2011 International Conference on Computer Science and Information Technology (ICCSIT 2011)
IPCSIT vol. 51 (2012) © (2012) IACSIT Press, Singapore
DOI: 10.7763/IPCSIT.2012.V51. 122

Binaural multiple sources localization with extracting high local SNR frequencies based on cepstrum

Mengqi Ren, Yuexian Zou*, Hong Liu and Bo Li

Advanced Digital Signal Processing Laboratory, Peking University Shenzhen Graduate School , Shenzhen, Guangdong Province, China

Abstract. This paper investigates the directions of arrival (DOA) estimation problem of multiple speech sources by using two microphones, which has promising applications for auditory scene analysis (ASA) in intelligent service robots. In this paper, we study the DOA estimation based on the inter-channel phase difference (IPD) versus signal frequency framework (IPD-frequency plot). A novel DOA estimation approach has been developed in frequency domain by extracting the high local SNR frequency information aiming to improve the DOA estimation accuracy and the robustness to the noise. The high local SNR frequency information is extracted effectively by exploring the harmonic structure of speech sources in cepstrum domain. The time delay is estimated from the IPD-frequency plot using clustering methods. Experimental results show that the proposed multi-source cepstrum-based DOA estimation algorithm is robust to the additive Gaussian noise, has less complexity and higher DOA estimation accuracy (especially under low SNR condition) compared to those of sinusoidal modeling based DOA estimation method.

Keywords: Cepstrum, Clustering, Underdetermined situation, Speech signals, Direction of arrival (DOA) estimation

1. Introduction

Direction of arrival (DOA) estimation is a basic and important technique in array signal processing, which has many emerging applications such as hands-free mobile telephones, intelligent meeting rooms, intelligent robots and hearing-aid devices, etc. It is well-known that binaural source localization (BSL) technique is one of the hot research topics for imitating human listening ability [1]. Binaural multiple source localization (BMSL) is a difficult and challenging research problem since it means that the number of sensors is less than or equal to the number of the sound sources. Obviously, BMSL problem cannot be solved by the classical array signal processing algorithms, where the number of sensors is asked for larger than the number of sound sources. In literature, BMSL problem can be categorized into the underdetermined problem.

One class of the most important DOA estimation methods in underdetermined situations is speech attributes based algorithms. This method can deal with speech signals which have wideband and quasi-stationary properties by making the best use of speech specific attributes, i.e., sparsity in time-frequency domain, to distinguish the information of different sources. Several algorithms have been developed in this field. D. Banks exploited the presence of gaps in the frequency spectrum of each source, and detected the vertical segments in the frequency versus path difference plot to locate two sources [2]. Making use of speech sparsity attributes, C. Liu et al. proposed a multiple sources' DOA estimation method based on a frequency-DOA histogram plot [3]. S. Araki et al. extended this algorithm to a 3-dimensional absolute DOA estimation using 3-dimensional sensor array [4]. O. Yilmaz and S. Rickard used both inter-channel time difference (ITD) and inter-channel intensity difference (IID) to generate the frequency-DOA histogram [5]. M. I. Mandel et al. introduced a statistical noise model on ITD by estimating the posterior distribution on discretized ITD to localize the sources [6]. Z. Wenyi and B. D. Rao utilized speech's sparsity attributes in both time and

* Corresponding author. Tel.: +86-0755-26032016; fax: +86-0755-26032015.
E-mail address: zouyx@szpku.edu.cn.

frequency domain, and transformed the DOA estimation of the multiple speech sources into a clustering problem by creating the inter-channel phase difference (IPD) versus frequency plot. The basic idea to improve the accuracy of DOA estimation lies in extracting the frequency points with high SNR. There are two methods have been proposed. One is based on sinusoidal modeling [7], and the other one selected the frequency points with higher power than the threshold. Their simulations demonstrated that the former method gave higher DOA estimation accuracy [8]. However, it is noted that the accuracy of DOA estimation still needs to be improved, especially in low SNR conditions.

In this paper, following the DOA estimation framework developed in [8], we propose a novel approach to extract high local SNR points in frequency spectrum by homomorphic speech processing. The paper is organized as follows. In section II, the signal model and the basic concepts of IPD-frequency algorithm are presented. The proposed cepstrum-based DOA estimatoin algorithm by extracting high local SNR frequency points is derived in section III. We conduct three expriments to evaluate the performance of our proposed algorithm in section IV, followed by conclusions in Section V.

