SOURCE LOCALIZATION AND SEPARATION USING RANDOM SAMPLE CONSENSUS WITH PHASE CUES

Łukasz Litwic *Philip JB Jackson*

Centre for Vision, Speech and Signal Processing, University of Surrey, UK
Guildford, Surrey, GU2 7XH
{l.litwic,p.jackson}@surrey.ac.uk

ABSTRACT

In this paper we present a system for localization and separation of multiple speech sources using phase cues. The novelty of this method is the use of Random Sample Consensus (RANSAC) approach to find consistency of interaural phase differences (IPDs) across the whole frequency range. This approach is inherently free from phase ambiguity problems and enables all phase data to contribute to localization. Another property of RANSAC is its robustness against outliers which enables multiple source localization with phase data contaminated by reverberation noise. Results of RANSAC based localization are fed into a mixture model to generate time-frequency binary masks for separation. System performance is compared against other well known methods and shows similar or improved performance in reverberant conditions.

1. INTRODUCTION

Separation of multiple acoustic sources in reverberant environments is a challenging task. The idea for using time-frequency masks for separating sources from underdetermined stereo mixtures has led to development of numerous algorithms based on interchannel or binaural cues [1],[2], namely, interaural phase differences (IPDs) and interaural level differences (ILDs).

The presented algorithm uses IPDs for this task. The main difficulty when dealing with IPDs is the so-called phase ambiguity problem. This is caused by wrapping of phase, and means that a single IPD value corresponds to multiple interaural time differences (ITDs), hence the ambiguity. The scale of the phenomenon depends on a distance between microphones as well as on a position of a source. Several methods have been proposed to alleviate the phase ambiguity problem. The most straightforward one, proposed in DUET [2], used low frequency range where phase ambiguity problem was not present. This was somewhat a crude solution to the problem, nevertheless, consistent with the Duplex Theory which says that IPDs contribute to localization primarily at lower frequencies. Alternative solution was proposed in [3] which used ILDs to resolve the ambiguity based on a relationship found between ILDs and spatial location. In MESSSL [1], instead of solving IPD to ITD ambiguity, hypothetical ITD values were tested against a given IPD value for goodness of fit. Finally, in Least-Squares based time delay estimation, an incremental phase unwrapping based on prediction from lower to higher frequency band was used [4].

In the first part of this paper we present a localization method which uses an representation called Cross-Phasogram (CPG) which includes all permutations of IPDs that would be allowed in a given microphone pair. This idea is similar to the time-delay graph in [5]. In addition, CPG allows for aggregation of data across multiple time frames to increase robustness against noise and reverberation. Hypothetical ITD models are tested against CPG in search for a global consistency of IPDs across the whole frequency range. Since the fitting is done against CPG data, it is inherently free of phase ambiguity problem. Model selection, search, and costings are based on Random Sample Consensus (RANSAC) approach. Once the location of sources is found, a mixture model is used to calculate binary separation masks. The algorithm is compared against other well known methods: DUET [2] and two variants of GCC-PHAT algorithm adopted for multi-source localization: PHAT-Histogram and PHAT-Sum [6]. The evaluation is done using multi speaker scenario in reverberant conditions.

2. ALGORITHM

For two signals from a microphone pair IPDs are calculated as follows:

$$\phi(k,t) = \arg(X_l(k,t) \cdot X_r^*(k,t)) \tag{1}$$

where $X_l(k,t)$ and $X_r(k,t)$ are time frequency representations of left and right channels. k and t are indices for frequency and time respectively. Relationship of time delay between the two channels τ and ϕ can be expressed as:

