interspeech 2008 incorporating SST 08

Speaker Orientation Estimation based on Hybridation of GCC-PHAT and HLBR

Carlos Segura[1], *Alberto Abad*[1,2], *Javier Hernando*[1] *and Climent Nadeu*[1]

[1] TALP Research Center
Universitat Politècnica de Catalunya (UPC), Barcelona, Spain
[2] L^2F - Spoken Language Systems Lab
INESC-ID / IST, Lisboa, Portugal
{csegura,alberto,javier,climent}@gps.tsc.upc.edu

Abstract

This paper presents a novel approach to speaker orientation estimation in a SmartRoom environment equipped with multiple microphones. The ratio between the high and low band energies (HLBR) received at each microphone has been shown in our previous work to be a potentially approach to estimate the direction of the voice produced by a speaker. In this work, for each microphone pair, a smoothed CPS phase is obtained by a proper windowing of the main peak of the cross-correlation sequence estimated with the GCC-PHAT method, and a HLBR is computed from the processed CPS. The proposed method keeps the computational simplicity of the HLBR algorithm while adding the robustness offered by the GCC-PHAT technique. Experimental preliminary results were conducted over a database recorded purposely in the UPC Smart room, and over the CLEAR head pose database. The proposed method performs consistently better than other state-of-the-art techniques with both databases.

Index Terms: Head orientation, Speaker orientation, Speaker localization,

1. Introduction

In recent years, significant research efforts have been devoted to the development of human-computer interfaces in intelligent environments aiming at supporting humans in various tasks and situations. The head orientation of a person provides important clues in order to give a better service in such scenarios. This knowledge allows a better understanding of what users do or what they refer to. Moreover, the development of enhanced microphone network management strategies for microphone selection based on both speaker position and orientation information would permit the improvement of speech technologies that are commonly deployed in smart-rooms

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 in [2] where the SRP-PHAT algorithm [3] is extended by incorporating the orientation as a new search parameter, and weighting the contribution of each microphone pair according to it. A similar approach named Oriented Global Coherence Field (OGCF) has been proposed in [4], which is also a variation of the SRP-PHAT algorithm. More recently, a work has been proposed by the same authors of [4] that tackles the problem of talker localization and estimation of head orientation from the perspective of the classification of SRP-PHAT or OGCF audiomaps [5].

On the other hand, other approaches are based on the head radiation pattern and propagation characteristics of the speech signal, and assume that the speaker position is known beforehand. They rely either on the acoustic energy received at each microphone [6], or in the High/Low Band Ratio (HLBR) measure proposed by the authors in [7] and compared to SRP-PHAT based methods in [8]..

This work proposes to use the frequency contribution of microphone pairs to the main peak in the global SRP-PHAT function as the fundamental information from which to derive the head orientation estimation. Experimental results reported in the paper 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. A normalization step similar to the HLBR, ensures the reliability of the performance for different distances between microphone pairs. The proposed method keeps the computational simplicity of the HLBR algorithm while adding the robustness offered by the GCC-PHAT algorithm.

Experimental results were conducted over a database recorded purposely in the UPC Smart room involving several speakers, positions and orientations. They are reported and compared with those from two alternative methods based on SRP-PHAT and HLBR, described in [8], proving the effectiveness of the proposed approach. The three methods are also evaluated with the CLEAR [9] head pose database. The proposed method performs consistently better than the other techniques with both databases, obtaining promising results in terms of accuracy and robustness of the estimation.

2. Head Orientation Estimation

The measurements reported in [10] show that human talkers do not radiate voice sound uniformly in all directions; more energy is radiated in talker's forward direction than towards the side

This work has been funded by the the Spanish project SAPIRE (TEC2007-65470)

or the rear direction. Additionally, the radiation pattern is frequency dependent being more directive for high speech frequencies. According to these observations, it becomes evident that the quality of the speech captured by a far-field microphone in an indoor environment, in adittion 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, and consequently, speech applications based on these signals are also affected by head orientation and non uniform speech radiation pattern.

The head orientation estimation method presented in this paper is intrinsically tied to the acoustic localization, since both are based on the coherence between the microphone of each pair. Therefore a two-step algorithm is proposed. First the position of the speaker is estimated using the SRP-PHAT algorithm, and the Time Delay Of Arrival (TDOA) for each microphone pair with respect to the detected position is computed. Then the cross-correlation function of each microphone pair nearby the estimated TDOAs is analysed to gather information about the head orientation.

