GCC-PHAT based Head Orientation Estimation

Carlos Segura[1][2], Javier Hernando[1]

[1]Universitat Politècnica de Catalunya, Barcelona, Spain
[2]Herta Security, S.L., Barcelona, Spain

Abstract

This work presents a novel two-step algorithm to estimate the orientation of speakers in a smart-room environment equipped with microphone arrays. First the position of the speaker is estimated by the SRP-PHAT algorithm, and the time delay of arrival for each microphone pair with respect to the detected position is computed. In the second step, the value of the cross-correlation at the estimated time delay is used as the fundamental characteristic from where to derive the speaker orientation. The proposed method performs consistently better than other state-of-the-art acoustic techniques with a purposely recorded database and the CLEAR head pose database.

Index Terms: Head pose; speaker orientation; acoustic source localization

1. Introduction

In recent years, significant research efforts have been focused on developing human-computer interfaces in intelligent environments that aim to support human tasks and activities. The knowledge of the position and the orientation of the speakers present in a room constitutes a valuable information allowing a better understanding of user activities and human interactions in those environments, such as the analysis of group dynamics or behaviors, deciding which is the active speaker among the participants or determining who is talking to whom.

The interest in this problem based on multi-channel speech observations is so recent that very few works can be found in the speech related literature. Most of the recent proposals have been done in relation to robust sound localization systems rather than stand-alone orientation estimation algorithms. The main motivation is that taking into account the possibly degrading effects of the head orientation into the localization algorithm may yield to more reliable source positions estimates [1]. This is the case of [2], that based on the SRP-PHAT algorithm, extends the exploration space with an orientation dimension by weighting the contribution of each microphone pair for different possible orientations. A similar approach also based on the SRP-PHAT algorithm can be found in [3], named the Oriented Global Coherence Field (OGCF) method. More recently, a work has been proposed by the same authors of [3] that tackles the problem of talker localization and estimation of head orientation from the perspective of the classification of SRP-PHAT or OGCF spatial likelihood functions [4].

On the other hand, other approaches to head orientation estimation are based on the measurement of the acoustic energy, relying on radiation and propagation characteristics of the

speech signal, given that the speaker position is known beforehand. Usually, these methods need frequency weighting to enhance the directional components of the voice and must account for bad microphone gain calibration and require accurate estimation of propagation attenuation. An example of this approach is presented in [5, 6, 7], employing a large-aperture array consisting of 512 microphones, which completely surrounds the speaker in the horizontal plane. In scenarios, where it is not possible to calibrate the gain of the microphones, the work presented in [8] proposes to normalize the energy at each microphone using the ratio between the energy of high band and low band of frequencies (HLBR). Low frequencies are being radiated by the human head with almost the same intensity in all directions, therefore this value is used as a normalizing value, that partly compensates for different microphone gains and propagation losses. The results obtained by the HLBR measure are compared to those obtained by SRP-PHAT based methods in [9]. A similar approach is followed by the authors in [10], where the HLBR is conducted using cross-correlation features instead of acoustic energy. Recently, the use of artificial neural networks (ANN) has been reported in [11] that uses the time delay estimates (TDEs), source position estimates, distance estimates, and energy features as parameters of the ANN.

In this work, the GCC-PHAT cross-correlation function between pairs of microphones is deeply studied as a basis for speaker orientation estimation. A two step algorithm is proposed for first estimating the position and then the speaker orientation based on cross-correlation orientational cues. The comparison of this orientational characteristic among the microphone pairs distributed across the room provides the most probable orientation of the speaker at each iteration. The proposed method has very low computational demand and has a good performance in adverse scenarios due to the robustness offered by the GCC-PHAT technique, that effectively reduces the impact of reverberation and background noise. Experimental results were conducted over the CLEAR head pose database and a secondary database recorded purposely in the UPC Smart room involving several speakers, positions and orientations.

2. Head radiation pattern

Human voice is mostly radiated from the mouth aperture. Other parts of the body also radiate energy such as the nose or throat in the articulation of certain phones. In addition to that, the radiated sound field is affected by the whole body, and in particular, the head, shoulders and chest refract and absorb part of the sound. Therefore, the human head radiation pattern depends on the physical characteristics of the person. Moreover, during normal speech the radiation pattern is constantly changing, being dependent on the articulated phoneme.