$$\phi(k,t) = \Big[\tau(t)\omega(k) + \epsilon\Big]_{-\pi}^{\pi} \tag{2}$$

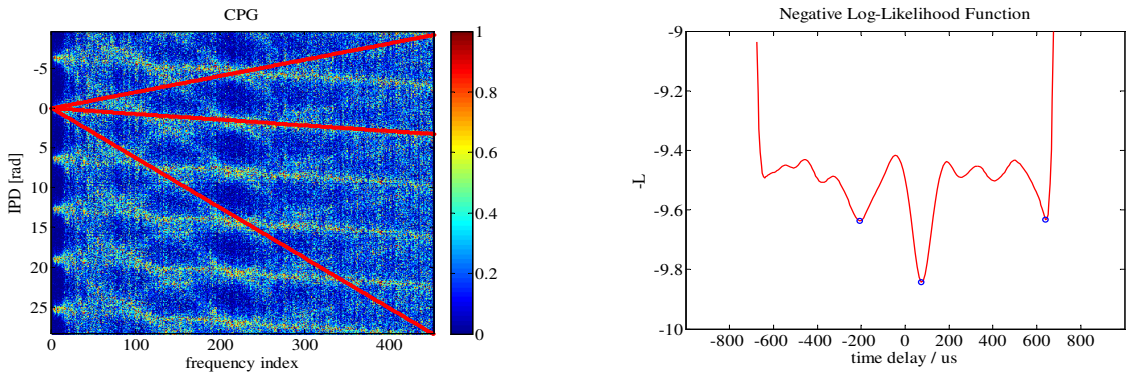
where $\phi \in [-\pi, \pi)$, $\omega(k)$ is an angular frequency and ϵ is error term. The error term represents diversions from the true time delay due to ambient noise, reverberation or interference from other sources.

2.1. Cross-Phasogram

The idea for using Cross-Phasogram (CPG) is to aggregate all IPD data in order to be able to perform search to find best parameters for model (2). Unlike in other methods where phase data is translated to a time delay histogram [2], Cross-Phasogram is built from separate histograms for each frequency band:

$$CPG(\phi_b, k) = \sum_t H(\phi(k,t), \phi_b) W(k,t) \tag{3}$$

where $\phi_b \in (-\pi, \pi)$ defines histogram bins, k is a frequency index, H is a histogram increment function and W is a weighting function defined for each time-frequency atom. For $W(k,t) = 1$ for all k, t CPG is equivalent to an estimate of probability density function (pdf) of ϕ. On the other hand, W could be used to emphasize contribution from those atoms, which have a reliable localization



(a) (b)

Figure 1: Plot a) shows an example of Cross-Phasogram representation with detected time-delay models (red line). The example is taken for a three speaker mixture with room reverberation $RT_{60} = 0.89s$. Sources were located at: 80°, 10° and -25°. Plot b) shows an example of negative log-likelihood function calculated during MLESAC search. Main modes are selected as the best matching time delay estimates. The log-likelihood is calculated for the case presented at plot a).

information to increase robustness of time delay estimates. In order to deal with data outside of $(-\pi, \pi)$ range, CPG could be treated as a pdf of a periodic random variable ϕ as in [1]. Alternatively CPG can be explicitly unwrapped by applying an iterative duplication with $\pm 2\pi$ period. This may make the analysis of IPDs more approachable as well as fits more nicely with RANSAC estimation of (2) presented in the next paragraph. Figure 1a shows an example of unwrapped CPG.

2.2. RANSAC based Model Selection

The linear relationship of frequency and IPDs through time delay τ (2) has been previously utilized in Least-Squares (LS) based time-delay estimation methods [4]. In this paper we propose to use RANSAC approach for this task. RANSAC was introduced by Fishler and Bolles in [7], and is a powerful model estimator widely used in the field of computer vision. Its main strength is ability to find a good fit for a model, even for highly contaminated data. Unlike in LS estimation, where a whole data set is included in model estimation, though different weights can be applied across the data, RANSAC uses minimal data sample sufficient for model estimation. For instance, line model estimation in 2D space requires only 2 points. A single RANSAC iteration consists of two steps: hypothesis model selection based on minimal data set and test for a goodness of fit for the selected model.

2.2.1. *Hypothesis Model Selection*

In the original RANSAC algorithm a hyphotesis model is calculated by randomly selected samples from data set. This process is iteratively run until a probability of finding better match is lower than a predefined threshold. There are a few adaptations that can be made to this process when running on Cross-Phasogram.

Firstly, since every model needs to go through $(0, 0)$ point, sample selection reduces to one point only. Secondly, instead of random selection of points, especially for noisy conditions, hypothesis models can be uniformly selected over a whole range of possible values of τ. The drawback of the latter is that in order to provide good resolution of τ estimates the number of iterations may be excessive when compared to the random selection process.

