

ON PRE-FILTERING STRATEGIES FOR THE GCC-PHAT ALGORITHM

Hong-Goo Kang¹, Michael Graczyk², Jan Skoglund²

¹ School of EE, Yonsei University, Seoul, Korea

² Google Inc., Mountain View, CA, USA

ABSTRACT

In this paper, we investigate the impact of the pre-filtering method to generalized cross-correlation (GCC) based direction of arrival (DOA) estimation. The role of pre-filtering is either to emphasize or deemphasize certain frequency components before computing cross power spectrum. However, its impact or relation to environmental variation, e.g., in noisy environments, has not been clearly studied yet. An efficient pre-filter should consider the relative importance of individual frequency components and adaptively change its related parameters based on environmental variations.

We first investigate the relationship between conventional pre-filtering functions and power normalization factors and its dependency on signal-to-noise-ratio (SNR) in various noisy environments. Then, we experimentally show that it is helpful to introduce a Wiener-like gain function into the pre-filtering process to efficiently design a generalized rule to determine control parameters. By analyzing the importance of parameter estimation errors in the estimation of Wiener filter parameters, e.g. noise power spectral density (PSD) and the *a priori* SNR, this paper proposes an efficient estimation strategy to design a pre-filter for robust GCC based DOA estimation.

1. INTRODUCTION

In multiple microphone based hands-free voice communication or voice activated human-computer-interaction (HCI) applications, it is important to accurately estimate locations of users to acquire desired speech signals [1]. The azimuth of the target user can be obtained by estimating time delays between the combination of microphone input pairs with a far-field assumption. Time delays between two microphone inputs are obtained by computing a cross-correlation, typically in the frequency domain to decrease computational complexity.

Since power levels and relevance of frequency contents vary based on signal characteristics and operating environments, many researchers have introduced various types of frequency-dependent weighting functions such as

the smoothed coherence transform [2], the maximum likelihood processor [3], the phase transform based methods (GCC-PHAT), and so on [4]. The power weighted or partial whitening based phase transform (GCC-PHAT _{β}) method that partially normalizes cross power spectral (CPS) components by their amplitudes has been frequently used in many applications because it softly reduces bias factors caused by the amplitude dynamics of frequency components [5]. Zhang et al proved that the GCC-PHAT converges to the ML estimator in high SNR conditions, which partially explain the method's consistent performance in low noise environments, even with reverberation [6].

To further increase the robustness of GCC-PHAT algorithm in noisy environments, several researchers introduced SNR dependent weights that are commonly used in single channel noise reduction modules, i.e. Wiener gain [7]. However, since the weights were independently determined only by the SNR of each frequency component and did not consider relative inter-component importance, they could not obtain consistent improvement in all noise conditions. Cross-correlation outputs are highly related to group delays or linear phase components in the frequency domain, so it is important to consider the relative importance of each frequency component. In addition, estimation error causes inconsistent results across noise and SNR conditions.

This paper investigates an efficient pre-filtering strategy for the GCC-PHAT _{β} method that adaptively controls relevant parameters depending on environmental variations. We considered two types of pre-filtering parameters, a power normalization factor and an SNR dependent frequency weighting function. We first experimentally analyze the impact of power normalization factors and SNR dependent frequency weighting functions on cross-correlation in various noise environments. The power normalization factor introduced in the GCC-PHAT _{β} method is effective in low noise conditions, but the optimal β differs in various noise environments. We then show that it is helpful to introduce a Wiener-like gain function into the pre-filtering process to design a generalized rule to determine the optimal β . By investigating the impact of estimation errors that are unavoidable during the Wiener estimation process, we propose an efficient strategy to design a unified pre-filter for DOA estimation.

This work was done while Hong-Goo Kang was working at Google as a visiting scientist.

