputs of the k'th bin of each FFT. Note that with P% overlap the number of blocks over which the covariance matrix is averaged is approximately: $(M / (K(1-P))$
(6) If the average power in any frequency bin exceeds the noise power $N(k)$ by a factor of T, then the DOA of the signal is calculated in that particular frequency bin. The average power in each frequency bin is determined by averaging the power of each microphone signal, obtained from the diagonal elements of the spatial covariance matrix for that frequency bin.
(7) The DOA in each frequency bin k is obtained from the spatial covariance matrix by finding the peak value of equation (1).
(8) The DOA's of all frequency components are arranged in a histogram $P(\theta)$ with bin spacing: θ_s.
(9) Steps 4 to 8 are repeated for each new data block, and an averaged version of the angle pdf is obtained from $P_{av}(\theta) = \alpha P_{av}(\theta) + P(\theta)$, where α is a forgetting factor.

There are a number of parameters that need to be set in the above algorithm including:

K: The length of the FFT.
M: The length of the data block from which to estimate the spatial covariance matrix for each frequency.
P: The amount of overlap between FFT's
T: The SNR in each frequency bin at which to calculate the DOA measurement for that particular frequency.
θ_s: The angular spacing of the histogram bins.
α: The forgetting factor in the averaging of the angular histograms.

The choice of the above parameters not only affects the performance but also the computational complexity and hence the speed at which the algorithm is able to run. To improve the speed another parameter was added to detect if speech was present before processing the data block by simply comparing the power of the entire data block to a predefined threshold.

3. EXPERIMENTAL RESULTS

The algorithm was tested on some real data collected from an eight element microphone array with five people sitting in-front of the array at different positions. The microphones in the array were spaced by 10 cm.

The sample rate was 44100 Hz, and the parameters used for this experiment were: $K = 512$, $M = 2000$, $P = 90\%$, $T = 4$, $\theta_s = 7$, $\alpha := 0.7$.

The first part of the data block had no speech so that the noise spectrum $N(k)$ could be evaluated following steps 1-3. The resulting noise spectrum for the bottom 10 KHz of the data collected is shown in Figure 1, where each of the individual FFT results are shown in colour (shades) and the maximum value $N(k)$, is shown in black.

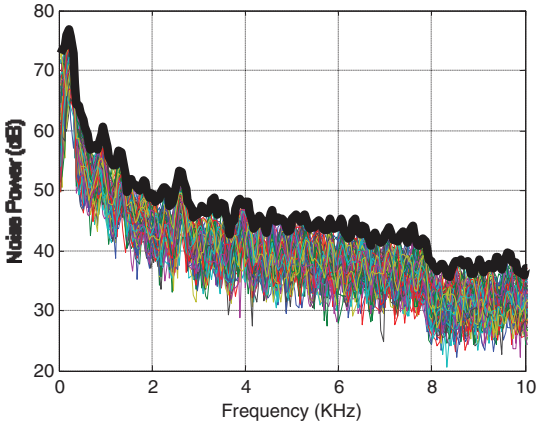

Figure 1: Noise power in each frequency bin $N(k)$

The first 2000 point data block that exceeds the power threshold for speech detection is processed according steps 5 and 6. This data block forms the first part of an utterance (the word "ten", as shown by the framed part of the waveform in Figure 2. This person was sitting about 30 degrees from the centre of the microphone array.

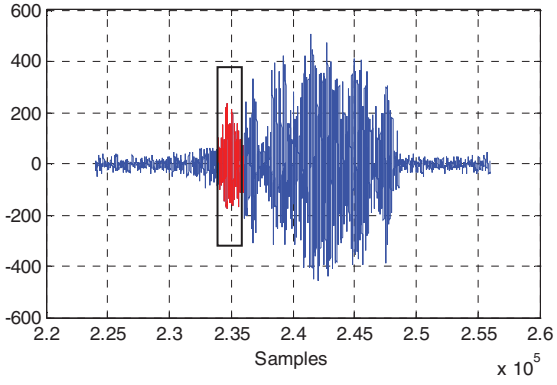

Figure 2: First data block in time domain

The average power in each frequency bin is compared with the noise threshold in Figure 3.