2.2.2. *Cost Function*

In [7] sample data was partitioned into two sets: data which supports a model called *inliers* and data not relevant for the model called *outliers*. Partitioning was based on a perpendicular distance e measured from each datum to a model. A support for the model was measured by calculating cardinality of a set of inliers. More robust behavior was obtained with the use of a following error term to cost each datum:

$$\rho(e^2) = \begin{cases} e^2 & e^2 < D_t^2 \\ D_t^2 & e^2 \geq D_t^2 \end{cases} \tag{4}$$

where, e^2 is squared perpendicular distance to the model and D_t^2 is a threshold. Best model was found by minimizing the cost function over whole data $C = \sum_i \rho(e_i^2)$. Similar cost function was adopted in Least-Squares based solution [4]. The problem with (4) is that it relies on the threshold D_t to be set correctly. If it is too large noisy data will bias the estimate. On the other hand, if it is too low relevant data may be excluded from estimate calculation. An alternative to the binary assignment of data, whether inliers or outliers, is to express the probability of error as a mixture of a zero-mean Gaussian and uniform distributions. This was proposed in MLESAC (Maximum Likelihood Estimation Sample Consensus) addition to RANSAC [8]:

$$P(e^2) = \gamma \mathcal{N}(e^2, \sigma^2) + (1 - \gamma)\frac{1}{\upsilon} \tag{5}$$

Where, γ is a mixing parameter, σ^2 is a parameter of a Gaussian and υ is a constant. In [8] all parameters of mixture model (5) $\Theta = \{\sigma, \gamma\}$ are found through Expectation-Maximization (EM) algorithm. In our algorithm, instead of associating γ with a model, we associate γ with each datum in CPG. Therefore each datum in CPG can be assigned a probability of supporting a given model based on distance to the model as well as on the evidence from IPDs accumulated in CPG value for this datum. Therefore best model is found by minimizing the negative log-likelihood over whole data:

$$-L = -\sum_i log\Big(\gamma(i)\mathcal{N}(e^2, \sigma^2) + (1 - \gamma(i))\frac{1}{\upsilon}\Big) \tag{6}$$

where $\gamma(i) = CPG(i)$. Example of negative log-likelihood function is shown in Fig. 1b.

2.3. Estimation of Time-Frequency Separation Mask

Once sources are localized and parameters of best time-delay models are found, a Gaussian Mixture Model can be applied to find time-frequency separation masks. For each TF atom value of IPD

$$p(\phi(k,t)|\mu,\sigma^2) = \sum_{n=1}^{N} w_n \cdot \mathcal{N}(\mu(n),\sigma^2(n)) \tag{7}$$

where N is a number of sources present in a mixture and w_n is a mixing coefficient. From this a binary separation mask is calculated from:

$$\mathcal{M}_i(k,t) = \begin{cases} 1 & r(i) > r(j) \quad \forall j \neq i \\ 0 & otherwise \end{cases} \tag{8}$$

where r_i is the posterior probability for source i:

$$r(i) = \frac{w_i p(\phi(k,t)|\mu_i,\sigma^2)}{\sum_{j=1}^{N} w_j p(\phi(k,t)|\mu_j,\sigma^2)} \tag{9}$$

3. EXPERIMENT

3.1. Experimental Settings

For the experiment we created a data set of audio mixtures where we varied the following parameters: number of speakers, spatial locations and room acoustics. Number of speakers was varied between two and four. For source signals, we used the GRID corpus [9]. We chose ten, five seconds long utterances for each of the speakers. Five of the speakers were male and five were female. Utterances were selected to maximize the phonetic coverage. The endpoints were aligned so there were no silence gaps other than those naturally occurring within utterances. For impulse responses, we used signals measured at Institute of Sound Recording at University of Surrey [10]. The impulse responses were captured in four different environments with the following RT_{60} times: $0.32s$, $0.47s$, $0.68s$ and $0.89s$. In addition an anechoic set was created. For all environments azimuths from -90° to 90° in a 5° step were recorded. For two and three source mixtures one source was always set randomly within -15° to 15° position while the other ones were set randomly outside of this region. In case of three source mixture one source was located at negative azimuths and the other at positive ones. For four source mixtures the median plane was partitioned into 4 overlapping regions in which sources where given random location. In total we run 450 iterations for each algorithm.