2. GCC BASED TIME DELAY ESTIMATION

Let $s(n)$ be a target signal, the impulse response from the target signal to microphone i be $h_i(n)$, and additive noise be $d_i(n)$. A simplified signal model and its frequency domain representation can be represented as follows:

$$x_i(n) = s(n) * h_i(n) + d_i(n), \quad i = 1, 2, \quad (1)$$

$$X_i(k) = S(k)H_i(k) + D_i(k), \quad k = 1..K, \quad (2)$$

where n denotes time index, k frequency index and K a number of frequency bins. The time difference between two microphone inputs $x_1(n)$ and $x_2(n)$, τ , can be estimated by finding the lag that maximizes the generalized cross-correlation (GCC), defined as:

$$\begin{aligned} r(l) &= \mathcal{F}^{-1}\{\Psi(k)G(k)\}, \\ G(k) &= X_1(k)X_2^*(k), \end{aligned} \quad (3)$$

where $r(l)$ denotes the output of cross-correlation at lag l , $G(k)$ denotes cross power spectrum, and $\Psi(k)$ denotes frequency weights. The commonly used phase transform method uses the weighting function of $\Psi(k) = 1/|G(k)|$.

In ideal conditions, i.e. $h_1(n)$ and $h_2(n)$ are simple integer sample delays and noises are uncorrelated, the GCC output becomes $\delta(t - \tau)$. Since the normalization process sometimes over-amplifies the noise spectral regions and de-emphasizes the contribution of the frequency components that have larger amplitudes than others, it is better to introduce a soft power normalization factor β :

$$\Psi^\beta(k) = \frac{1}{|G(k)|^\beta}. \quad (4)$$

The heuristically determined fixed β term plays a role in partial pre-whitening [5]. The approach is called the β -weighted phase transform based GCC, i.e. GCC-PHAT $_\beta$.

It is not easy to quantitatively analyze the characteristics of the GCC-PHAT $_\beta$ algorithm because outputs vary depending on the value of β as well as the frequency characteristics of target speech and background noise components. Analysis is particularly complicated when the convolutional term $H_i(k)$ and the additive term $D_i(k)$ are non-linearly coupled. In general, two step removal approaches are used to minimize the coupling effect [7]. To simplify the analysis, we only consider the additive noise term and assume that $H_i(k)$ is a simple phase shift.

3. GCC-PHAT IN NOISY ENVIRONMENT

3.1. Impact of power normalization factor

In this section, we experimentally analyze the variation of the optimal power normalization factor in different types of noise and signal-to-noise ratio (SNR) conditions. The target speech samples are taken from the TIMIT database [8], and noise

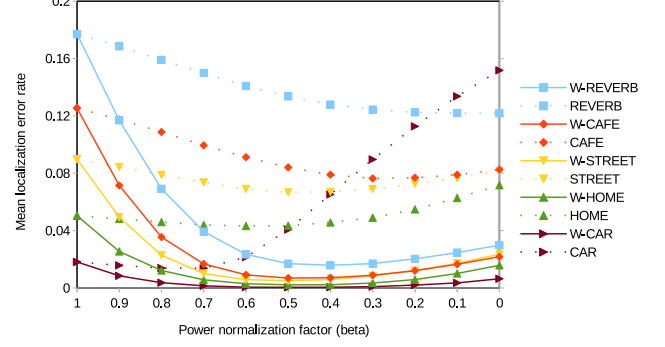


Fig. 1. Average localization error rate between oracle and estimated sample delays depending on the power normalization factor in various noisy conditions. ‘W-’ denotes the outputs that include Wiener filtering.

samples from the QUT-NOISE database [9]. To easily define oracle sample delays for reference, multiple pairs of the test samples are generated by varying the sample delays of the second microphone from -5 to 5 integer samples. The generated data resembles typical localization system inputs. Then, the target speech signals are mixed with various types of noise components by varying the SNRs from 0 to 30 dB with a step size of 5 dB. Although the type of microphones and recording setups of speech and noise samples are not perfectly matched in our experimental condition, it is good enough to analyze the overall trends. Finally, we obtained the average localization error rate by changing the power normalization factor from 0 to 1 with a step size of 0.1. The results were obtained at every 16ms with an analysis frame length of 32 ms.

The dotted line marks depicted in Figure 1 show the average localization error rate between oracle and estimated sample delays in various noisy conditions. In most cases, the best results were obtained when the power normalization factor was set between around 0.4 to 0.8, which agrees with previous research. However, because the optimal beta differs across scenarios, performance can be further improved with an adaptive power normalization factor depending on noise types. In other words, it is necessary to design an efficient way of controlling the power normalization factor depending on input noise characteristics and SNR variations.

The dotted line marks depicted in Figure 2 show the optimal power normalization factor that has the highest accuracy in different SNR conditions. Similar to the previous results [10], the optimal β value becomes smaller as the input SNR becomes lower. However, there is no consistent relationship between the β value and input SNRs. The reason is that the frequency characteristics of noise components are not white in general. In summary, it is better to change the power normalization factor, but it can not be generalized by frame SNRs alone.