2.1. Acoustic Localization

The SRP-PHAT algorithm [3] tackles the task of acoustic localization in a robust and efficient way. The basic operation of the SRP-PHAT algorithms consists of exploring the 3D space, searching for the maximum of the global contribution of the PHAT-weighted cross-correlations from all the microphone pairs. The SRP-PHAT algorithm performs very robustly due to the PHAT weighting, and actually, it has turned out in one of the most successful state-of-the-art approaches to microphone array sound localization.

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 $TDOA_{i,j}$ of an hypothetic acoustic source located at $\mathbf{x}$ between two microphones i, j with position $\mathbf{m}_i$ and $\mathbf{m}_j$ is:

$$TDOA_{i,j} = \frac{\parallel \mathbf{x} - \mathbf{m}_i \parallel - \parallel \mathbf{x} - \mathbf{m}_j \parallel}{s}, \qquad (1)$$

where s is the speed of sound.

The 3D room space is then quantized into a set of positions with typical separation of 5-10cm. The theoretical TDOA $\tau_{\mathbf{x},i,j}$ from each exploration position to each microphone pair are precalculated and stored.

Generalized cross-correlations (GCC) [11] of each microphone pair are estimated for each analysis frame with the PHAT weighting. It can be expressed in terms of the inverse Fourier transform of the estimated cross-power spectrum (CPS) ($G_{m_1 m_2}(f)$) as follows,

$$R_{m_i m_j}(\tau) = \int_{-\infty}^{\infty} \frac{G_{m_i m_j}(f)}{|G_{m_i m_j}(f)|} e^{j2\pi f \tau} df, \qquad (2)$$

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

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmax}} \sum_{i,j \in \mathbb{S}} R_{m_i m_j}(\tau_{\mathbf{x},i,j}), \qquad (3)$$

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.

2.2. GCC-PHAT Peak Analysis based Orientation Estimation

The contribution of every microphone pair to the main peak of the audio map in the SRP-PHAT algorithm depends on the influence of noise, reflexions and reverberations on the signal received by the microphones. If we take into account the frequency dependence of the head radiation pattern and the fact that the reverberated sound energy field is position independent, microphones placed in front of the speaker will have a higher SRR than microphones placed outside the main radiation lobe. The consequence is that the contribution to the peak at the hypothetical TDOAs in the cross-correlation from a microphone pair depends on the orientation of the speaker with respect to the microphone pair as showed in our previous work [1]. In this work we propose a novel method for extracting the speaker orientation information from the analysis of the CPS at the estimated TDOAs.

Referring back to the GCC-PHAT function (Eq. (2)), the information about the delay between two microphones i, j relies on the phase $\phi(f)$ of the cross-power spectrum $G_{m_i m_j}(f)$, since its magnitude is equalled to 1 for all frequencies in the computation of $R_{m_i m_j}(\tau)$. Ideally, $\phi(f)$ would be a line with a slope proportional to $\tau_{\mathbf{x},i,j}$. In Fig.1 we can observe the ideal CPS phase and the measured CPS phase in a real case of a person speaking in front of a microphone pair. As illustrated in this example, the phase scattering is strong and unevenly distributed along the frequency axis.

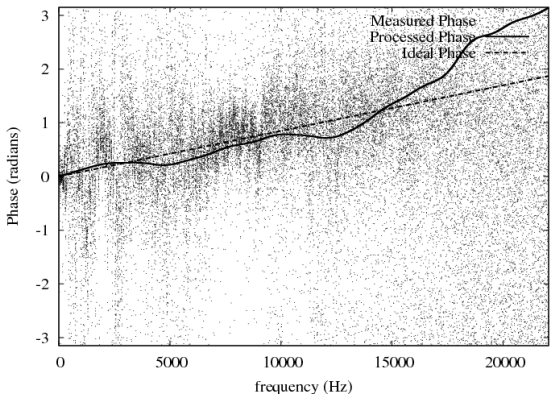

Figure 1: *Real, processed and ideal phase of the CPS for a person directly speaking to a pair of microphones.*

In this work we propose a way to reduce the CSP phase scattering and obtain a processed phase that is closer to the ideal straight line, so indirectly attenuating the effects of noises and reverberation. This could be done by some kind of smoothing in the frequency domain, but it is more efficient computationally to reduce the phase scattering in the time domain by windowing the main peak at $\tau_{\hat{\mathbf{x}},i,j}$ in the cross-correlation with a short window $w(t)$, and then computing the Fourier transform back to the frequency domain:

$$O(f) = \mathcal{F}\left(w(t - \tau_{\hat{\mathbf{x}},i,j}) \cdot \mathcal{F}^{-1}\left(\frac{G_{m_i m_j}(f)}{|G_{m_i m_j}(f)|} \right) \right), \qquad (4)$$

where $\tau_{\hat{\mathbf{x}},i,j}$ is the TDOA for the microphones i and j to an active speaker estimated in the localization step.