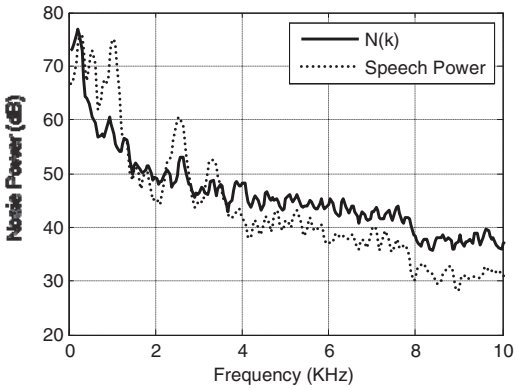

Figure 3: Power in each frequency bin compared with noise floor

Applying the beamforming algorithm to each frequency bin where the speech power exceeds the noise threshold by a factor of $T = 4$ gives the DOA estimates in Figure 4.

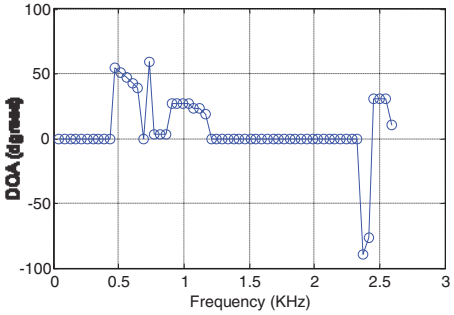

Figure 4: DOA estimates against frequency

Applying steps 5 and 7 to the first 6 data blocks in the utterance gives the DOA estimates in Figure 5, where each colour (shading) corresponds to a different data block. These first 6 data blocks cover almost the entire utterance. Clearly the DOA estimates in each frequency bin do not stay constant even within one utterance. This is expected to be due to the complicated multi-path environment. However the majority of the DOA measurements do cluster around the correct value of 30 degrees.

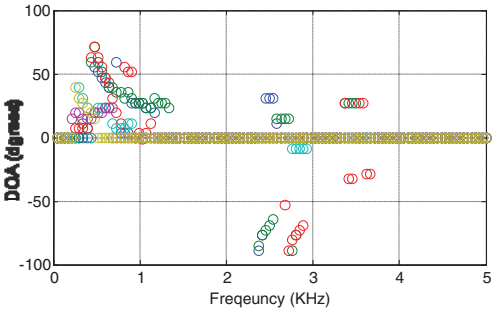

Figure 5: DOA estimates over multiple data blocks

For each of the six data blocks, a histogram of the DOAs is generated using a bin spacing of 7 degrees. These histograms are then averaged using a forgetting factor of 0.7 according to steps 8 and 9. The resulting histogram is shown in Figure 6 and shows a clear peak near 30 degrees.

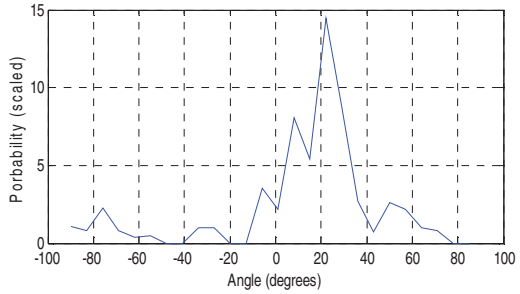

Figure 6: Averaged DOA Histogram over first 6 data blocks

This procedure is repeated for the remainder of the recording, in which four participants at different directions were counting numbers in turn. The first person was near 30 degrees the second near 0 the third near -20 and the fourth near -50 degrees. The resulting averaged histogram is shown as a function of time in Figure 7.

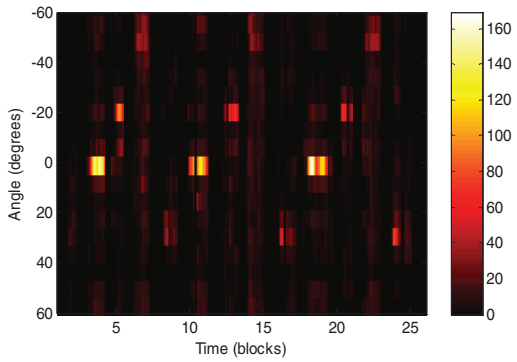

Figure 7: Angle histogram as a function of time with four people talking in turn

The speaker at 0 degrees was the loudest and dominates the histograms, while the speaker at -20 degrees is also quite clear, but the speakers at 30 and -50 are barely distinguishable. If the limits on the above plot is scaled so that the values lie between 30 and 50 then the stronger speaker no longer dominates and all four speakers are relatively clear as shown in figure 8.