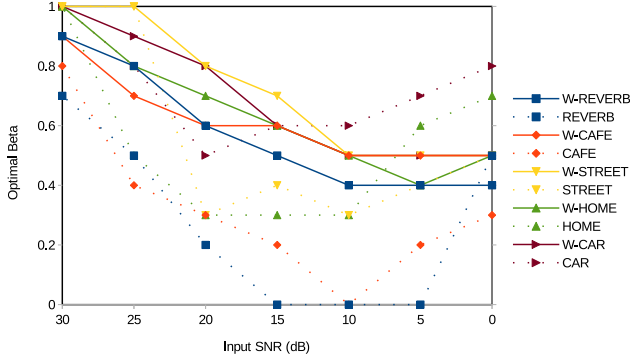


Fig. 2. Optimal power normalization factor in different noise and SNR conditions. ‘W-’ denotes the outputs that include Wiener filtering.

3.2. Signal-to-noise ratio (SNR) based weighting

A simple way of reducing the impact of background noise on delay estimation processing is to introduce an SNR dependent frequency weighting function to each microphone input [11]. Assuming that the source and noise are uncorrelated, the magnitude of the cross-correlation, $G(k)$, given in Eq.(3) can be represented as:

$$|G(k)| = |S(k)|^2 + |D(k)|^2 = |S(k)|^2 \left(1 + \frac{1}{\xi(k)} \right), \quad (5)$$

where $\xi(k)$ is the *a priori* SNR of each frequency component. The noise terms are correlated in general, thus it is better to apply the weighting function, i.e. Wiener gain to each channel such as:

$$W_i(k) = \frac{\xi_i(k)}{1 + \xi_i(k)}, \quad i = 1, 2, \quad (6)$$

$$\xi_i(k) = \frac{E\{|S(k)|^2\}}{E\{|D_i(k)|^2\}} = \frac{\lambda_s(k)}{\lambda_{d_i}(k)}, \quad (7)$$

where $\xi_i(k)$ is the *a priori* SNR of microphone input i . Since the $W_i(k)$ reduces the effect of noise components of each channel, the output of the pre-filtering can be regarded as a clean input signal if the Wiener gains are estimated accurately. The solid marks depicted in Fig.1 and 2 show the results when the ideal Wiener gains, i.e. assume that the PSDs of target speech and noise are known, are used for the pre-filtering. The average localization error rate reduces significantly, and the relationship between optimal β s and input SNRs becomes more uniform. Thus, it is much easier to design a rule to determine the optimal power normalization factor in different noise types and SNR conditions.

To estimate the *a priori* SNR required for computing the Wiener filter, the noise power spectral density (PSD) needs to be estimated first [12] [13]. Since the power spectrum density of target speech signal is also unavailable, the *a priori* SNR

is estimated using a decision-directed (DD) approach [14]. However, all that processing brings estimation errors that depend inconsistently on the types of noise and SNRs. Therefore, it is necessary to analyze the impact of estimation errors on the final DOA estimation.

4. RELATIONSHIP BETWEEN WIENER ESTIMATION ERROR AND GCC-PHAT

In this section, we investigate the impact of Wiener gain estimation error to the computation of GCC-PHAT in various SNR conditions. To represent the error terms caused by estimation errors, we replace the ideal Wiener gain $W_i(k)$ with the estimated one, $\hat{W}_i(k)$. Then, the GCC-PHAT $_{\beta}^W$ becomes:

$$r^w(l) = \mathcal{F}^{-1}\{\Psi^w(k)G^w(k)\}, \quad (8)$$

$$G^w(k) = X_1(k)X_2^*(k)\hat{W}_1(k)\hat{W}_2(k), \quad (9)$$

$$\Psi^w(k) = \frac{1}{|G^w(k)|^{\beta}}, \quad (10)$$

where all the variables denote similar terms described in Section 2. Wiener estimation errors are caused by inaccuracies in the noise power spectrum density (PSD) and *a priori* SNR estimation processes. We analyze the impact of each estimation error to the GCC-PHAT output in the following subsections. Note that all the experimental setups are the same as the one described in the previous section.