2. Signal Model

In this section, binaural source DOA estimation using inter-channel phase difference will be formulated, which includes the single source as well as multiple sources DOA estimation.

2.1. Inter-Channel Phase Difference-Based DOA Estimation

To make the presentation clear, we introduce the principle of inter-channel phase difference-based DOA estimation under the far-field non-attenuation and non-reverberation conditions. The signals captured by two microphones can be modeled as

$$y_1(t) = s(t) + n_1(t) \tag{1}$$

$$y_2(t) = s(t-\tau) + n_2(t) \tag{2}$$

where *s*(*t*) is the desired source signal and also propagating signal arrives the microphone one and $s(t-\tau)$ represents a delayed version of $s(t)$, τ is the propagation time difference between the microphone one and microphone two. $n_1(t)$ and $n_2(t)$ represent additive noise at the two microphones, respectively. The Fourier transform (FT) of $y_1(t)$ and $y_2(t)$ can be expressed as

$$Y_1(\omega) = S(\omega) + N_1(\omega) \tag{3}$$

$$Y_2(\omega) = S(\omega)e^{-j\omega\tau} + N_2(\omega) \tag{4}$$

where $S(\omega)$ denotes the FT of $s(t)$, $N_1(\omega)$ and $N_2(\omega)$ denote the FT of $n_1(t)$ and $n_2(t)$, respectively. The inter-channel phase difference $\psi_{12}(\omega)$ is determined as

$$\begin{aligned} \psi_{12}(\omega) &= \angle Y_1(\omega) - \angle Y_2(\omega) \\ &= \omega\tau + 2\pi n + v(\omega) \end{aligned} \tag{5}$$

where $\angle Y_1(\omega)$ and $\angle Y_2(\omega)$ denote the phase of $Y_1(\omega)$ and $Y_2(\omega)$. n is an integer number, $2\pi n$ represents possible phase unwrapping for $\psi_{12}(\omega)$ which is constrained in the range of $[-\pi, \pi]$ after the $\mathrm{mod}\ (2\pi)$. $v(\omega)$ denotes the IPD error which is introduced by the additive noise. More specifically, $N_1(\omega)$ and $N_2(\omega)$ can be modeled as Gaussian distribution or Laplacian distribution. Further study showed that the different modeling of the additive noise will give different DOA estimation result [8]. More discuss will be given later. From Fig. 1, the relationship between DOA θ and τ is given by

$$\tau = d\cos\theta / c \tag{6}$$

where d denotes the inter-space between two microphones, c denotes the sound speed in the air.

2.2. Speech's Sparsity Attribute and IPD

Literature studies showed that the speech signal has sparse attributes in the time-frequency domain, which has been used to achieve the DOA estimation of multiple sources using two microphones [2, 3]. Concretely, there are two kinds of sparsity characters of speech signals [7, 8].

- *Sparsity in time domain*: Natural speech generally has many short pauses and silent segments, that may appear more than half of the total time sequence.

- *Sparsity in frequency domain*: Unlike white signal, speech signal exhibits strong short correlation and the signal power is concentrated on harmonics of the pitch frequency, especially for voiced speech signal.

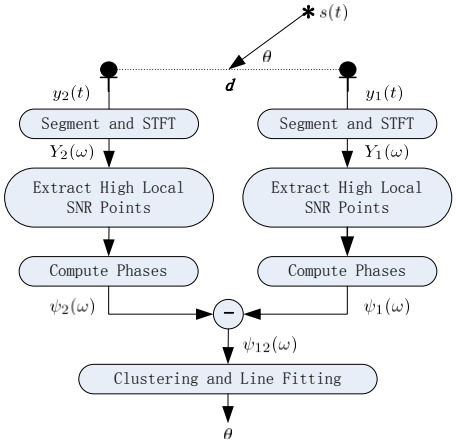

Fig. 1. IPD-frequency algorithm for DOA estimation

From speech sparsity attributes in time-frequency domain, it is easy to understand that we may distinguish different sources from the data captured by microphones according to the time-frequency relationship for multiple speech sources situation. This is the basic principle for using IPD-frequency plot to estimate the DOA of multiple sources. To give the clear description, this algorithm is shown in Fig. 1.

In this algorithm, as discussed before, the IPD error $v(\omega)$ will affect the accuracy of DOA estimation. Analysis shows that we can ignore this adverse noise influence on $\psi_{12}(\omega)$ given in (5) when the SNR is very high. One simple idea is to extract the high local SNR points in frequency domain and just use these extracted points to compute $\psi_{12}(\omega)$ accordingly.