The above results are for single words spoken from each location. The performance of the algorithm is expected to increase if longer utterances are spoken at each location as longer averages can be used. In this scenario the forgetting factor was set to be quite small at 0.7 to handle the rapidly changing speaker directions. With longer utterances a forgetting factor closer to unity would give longer averaging times.

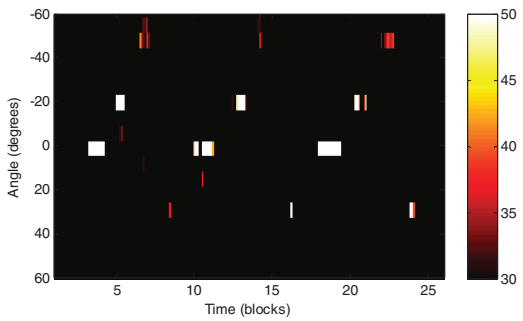

Figure 8: Angle histogram scaled to show weaker signals

As well as data testing reported earlier, the performance of our eight element array was compared with that of Microsoft Kinect that has a built-in four-microphone array. The two arrays were placed in turn, 1.7 meters away facing the centre position of four participants sitting equidistant across a table 2.1 meters wide. With only one person talking at a time, direction finding accuracy of 90% was achieved from each array, but the precision of our array in selecting the direction of individuals was greater.

Future work could include optimising the various parameters in the algorithm to optimise its performance in terms of reliability and speed, also more advanced averaging/tracking algorithms for generating the angular pdf functions could be explored including techniques where multiple angular pdfs are maintained, one for each supposed speaker location with some mechanism for associating each data block with a particular angular pdf. Finally this algorithm also provides a good basis for generating the probability distribution of the person's position in the entire room based on the angular probability distribution obtained from several microphone arrays.

4. CONCLUSION

A DOA estimation algorithm for microphone arrays has been introduced that was found to be relatively robust to reverberation by estimating the DOA of the speaker signal independently in each frequency band and using all these DOA measurements to estimate the likelihood of the speaker direction. As reverberation affects all frequency components differently, considering the DOA at all frequencies gives extra robustness against reverberation. The technique has been applied to an eight element array and found to produce relatively reliable angular probability distributions for a number of different speaker directions. Future work could include fine-tuning the algorithm parameters, improving the way the probability functions are averaged, especially in the case of multiple speakers and incorporating the angular probabilities from several microphone arrays together to determine the likelihood function for the speaker location.

5. REFERENCES

[1] J. Chen, J. Benesty, Y. Huang, "Time Delay Estimation in Room Acoustic Environments: An Overview", EURASIP Journal on Applied Signal Processing Vol 2006, pp 1-19.

[2] S. Doclo, M. Moonen, "Robust Adaptive Time Delay Estimation for Speaker Localization in Noisy and Reverberant Acoustic Environments", EURASIP Journal on Applied Signal Processing Vol 2003, pp 1110-1124.

[3] Y.T. Chan and K.C. Ho "A Simple and Efficient Estimator for Hyperbolic Location", IEEE Transactions on Signal Processing, 42(8), August 1994 pp. 1905-1915.

[4] P. Aarabi, "The Fusion of Distributed Microphone Arrays for Sound Localization", EURASIP Journal on Applied Signal Processing Vol 2003:4, pp 338-347.

[5] Y. Bucris, I. Cohen, M. Doron, "Bayesian Focusing for Coherent Wideband Beamforming", IEEE Transactions on audio, speech, and language processing, vol. 20, no. 4, May 2012.

[6] A. Johansson, N. Grbić, S. Nordholm, "Speaker Localisation Using the Far-Field SRP-PHAT in Conference Telephony", 2002 International Symposium on Intelligent Signal Processing and Communication Systems, Kaohsiung, Taiwan ROC, 2002.