3.2. Algorithmic Settings

We used two well known algorithms for comparison with our algorithm: DUET and two variants of PHAT [6]. For the former one, our DUET implementation was based on [2]. For the latter, the time-delay estimate was found using the PHAT-Histogram algorithm for each time frame and results aggregated in a histogram. Most significant modes of the histogram were selected as time-delay estimates for each source. An alternative to this was PHAT-Sum, where a sum of cross-correlation terms for all time frames was done before the search for most significant peaks. For both PHAT-Histogram and PHAT-Sum localization, results were fed into the separation model(7)for time-frequency mask calculation. For our algorithm, we used MLESAC probability (5) calculated for each point in CPG for an evaluated model. Again we run two variants: MLESAC-Fixed ran a deterministic search over every azimuth location between -90 degrees and 90 degress with increment of 1 degree; MLESAC-Random tested models that were chosen in a random fashion. In the latter case, we allowed maximum 300 iterations. Minima from the negative log-likelihood function (6) were taken as time-delay estimates and fed into the separation model (7). In the separation model, we used the same variance as for all algorithms. In this way, any difference in separation performance could be directly related to localization performance. Finally, we also measured separation results using a Ground Truth binary mask. For all methods, the number of sources was assumed to be known. The aggregation time for all methods was 5s , i.e., the whole duration of each mixture. The algorithms are denoted as DUET, PHAT-H (Histogram), PHAT-S (Sum),MLESAC-F (Fixed) and MLESAC-R (Random) respectively.

3.3. Performance Analysis

We used two metrics to measure localization performance: Root Mean Square (RMS) Error to measure precision of localization and Gross Error as a percentage of outliers (anomalous estimates). The threshold for an estimate to become an outlier was 5°. This value was determined by the smallest spatial distance between the sources in our mixtures which was 10°. We used Source to Distortion Ratio (SDR) metric defined in [11] to measure separation performance of the algorithms.

4. RESULTS

Results are presented in Fig.2 and show improvement given by both MLESAC methods over DUET and PHAT-H. Separation performance of the MLESAC methods and PHAT-S is very close. Total SDR result were: MLESAC-R 3.43dB, PHAT-S 3.37dB, MLESAC-F 3.32dB, PHAT-H 2.72dB and DUET 1.59dB. For reference, the Ground Truth separation result was 7.2dB. Difference in performance between PHAT-H and PHAT-S, as well as the MLESAC methods, stems from the fact that PHAT-H keeps only one time-delay estimate per time frame for accumulation into its histogram. Comparison of MLESAC-R and MLESAC-F shows that evaluation of time-delay models aligned with points in CPG gives improvement over preselected time-delay models. Comparison of PHAT-S and MLESAC-R shows that a minor improvement in localization performance is not necessarily translated to an improvement in separation performance for a reverberant environment (see $RT_{60} = 0.47$s for example). On the other hand, for the anechoic case, a similar scale of improvement in localization performance yielded a small yet noticeable improvement in separation performance.

5. CONCLUSIONS

We have presented two variants of an algorithm for localization and separation of multiple sources in reverberant conditions. The proposed algorithm uses a RANSAC approach to find consistency in phase data aggregated in the Cross-Phasogram (CPG). We used the MLESAC version of RANSAC which employs a mixture model to express the probability of each CPG datum in support of a given time-delay estimate. The advantage of this approach is that it evaluates the time-delay model for each source in isolation from other