4.1. Impact of noise PSD estimation error

The noise PSD in non-stationary conditions can be estimated using minimum statistic [12] [13] or minimum mean square error (MMSE) estimator based approaches [15]. Since the approaches consist of multiple stages of parameter estimation processes, it is not easy to quantitatively characterize the estimation error terms.

To simplify the analysis, we introduce a first-order smoothing coefficient, α_d , that is related to the degree of estimation accuracy:

$$\hat{\lambda}_d^{(n)}(k) = \alpha_d \hat{\lambda}_d^{(n-1)}(k) + (1 - \alpha_d) \lambda_d^{(n)}(k), \quad (11)$$

where $\hat{\lambda}_d^{(n)}(k)$ and $\hat{\lambda}_d^{(n-1)}(k)$ denote the estimated noise PSD of the current and previous frames and $\lambda_d^{(n)}(k)$ true noise PSD of the current frame. The estimation error increases as the α_d is set close to one. Fig. 3 depicts average localization error rates between oracle and estimated sample delays obtained by averaging the results with the optimal β s in all of the noise conditions. The impact of noise estimation error on DOA estimation is not significant if the input SNR is higher than 5 dB. In addition, the performance does not degrade much even if we use large α_d s that simulate high estimation error conditions.

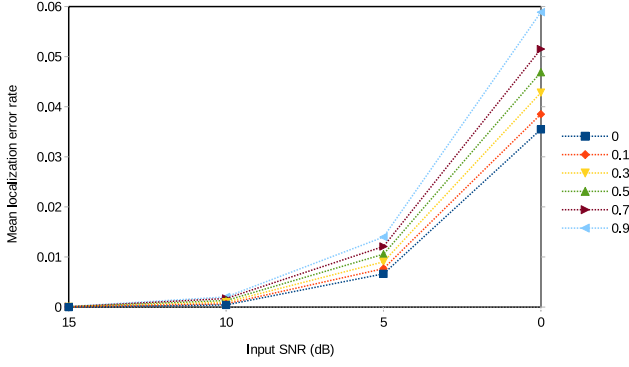


Fig. 3. Average localization error rate between oracle and estimated sample delays depending on the variation of smoothing coefficient for noise PSD estimation.

4.2. Impact of *a priori* SNR estimation error

In this section, we investigate the impact of *a priori* SNR estimation error to the computation of GCC-PHAT in various SNR conditions. Assuming that the noise PSD is obtained by the methods given in the previous section, the estimation error of the *a priori* SNR, $\xi(k)$ is modeled by the decision-directed approach that utilizes the *a posteriori* SNR.

$$\begin{aligned} \hat{\xi}^{(n)}(k) &= \alpha^{(n)}(k) \hat{\xi}^{(n-1)}(k) \\ &+ (1 - \alpha^{(n)}(k))(\max(|\gamma^{(n)}(k) - 1|, \epsilon)), \end{aligned} \quad (12)$$

where $\alpha^{(n)}(k)$ denotes a time and frequency-dependent smoothing coefficient and $\gamma^{(n)}(k) = \lambda_x^{(n)}(k)/\lambda_d^{(n)}(k)$ denotes *a posteriori* SNR.

At first, we performed similar experiments given in the previous subsection by fixing $\alpha^{(n)}(k)$ across both time and frequency domains, i.e. $\alpha^{(n)}(k) = \alpha$. The dotted line marks given in Fig. 4 depicts average localization error rates between oracle and estimated sample delays by varying the α term. The smoothing coefficient for noise PSD, α_d given in Eq. (11) is set to 0.7. The localization error rate becomes much worse than before, which confirms the importance of the *a priori* SNR estimation. The results also tell us that we should weight parameters estimated in the current frame more heavily. It is different from conventional noise reduction systems for voice communication that commonly adopt a less aggressive strategy to minimize the spectral fluctuation of the enhanced speech and musical background noise effect.