3. Proposed Method

It is without doubt that the method to extract the high local SNR points in frequency spectrum will finally affect the DOA estimaton accuracy since this algorithm just uses partial information of the data. In this section, we explore to get a novel and efficient method for the high SNR frequency point extraction by considering the nature speech generation mechanism. Basically, speech signal is the response of human vocal tract to the excitation of glottal flow. Research shows that the speech cepstrum has the ability to separate the human glottal flow and the resonance of human vocal tract [9], or the harmonic structure of the human voice can be determined. More specifically, it can be seen that, for human speech, its power is concentrated on harmonics of the pitch frequency. This comes to the conclusion that the high local SNR frequency points directly reflect the harmonic structure of the human voice and they can be calculated from the cepstrum accordingly.

Let's define the log frequency spectrum $\hat{Y}_n(\omega, j)$ as

$$\hat{Y}_n(\omega, j) = \log[Y_n(\omega, j)], \qquad n = 1, 2 \tag{7}$$

where n and j is the channel index and frame index, respectively, $Y_n(\omega, j)$ denotes the *n*th channel frequency spectrum. The cepstrum $C_n(t, j)$ is defined as the Discrete Cosine Transformation (DCT) of $\hat{Y}_n(\omega, j)$ given as

$$C_n(t, j) = \mathcal{DCT}[\hat{Y}_n(\omega, j)] \tag{8}$$

It is noted that the lower and the upper cepstra correspond to the resonance of human vocal tract and the human glottal flow, respectively. As the result, the harmonic structure of human voice can be extracted as follows

$$\hat{C}_n(t, j) = \begin{cases} \epsilon \cdot C_n(t, j) & \text{for } t < lower\_cep \\ C_n(t, j) & \text{otherwise} \end{cases} \tag{9}$$

where ϵ is a small constant (close to zero), $lower\_cep$ is a threshold that separates the harmonic structure and the resonance of human vocal tract. In this study, we set ϵ to be 10^{-3} and $lower\_cep$ to be 5ms while the frame length is set to be 30ms. It can be seen that the filtered cepstrum $\hat{C}_n(t, j)$ only maintains the harmonic

structure information of the human voice, which will be used to generate the high local SNR frequency points. Firstly, the filtered log frequency spectrum $R_n(\omega, j)$ keeping the harmonic structure of the human voice can be determined by inverse DCT (IDCT)

$$R_n(\omega, j) = \mathcal{IDCT}[\hat{C}_n(t, j)] \tag{10}$$

Secondly, the filtered linear frequency spectrum $\hat{R}_n(\omega, j)$ then can be given by

$$\hat{R}_n(\omega, j) = \exp[R_n(\omega, j)] \tag{11}$$

In (11), it is noted that *j* is the frame index and the process described above should be performed for each frame.

Thirdly, to extract the high local SNR points from the $\hat{R}_n(\omega, j)$, we need to sort $\hat{R}_n(\omega, j)$ across all the frames and only keep the top X% frequency points in power.

Finally, the extracted frequency points will be used to estimate DOA.

Considering the spatial sampling theorem, we only take the frequency below $c/(2d)$ into account to avoid the spatial aliasing. The critical frequency f_a is determined as

$$\begin{aligned} |\omega\tau| = \pi \;&\Rightarrow\; |2\pi f_a \frac{d\cos\theta}{c}| = \pi \\ &\Rightarrow\; f_a = \frac{c}{2d|\cos\theta|} \end{aligned} \tag{12}$$

It is easy to understand that the minimum of f_a from (12) ends at $c/(2d)$. As an example, if *d=8cm*, we get f_a*=2kHz*. It can be seen that almost all the extracted frequency points are smaller than the minimum critical frequency f_a because the power of speech are concentrated at relative low frequencies. Thus, we can conclude this constraint will not degrade the final performance of DOA estimation for this algorithm.

For giving more clear presentation, here we summarize the proposed cepstrum-based DOA estimation algorithm.