The measurements reported in the literature show that human talkers do not radiate voice sound uniformly in all direc-
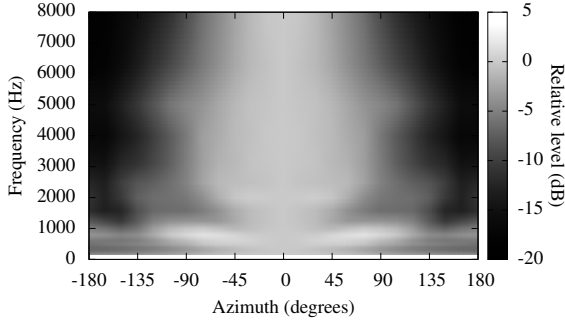
Figure 1: *Mean broadband directivity measures of a human talker in the horizontal plane, based on measures from* [12].

Figure 2: *Mean broadband GCC-PHAT-P measures of a human talker in the horizontal plane.*

tions; more energy is radiated in talker's forward direction than towards the side or the rear direction. The head radiation pattern is also directional with a higher directivity with increasing frequency and mouth aperture size. The phase of the sound field is also affected depending on the angle. The broadband directivity in the horizontal plane as measured in [12] is depicted in Fig. 1, where one can observe a difference of 20 dB depending on the frequency and the angle to the head.

According to these observations, it becomes evident that the quality of the speech captured by a far-field microphone in an indoor environment, in addition to be dependent on the noise and reverberation characteristics of the room, it is also dependent on the relative orientation of the speaker with respect the recording microphone. Consequently, speech applications based on these signals are also affected by head orientation and non uniform speech radiation pattern. Moreover, microphones located at a significant angle from a speaker capture a low-passed version of the signal picked by microphones on the front of the mouth. This has strong consequences for beamforming techniques, whose input signals may include a low-passed version of the speech. The phase change also decreases the performance of beamformers that compute the channel beamforming delays based on the estimated location of the speaker, since the signals are not combined coherently. Compensating for this harming effects, requires the accurate estimation of the time varying radiation pattern of a speaker.

3. Acoustic source localization

3.1. GCC-PHAT algorithm

In a multi-microphone environment, one of the observable clue with positional information more commonly used in acoustic localization algorithms is the time delay of arrival (TDOA) of the signal between microphone pairs. Consider a smart-room provided with a set of N microphones from which we choose M microphone pairs. Let $\mathbf{x}$ denote a $\mathbb{R}^3$ position in space. Then the time delay of arrival $\tau_{\mathbf{x},i,j}$ of an hypothetical acoustic source located at $\mathbf{x}$ between two microphones i, j with position $\mathbf{m}_i$ and $\mathbf{m}_j$ is:

$$\tau_{\mathbf{x},i,j} = \frac{\parallel \mathbf{x} - \mathbf{m}_i \parallel - \parallel \mathbf{x} - \mathbf{m}_j \parallel}{c}, \quad (1)$$

where c is the speed of sound in air.

The cross-correlation function is well-known as a measure of the similarity between signals for any given time displacement and ideally it should exhibit a prominent peak in correspondence to the delay between the pair of signals [13]. A commonly used weighting function in acoustic event localization is the Phase Transform (PHAT), also known in the literature as crosspower-spectrum phase (CSP) technique [14], that
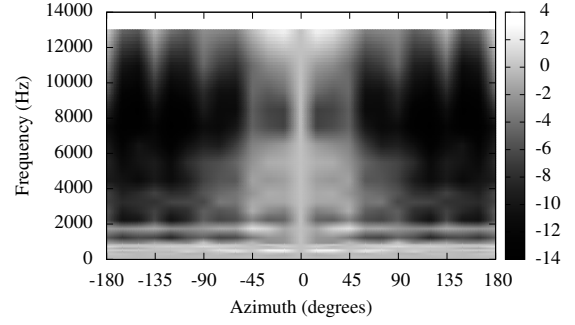
is usually considered useful in reverberant conditions. It can be expressed in terms of the inverse Fourier transform of the estimated CPS ($G_{m_i m_j}(f)$) with the following equation,

$$R_{ij}(\tau) = \int_{-\infty}^{\infty} U(f_1, f_2) \frac{G_{ij}(f)}{|G_{ij}(f)|} e^{j2\pi f \tau} df, \quad (2)$$

and the estimation of the TDOA is as follows:

$$\hat{\tau}_{i,j} = \underset{\tau}{\operatorname{argmax}} R_{ij}(\tau) \quad (3)$$

In practice, the frequency range used to compute $R_{ij}(\tau)$ can be reduced to the speech-band to increase the accuracy [15], employing the rectangular band-pass filter $U(f_1, f_2)$ with unitary value for frequencies $f_1 \leq |f| \leq f_2$, and zero otherwise.