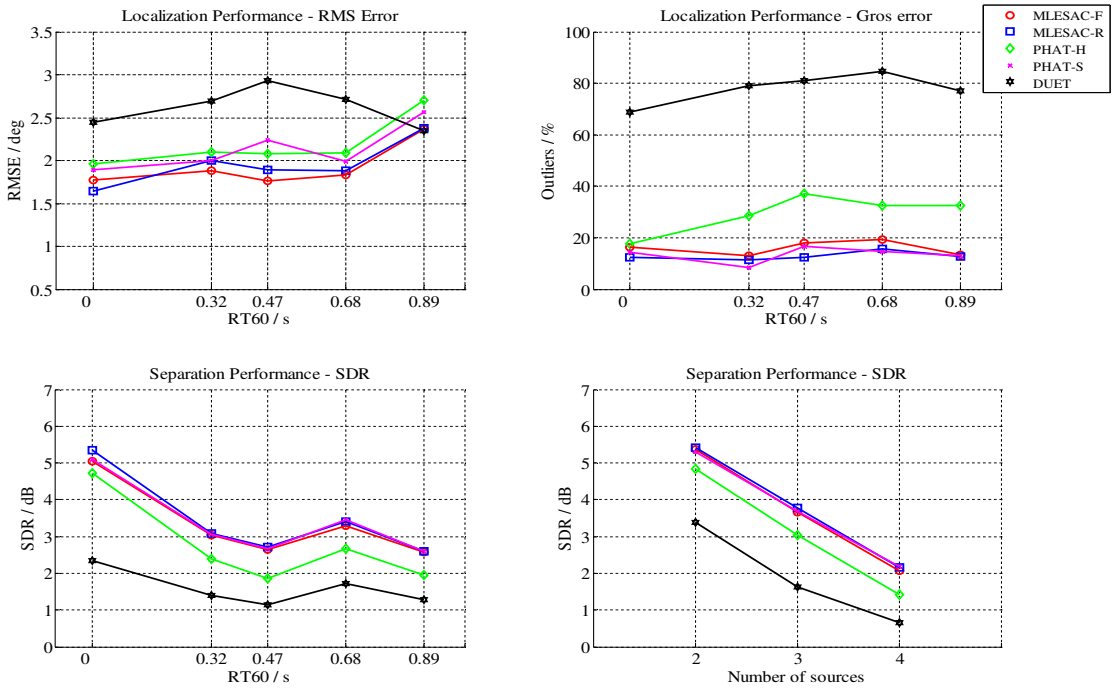


Figure 2: The top plots present comparison of localization performance using RMSE and Gross Error. The bottom plots show comparison of separation performance using SDR as a function of RT_{60} (bottom-left) and number of sources (bottom-right).

sources yet optimizes the parameter of the model using the evidence from all the data. The algorithm proved to be robust against reverberation in a multi talker scenario. Localization and separation performance was found to be better than that of DUET and PHAT-Histogram, and on par with PHAT-Sum.

The work on the algorithm could be taken forward to optimize the random sampling part of RANSAC approach to reduce the complexity of the algorithm, e.g. fewer iterations are required for anechoic localization than in reverberant conditions. Another area of investigations could be to reduce aggregation time (latency), currently $5s$.

6. ACKNOWLEDGMENTS

Authors would like to thank Chris Hummersone for sharing with us his recorded Binaural Impulse Responses.

7. REFERENCES

[1] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation maximization source separation and localization," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 18, no. 2, pp. 382–394, Feb. 2010.

[2] S. Rickard, *The DUET Blind Source Separation Algorithm*. Springer, 2007, ch. 8, pp. 217–237.

[3] H. Viste and G. Evangelista, "On the use of spatial cues to improve binaural source separation," in *in Proc. Int. Conf. on Digital Audio Effects (DAFx)*, 2003, pp. 209–213.

[4] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *in Proc. of the IEEE Conf. on Acoust., Speech, and Sig. Proc.*, 1997, pp. 375–378.

[5] H. F. Silverman and J. M. Sachar, "The time-delay graph and the delayogram-new visualizations for the time delay," *IEEE Sig. Proc. Letters*, vol. 12, no. 4, pp. 301–304, 2005.

[6] P. Aarabi, "Self-localizing dynamic microphone arrays," *IEEE Trans. on Syst., Man, and Cybernetics - part C*, vol. 32, no. 4, pp. 475–484, 2002.

[7] M. Fischler and R. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[8] P. H. S. Torr and A. Zisserman, "Mlesac: A new robust estimator with application to estimating image geometry," *Computer Vision and Image Understanding*, vol. 78, pp. 138–156, 2000.

[9] M. P. Cooke, J. Barker, S. P. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. of Acous. Soc. of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[10] C. Hummersone, T. Brookes, and R. Mason, "A comparison of computational precedence models for source separation in reverberant environments," in *AES Convention 128*, 5 2010.

[11] E. Vincent, C. Fevotte, and R. Gribonval, "Performance measurement in blind audio source separation," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 14, no. 4, pp. 1462–1469, 2006.