Indeed, the magnitude of the new CPS will have changed and will no longer be equal to 1 for all frequencies. The new

CPS phase resulting from the peak-centered time windowing operation is also plotted in Fig.1, where we can see that it is very similar to the ideal one, at least for other bands than the highest one. The magnitude of $O(f)$ represents the frequency contribution to the main peak in the cross-correlation. Using the database described later, the mean and variance of $O(f)$ for different relative angles of the speaker to the microphone pairs and are computed from more than 10^5 samples as shown in Fig. 2.

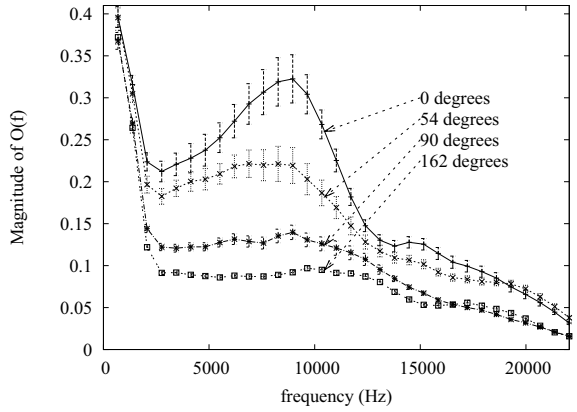

Figure 2: *Magnitude of $O(f)$ for 4 orientation angles between speaker and microphone pair 0, 54, 90 and 162 degrees.*

From Fig.2 it is clear that the magnitude measure of $O(f)$ is nearly independent on the orientation for a low frequency band range (up to 1.5kHz), while for high band frequencies (from 2.5kHz to 12 kHz) is strongly affected by the orientation. Consequently, the mean value of the magnitude of $O(f)$ for the high band frequencies itself could be used as a cue measure for orientation estimation. However, we propose to compute the ratio GCCPHAT-HLBR between the mean value of the high frequency band and the mean value of the low band, similarly as it was done in our previous work [7]. By means of this normalization, the effect of the different distances between different microphone pairs and the speaker in the cross-correlation function computation are partially cancelled. The mean and variance of the estimates of the GCCPHAT-HLBR measures in terms of the angle is depicted in Fig. 3, where similarties to speaker radiation pattern can be clearly observed.

In order to estimate the orientation based on the GCCPHAT-HLBR measures we propose a simple vectorial method like in [8]. First, the vectors $\mathbf{v}_n$ from the speaker to the center of each microphone pair $\mathbf{p}_n$ with module $|\mathbf{v}_n|$ equal to the GCCPHAT-HLBR measure of the microphone pair are computed. Then, the angle of the sum vector of all the GCCPHAT-HLBR of each microphone pair is considered the estimated head orientation $\hat{o}$:

$$\mathbf{v}_{sum} = \sum_{n=1}^{N} \mathbf{v}_n \qquad \hat{o} = \angle \mathbf{v}_{sum} \tag{5}$$

3. Database description and evaluation metrics

In order to carry out a detailed experimental investigation, aimed at verifying the performance and robustness of the proposed method in comparison with other state-of-the-art approaches, a database was recorded at the UPC's Smart room using the existent sensor set up and involving several speakers at several positions and orientations.