3.2. SRP-PHAT algorithm

The contributions of each microphone pair can be combined to derive a single estimation of the source position. However, in the general case, the availability of multiple TDOA estimations leads to a minimization of an over-determined and non-linear error function. A very efficient approach is the SRP-PHAT or Global Coherence Field introduced in [15], which also performs performs very robustly in reverberant environments.

The basic operation of the SRP-PHAT algorithms consists of exploring the 3-dimensional (3D) space, searching for the maximum of the global contribution of the PHAT-weighted cross-correlations from all the microphone pairs. The 3D room space is quantized into a set of positions with typical separation of 5-10 cm. The theoretical TDOA $\tau_{\mathbf{x},i,j}$ from each exploration position to each microphone pair are precomputed and stored.

The estimated acoustic source location is the position of the quantized space that maximizes the contribution of the cross-correlation of all microphone pairs:

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmax}} \sum_{i,j \in \mathbb{S}} R_{m_i m_j}(\tau_{\mathbf{x},i,j}), \quad (4)$$

where $\mathbb{S}$ is the set of microphone pairs. Then the TDOA for each microphone pair $\tau_{\hat{\mathbf{x}},i,j}$ is estimated using the obtained location.

4. GCC-PHAT based speaker orientation

4.1. GCC-PHAT-P

To investigate the usefulness of the GCC-PHAT Peak value (GCC-PHAT-P) as an orientational clue, first the value of the peak for different relative angles of the speaker to the microphone pairs and its frequency behavior is analyzed. Using the database recorded for this purpose described later, the information regarding the distance of the speaker and its relative orientation to every microphone pair is extracted from the ground
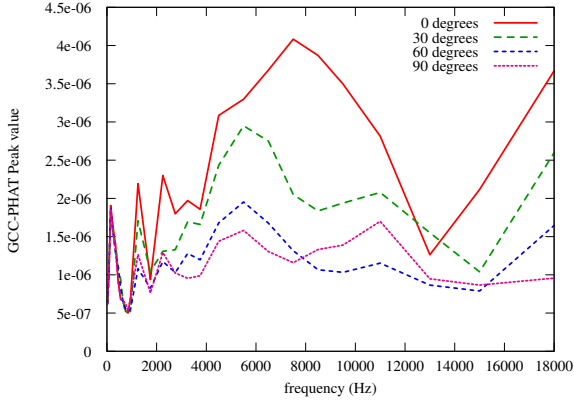
Figure 3: *magnitude of GCC-PHAT-P for 4 orientation angles between speaker and microphone pair 0, 30, 60 and 90 degrees.*

truth annotations. Then the GCC-PHAT cross-correlation function is computed for different frequency ranges using Eq. 2.

The mean value of the band filtered cross-correlation at the delay $\tau_{\hat{\mathbf{x}},i,j}$ corresponding to the location of the speaker is shown in Fig 3 for 4 different orientation angles and frequencies. Experimental results show that signals from microphone pairs placed directly in front of a speaker exhibit a higher coherence over the cross-spectrum than signals from microphones placed outside the main radiation lobe, which are attenuated by the head of the speaker and are more affected by noise and reverberation.

Fig. 3 reveals that the measures of the global peak of the GCC-PHAT function if strongly affected by the orientation for frequencies in the range between 2 kHz and 12kHz. Consequently, the mean values of GCC-PHAT-P for in this frequency band can effectively be used as a cue measure for orientation estimation. The mean broadband GCC-PHAT-P measures of a human talker in the horizontal plane is depicted in Fig. 2. The obtained pattern has a similar shape to the one of speech radiation diagram from Fig. 1.

4.2. Orientation angle estimation

In order to estimate the orientation of a speaker based on the GCC-PHAT-based orientational measures we propose a simple vectorial method similar to that described for the energy-based approach [9]. The proposed technique first needs the position of the active person to be known beforehand or estimated by means of the SRP-PHAT or any other source localization method. In this work, the localization step is performed by the SRP-PHAT algorithm. Then, the vectors $\mathbf{v}_n$ from the speaker to the center of each microphone pair $\mathbf{p}_n$ are computed, adjusting their magnitude $|\mathbf{v}_n|$ to the orientational measure of the microphone pair. The weighted sum of the vectors formed by all the orientational measures of each microphone pair is considered the estimated head direction $\mathbf{v}_{sum}$ as follows:

$$\mathbf{v}_{sum} = \sum_{n=1}^{N} w_n \mathbf{v}_n \qquad (5)$$

The estimated head orientation angle $\hat{o}$ is computed from the $x-$ and $y-$components of $\mathbf{v}_{sum}$:

$$\hat{o} = \angle \mathbf{v}_{sum}, \qquad (6)$$

where $\angle \mathbf{v}_{sum}$ denotes the angle of the projection of $\mathbf{v}_{sum}$ in the xy-plane with the x-axis.