- Set a threshold $\xi \in (0, 1]$, which determines the number of frequency points needed to be extracted.
- Segment $y_1(t)$ and $y_2(t)$ into frames.
- Calculate FT $Y_1(\omega)$ and $Y_2(\omega)$ for each frame.
- Calculate frequency spectrum $\hat{R}_1(\omega)$ and $\hat{R}_2(\omega)$ by (11).
- Extract high local SNR frequency index as follows

  $I_1 = \{(i, j)|\hat{R}_1(\omega_i, j) \geqslant \xi \cdot \max(\hat{R}_1(\omega))\}$
  $I_2 = \{(i, j)|\hat{R}_2(\omega_i, j) \geqslant \xi \cdot \max(\hat{R}_2(\omega))\}$
  $I = I_1 \cap I_2$
- Calculate IPD-frequency information: $\psi_{12}(\omega) = \angle Y_1(\omega_i, j) - \angle Y_2(\omega_i, j)$, where $(i, j) \in I$.
- Calculate θ by clustering and line fitting of $\psi_{12}(\omega)$ using the algorithm GMDA_Laplace [8].

4. Simulation

4.1. Simulation Setup

In this simulation study, three experiments are given to evaluate the performance of the proposed cepstrum-based DOA estimation algorithm compared to the sinusoidal modeling based DOA estimation algorithm [10]. The array setup is shown in Fig. 1, except two omnidirectional microphones spaced 8cm are used. Two far-field independent speech sources are considered, which is name as A-source and B-source with the impinging angles denoted as θ_A and $\theta_B = \theta_A + \Delta\theta$ ($\Delta\theta$ is the separation angle), respectively. The speech sources' sample rate is at 8kHz. The duration of each speech source is about 3 seconds. The 256-points FT is used to calculate the frequency spectra. For signal segmentation, the hamming window is adopted with a size of 30ms, and the overlap is set to 20ms. Additive Gaussian noise is considered for each microphone and channel noise is independent to each other. In order to eliminate the robustness to environment noise, the experiments are carried out 100 times independently.

4.2. Experiment 1: Robustness of DOA Estimation Algorithm

In this simulation, the robustness of the DOA estimation algorithm under different additive Gaussian noise levels is evaluated. We vary SNR levels from -5dB to 30dB stepsize by 5dB. Two speech sources are located at θ_A =60 ° and θ_B=135 °, respectively. For our proposed algorithm, set the threshold $\xi = 0.8$. This

means the top 20% frequency points are selected to estimate DOAs. The root mean squared (RMS) error is used as the performance measurement. Simulation result is shown in Fig. 2. From Fig. 2, it is clear to see that our proposed cepstrum-based algorithm outperforms to that of sinusoidal modeling based algorithm, especially in the low SNR levels. The higher the SNR is, the better the DOA estimation accuracy will be. From this simulation result, it can be seen that the highest DOA estimation RMS error of our proposed cepstrum algorithm for two sources will be below 1 ° when SNR is higher than 10dB. This is an encouraged result.

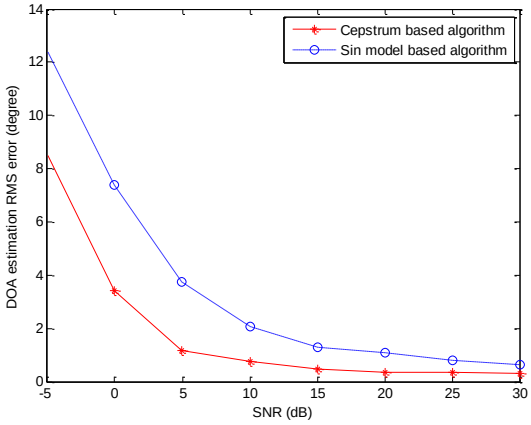

Fig. 2. Performance of two algorithms versus SNR

4.3. Experiment 2: Capability of DOA Separation

In this simulation, we evaluate the capability of two algorithms to separate two spatial sources when SNR is set to be 30dB. Two sources are placed symmetric to 90 °, and the separation angle varies from 4 ° to 160 °. All other parameters are set as the same as those in Experiment 1. Simulation result is presented in Table I.