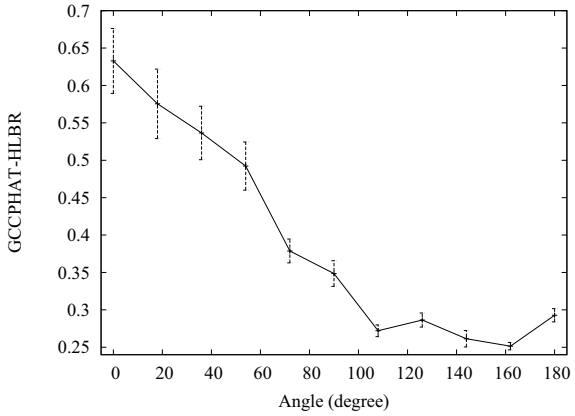

Figure 3: *Experimental mean and variance of the GCCPHAT-HLBR value for different angles.*

Additionally, the performance of the proposed head orientation estimation algorithm was evaluated with the CLEAR head pose database [12]. It consists of an extract of 3 seminars from the data collected by the CHIL consortium for the CLEAR 2006 evaluation that was labelled 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.

3.1. UPC Smart-Room

The testing database was collected in UPC smart-room. It is a room equipped with several multimodal sensors such as microphone arrays, table-top microphones, and fixed or pantilt-zoom video cameras. The room dimensions in the x, y, z coordinates are 3966 x 5245 x $4000mm$, and its measured reverberation time is approximately $400msec$.

The sensor network used by the speaker localization and head orientation algorithms consists of 6 T-shaped microphone clusters of 4 microphones (Shure Microflex).

3.2. Database recording

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 WSJ database, 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 posibility to analyse 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.

3.3. 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.

4. Experimental results and discussion

4.1. Results

Table 1 and table 2 summarize the averaged results obtained by the proposed method in comparison to the methods described in our previous work [8] using both the new UPC database and the CHIL head pose database. The new GCCPHAT-HLBR technique exhibits better overall performance than all previous methods.

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%
GCCPHAT-HLBR	29.07°	50.54%	84.49%

Table 1: *Head pose orientation results for the 5 methods evaluated with the UPC database.*

Figure 4 depicts the PMAE scores for every head pose angle obtained by the GCCPHAT-HLBR and SRPPHAT-J methods with the UPC database. From the graph we can conclude that the proposed GCCPHAT-HLBR method has a generalized better performance across angles, so it shows a high robustness and independence from the environment conditions like the proximity of the speaker to the walls, reverberation and noises.

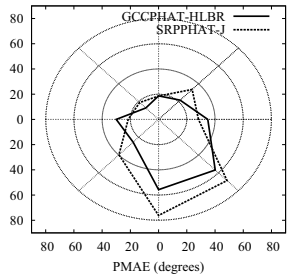

Figure 4: *PMAE scores for SRPPHAT-J and GCCPHAT-HLBR methods with the UPC database.*

Finally the results obtained with the CLEAR database are showed in Table 2. Again, the proposed GCCPHAT-HLBR method performs consistantly better than the other four previously-reported techniques.

Method	PMAE	PCC	PCCR
SRPPHAT-J	44.68°	37.32%	73.38%
SRPPHAT-F	44.23°	37.71%	73.89%
HLBR-B	52.92°	29.85%	67.99%
HLBR-V	50.98°	32.61%	68.94%
GCCPHAT-HLBR	36.52°	37.25%	86.13%

Table 2: *Head pose orientation results of the five methods evaluated with CHIL head pose database.*

5. Conclusions

This paper presents a novel approach to speaker orientation estimation in a SmartRoom environment equipped with multiple microphones based on the hybridation of the High/Low Band Ratio (HLBR) and GCC-PHAT algorithms. In preliminary experiments, the proposed method performs consistently better than other state-of-the-art techniques with two databases, obtaining promising results.

6. 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*, 2005.

[2] B. Mungamuru and P. Aarabi, "Enhanced sound localization," in *IEEE Transactions on Systems, Man and Cybernetics*, 2004, vol. 34(3), pp. 1526–1540.

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

[4] 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. Interspeech*, 2005.

[5] 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 *ICASSP*, 2007, vol. 4, pp. 493–496.

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

[7] C. Segura, C. Canton-Ferrer, A. Abad, J.R. Casas, and J. Hernando, "Multimodal head orientation towards attention tracking in smart rooms," in *ICASSP*, 2007.

[8] A. Abad, C. Segura, C. Nadeu, and J. Hernando, "Audio-based approaches to head orientation estimation in a smart-room," in *Interspeech*, 2007.

[9] "CLEAR Evaluation Campaign," http://www.clear-evaluation.org.

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

[11] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," in *IEEE Trans. on Speech and Audio Processing*, 1997.

[12] 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 *Lecture Notes in Computer Science*, 2007, vol. 4122.