The purpose of the weights w_n is to normalize the magnitude of all microphone pairs enabling them to lie between the range $[-\alpha, (1-\alpha)]$ employing the Min-Max normalization:

$$w_n = \frac{|\mathbf{v}_n| - |\mathbf{v}_{min}|}{|\mathbf{v}_{max}| - |\mathbf{v}_{min}|} - \alpha, \qquad (7)$$

where $\mathbf{v}_{min}$ is the vector with the minimum magnitude from the set of $\mathbf{v}_n$, and $\mathbf{v}_{max}$ is the vector with maximum magnitude.

This weighting models the fact that the microphone pairs with lowest orientational cue value are probably behind the speaker and by giving those pairs a negative value, its resulting vector would help point to the correct direction. In our experiments we obtained good results with $\alpha = 0.3$.

5. Experimental setup

5.1. Database description

The testing database was collected in UPC smart-room. The room dimensions in the x, y, z coordinates are 3966 x 5245 x 4000 mm, and its measured reverberation time is approximately 400 ms. The sensor network used by the speaker localization and head orientation algorithms consists of 6 T-shaped microphone clusters of 4 microphones covering the room.

Collected data consisted of a sequence of sentences uttered by six male speakers at six different positions for eight orientations in steps of about 45 degrees. Eight phonetically rich sentences (of about 3.5 seconds length) were extracted from the wall street journal (WSJ) database [16], one sentence for each orientation. The speakers were split in groups of 2 speakers, and each group had a different sequence of sentences, thus enabling the possibility to analyze the impact of the sentence content on the orientation estimation and also differences among speakers.

The speakers repeated each sentence twice at every location and orientation, following his scheduled sequence of sentences. Signals were sampled at 44.1 kHz. The total database consists of about 32 minutes of audio.

Additionally, the performance of the proposed head orientation estimation algorithm was evaluated with the CLEAR head pose database [17]. It consists of an extract of 3 seminars from the data collected by the CHIL consortium for the CLEAR 2006 evaluation that was labeled for particular head pose evaluation purposes. The seminars were recorded in a non-interactive indoor scenario where a person was giving a talk, for a total of approximately 15 min.

5.2. Evaluation metrics

Metrics and scoring of the systems has been done following the common agreement of the CHIL consortium for head pose evaluation. Three basic metrics are defined:

Pan Mean Average Error (PMAE) [degrees]: the precision of the head orientation angle estimation.

Pan Correct Classification (PCC) [%]: the ability of the system to correctly classify the head position within 8 classes spanning $45°$ each.

Pan Correct Classification within a Range (PCCR) [%]: the ability of the system to correctly classify the head position within 8 classes spanning $45°$ each, allowing a classification error of ± 1 adjacent class.

6. Results

Table 1 shows different results for the proposed GCC-PHAT-P method varying the cut-off frequencies f_1 and f_2 in Eq. 2, eval-

uated with the UPC database. Best performance was obtained with the frequency range 100 - 6000 Hz. However, for the sake of computational complexity, in subsequent experiments the same frequency range as the localization algorithm was employed, which is from 100 to 8000 Hz.

Table 1: *PMAE for different results for the proposed GCC-PHAT-P method varying the cut-off frequencies f_1 and f_2.*

$f_2\backslash f_1$	100 Hz	500 Hz	1 kHz	2 kHz	3 kHz
6 kHz	**11.21°**	12.68°	12.50°	12.44°	14.54°
8 kHz	11.80°	14.46°	13.39°	14.53°	15.86°
10 kHz	15.47°	17.62°	17.70°	18.02°	19.02°
12 kHz	15.52°	17.12°	17.22°	17.71°	18.49°

The results obtained the proposed method are compared with those from two alternative methods based on SRP-PHAT, described in [9], and also compared with best results published in the CLEAR Evaluation [18] involving video algorithms.

Table 2: *Head pose orientation results for the 5 methods evaluated with the UPC database.*

Method	PMAE	PCC	PCCR
SRPPHAT-J	34.70°	37.75%	84.31%
SRPPHAT-F	35.58°	33.46%	83.84%
HLBR-B	57.83°	26.01%	60.48%
HLBR-V	58.72°	25.28%	59.03%
GCC-PHAT-P	11.80°	76.87%	99.46%

Table 2 and table 3 summarize the averaged results obtained by the proposed method using both the new UPC database and the CLEAR head pose database. The new SRP-PHAT-P technique for estimating the orientation of a speaker exhibits better overall performance than the other state of the art acoustic methods and achieves a very similar performance to video algorithms in the CLEAR head pose database.