Table 1. RMS Error versus Angular Separation

Angular Separation	4 °	6 °	8 °	30 °	80 °	140 °	160 °
θ_A	88 °	87 °	86 °	75 °	50 °	20 °	10 °
RMS error of Cepstrum based algorithm	0.52 °	0.36 °	0.49 °	1.28 °	0.23 °	0.96 °	7.72 °
RMS error of Sinusoidal modeling based algorithm	0.80 °	0.54 °	0.66 °	1.39 °	1.26 °	7.47 °	7.82 °

From Table I, we can see that our proposed cepstrum based algorithm gives better ability than that of sinusoidal modeling based algorithm at all angles. It is also noted that, for both of the two algorithms, the DOA estimation RMS error have similar tendency. More specifically, it can be seen that our proposed algorithm is able to estimate the correct DOAs for two sources when θ_A=88 ° and θ_B=92 ° at the RMS error of 0.52 °. It is interested to see that the estimation accuracy decreases when θ_A varies from 87 ° to 70 °, but increases slightly when θ_A goes from 70 ° to 50 °. But the DOA estimation accuracy dramatically degrades when θ_A goes smaller than 20 °, which satisfies the array theorem since the two sources are approaching the end-fire. The lowest DOA estimation RMS error is 0.23 ° when θ_A=50 °. The smallest separation angle is $\Delta\theta$=4 ° at the RMS error of 0.52 °. This also indicates the ability of our proposed algorithm that it is unable to distinguish two sources when $\Delta\theta$<4 °. In this simulation, it is concluded that the DOA estimation RMS error of our proposed algorithm is about 1 ° when 4 °<$\Delta\theta$<140 °.

4.4. Experiment 3: Computational Complexity

In this simulation, the computational complexity of two DOA estimation algorithms is studied. All simulation parameters are the same as those in Experiment 1 except SNR = 0dB. The simulation results are shown in Table II using MATLAB program. We can see clearly that the execution time of DOA estimation algorithm is reduced 71% by our proposed cepstrum-based DOA estimation algorithm.

Table 2. Algorithm Execution Time

	Our Algorithm	Algorithm [8]
Time	1.5194s	5.2338s

5. Conclusion

In this paper, we have introduced an effective method to extract the high local SNR frequency points by using cepstrum of two microphone output signals, which results in a novel cepstrum-based DOA estimation algorithm. In our algorithm, to get better IPD-frequency plot, the high local SNR frequency points is determined by selecting the high frequency components from its cepstrum, which essentially reflects the glottal flow components of the speeches. Then a GMDA_Laplace clustering method is employed to fit the IPD-frequency plot and give the estimation of the DOAs accordingly. Simulations using speech sources showed that, comparing to sinusoidal modeling based DOA estimation algorithm, the proposed cepstrum-based algorithm improves the DOA estimation accuracy across all SNR levels and under all angular separations. The resolution of our proposed algorithm could achieve 4 °when SNR equals to 30dB and sources are around the broadside. The complexity of our algorithm is also lower than sinusoidal modeling based algorithm.

6. Acknowledgements

This work is supported by the Chinese National 863 Program (2007AA11Z224), Shenzhen Science &Technology Program (SZKJ-200716) and the special foundation of president of PKUSZ (2010009).

7. References

[1] Y. Nagata, et al., "Binaural Localization Based on Weighted Wiener Gain Improved by Incremental Source Attenuation," IEEE Transactions on Audio, Speech, and Language Processing, vol. 17, pp. 52-65, 2009.

[2] D. Banks, "Localisation and separation of simultaneous voices with two microphones," Communications, Speech and Vision, IEE Proceedings I, vol. 140, pp. 229-234, 1993.

[3] C. Liu, et al., "Localization of multiple sound sources with two microphones," The Journal of the Acoustical Society of America, vol. 108, pp. 1888-1905, 2000.

[4] S. Araki, et al., "Doa Estimation for Multiple Sparse Sources with Normalized Observation Vector Clustering," in Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on, 2006, pp. V-V.

[5] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," IEEE Transactions on Signal Processing, vol. 52, pp. 1830-1847, 2004.

[6] M. I. Mandel, et al., "An EM algorithm for localizing multiple sound sources in reverberant environments," in Advances in Neural Information Processing Systems, B. Schölkopf, et al., Eds., ed. Cambridge, MA: MIT Press, 2007, pp. 953-960.

[7] R. McAulay and T. Quatieri, "Speech analysis/Synthesis based on a sinusoidal representation," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 34, pp. 744-754, 1986.

[8] Z. Wenyi and B. D. Rao, "A Two Microphone-Based Approach for Source Localization of Multiple Speech Sources," IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, pp. 1913-1928, 2010.

[9] T. F. Quatieri, Discrete-Time Speech Signal Processing: Principles and Practice Prentice Hall, 2001.

[10] D. P. W. Ellis. Sinewave and Sinusoid+Noise Analysis/Synthesis in Matlab [Online]. Available: http://www.ee.columbia.edu/~dpwe/resources/matlab/sinemodel/