Table 3: *Head pose orientation results evaluated with CLEAR head pose database.*

Method	PMAE	PCC	PCCR
SRPPHAT-J	44.68°	37.32%	73.38%
SRPPHAT-F	44.23°	37.71%	73.89%
Best video CLEAR 2006	33.56°	44.8%	86.6%
GCC-PHAT-P	32.52°	48.31%	85.44%

7. Conclusions

The cross-correlation function between pairs of microphones is studied as a basis for speaker orientation estimation, which is stated as having a strong dependence with the speaker orientation and frequency. A two step algorithm is proposed for first estimating the position and then the speaker orientation based on cross-correlation orientational cues. The proposed method performs consistently better than the other audio techniques with both databases, obtaining promising results in terms of accuracy and robustness of the estimation very similar to those obtained with video algorithms.

8. References

[1] A. Abad, D. Macho, C. Segura, J. Hernando and C. Nadeu, "Effect of head orientation on the speaker localization performance in smart-room environment," in *Proc. INTERSPEECH, Lisbon*, 2005.

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

[3] A. Brutti, M. Omologo, and P. Svaizer, "Oriented global coherence field for the estimation of the head orientation in smart rooms equipped with distributed microphone arrays," in *Proc. INTER-SPEECH*, 2005.

[4] A. Brutti, M. Omologo, P. Svaizer, and C. Zieger, "Classification of acoustic maps to determine speaker position and orientation from a distributed microphone network," in *Proc. ICASSP*, vol. 4, Honolulu, Hawaii, USA, April 2007, pp. 493–496.

[5] A. Levi and H. Silverman, "A new algorithm for the estimation of talker azimuthal orientation using a large aperture microphone array," in *Multimedia and Expo, 2008 IEEE International Conference on*, April 2008, pp. 565 –568.

[6] J. Sachar and H. Silverman, "A baseline algorithm for estimating talker orientation using acoustical data from a large-aperture microphone array," in *Proc. ICASSP*, vol. 4, 2004, pp. 65–68.

[7] A. Levi and H. Silverman, "A robust method to extract talker azimuth orientation using a large-aperture microphone array," *IEEE Trans. on ASLP*, vol. 18, no. 2, pp. 277–285, February 2010.

[8] C. Segura, C. Canton-Ferrer, A. Abad, J. Casas, and J. Hernando, "Multimodal head orientation towards attention tracking in smartrooms," in *Proc. ICASSP*, vol. 2, April 2007, pp. 681–684.

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

[10] C. Segura, A. Abad, J. Hernando, and C. Nadeu, "Speaker Orientation Estimation based on GCC-PHAT and HLBR hybridation," in *Proceedings International Conference on Spoken Language Processing (ICSLP'08)*, Brisbane, Australia, 2008, pp. 1325–1328.

[11] A. Nakano, S. Nakagawa, and K. Yamamoto, "Automatic estimation of position and orientation of an acoustic source by a microphone array network," *JASA*, vol. 126, pp. 3084–3094, 2009.

[12] W. T. Chu and A. Warnock, "Detailed directivity of sound fields around human talkers," Institute for Research in Construction, Tech. Rep., 2002.

[13] P. Svaizer, M. Matassoni, and M. Omologo, "Acoustic source location in a three-dimensional space using crosspower spectrum phase," in *Proc. ICASSP*, vol. 1, Munich, Bavaria, Germany, April 1997.

[14] M. Omologo and P. Svaizer, "Acoustic event localization using a crosspower-spectrum phase based technique," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Adelaide, South Australia, Australia, April 1994.

[15] J. DiBiase, H. Silverman, and M. Brandstein, *Microphone Arrays. Robust Localization in Reverberant Rooms*. Springer, 2001.

[16] D. Paul and J. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proceedings of the workshop on Speech and Natural Language*. Morristown, NJ, USA,: Association for Computational Linguistics, 1992, pp. 357–362.

[17] R. Stiefelhagen and J. Garofolo, "Multimodal technologies for perception of humans. first international evaluation workshop on classification of events, activities and relationships, clear 2006," in *LNCS*, vol. 4122, 2007.

[18] "CLEAR - Classification of Events, Activities and Relationships Evaluation and Workshop," http://www.clear-evaluation.org, 2007.