To investigate the effectiveness of time and frequency-dependent smoothing coefficient, $\alpha^{(n)}(k)$, we performed the same experiments using an adaptive estimation approach proposed in [16]. The solid line marks given in Fig. 4 depict the results. The *adaptive*(0.7) and *adaptive*(0.0) denote the results when α_d is set to 0.7 and 0.0, respectively. The adaptation process was not effective in DOA estimation, which leads us to conclude that the specific criterion to design a weighting

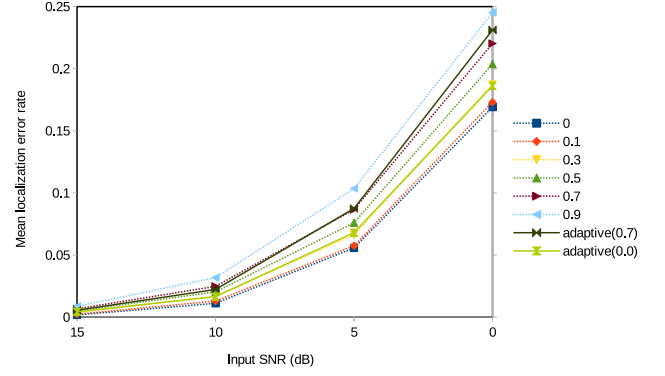


Fig. 4. Average localization error rate between oracle and estimated sample delays depending on the variation of smoothing coefficient for *a priori* SNR estimation.

filter for DOA estimation needs to be modified.

4.3. An idea to design a unified pre-filter

The experimental results given above provide an interesting guideline to design a new type of unified pre-filter for the GCC-PHAT _{β} algorithm. Using the fact that the relationship between the power normalization factor and input SNRs becomes similar if the Wiener gain is applied to each channel, two types of pre-filters are merged into a single unified form. Therefore, the unified pre-filter can be regarded as a modified form of Wiener gain estimator that also considers the PSD of target speech signal, e.g. $W_i^u(k) = J\{\lambda_s(k), \xi_i(k)\}$, where $J\{\}$ denotes a criterion or a mapping rule that needs to be estimated. A simple rule-based approach can be designed, but it would be better to adopt a training based approach using a large database of recordings because the relationship between two related parameters changes with environmental variations.

5. CONCLUSIONS

We experimentally analyzed the impact of the pre-filtering method to generalized cross-correlation (GCC) based direction of arrival estimation in various types of noise environments. After investigating the relationship between the power normalization factor of GCC-PHAT _{β} and frame SNRs, we experimentally showed that it is useful to introduce a Wiener-like gain function to determine the optimal power normalization factor in various types of noise conditions. From the results, we concluded that an efficient pre-filter should consider the relative importance of individual frequency components and needs to adaptively change its related parameters based on environmental variations. We also suggested an idea to design a unified pre-filtering paradigm.

6. REFERENCES

- [1] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: An overview," *EURASIP Journal on Applied Signal Processing*, pp. 1-19, 2006.
- [2] G. C. Carter, A. H. Nuttall, and P. G. Cable, "The smoothed coherence transform," *Proc. of the IEEE*, vol. 61, no. 10, pp. 1497-1498, 1973.
- [3] H. Saarnisaari, "ML time delay estimation in a multipath channel," *IEEE 4th International Symposium on Spread Spectrum Techniques and Applications Proceedings*, vol.3 pp. 1007-1011, 1996.
- [4] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-24, no. 4, pp. 320-327, 1976.
- [5] K. D. Donohue, A. Agrinoni and J. Hannemann, "Audio signal delay estimation using partial whitening," *Proc. 2007 IEEE SoutheastCon*, pp. 466-471, 2007.
- [6] C. Zhang, D. Florencio, and Z. Zhang, "Why Does Phat work well in low noise, reverberative environments," *Proc. ICASSP2008*, pp. 2565-2568, 2008.
- [7] Y. Rui and D. Florencio, "Time delay estimation in the presence of correlated noise and reverberation," *Proc. ICASSP2004*, vol. 2, pp. 133-136, 2004.
- [8] W. M. Fisher, G. Doddington, and K. M. Goudie-Marshall, "The DARPA speech recognition research database: specifications and status," *Proc. DARPA Workshop on Speech Recognition*, pp. 93-99, 1986.
- [9] D. B. Dean, S. Sridharan, R. J. Vogt, and M. Mason, "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," *Interspeech 2010*, Makuhari, Japan, 2010.
- [10] B. Qin, H. Zhang, Q. Fu, and Y. Yan, "Subsample time delay estimation via improved GCC PHAT algorithm," *Proc. ICSP2008*, pp. 2579-2582, 2008.
- [11] J. Valin, F. Michaud, and J. Rouat, "Robust 3D localization and tracking of sound sources using beamforming and particle filtering," *Proc. ICASSP2006*, pp. 841-844, 2006.
- [12] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504-512, 2001.
- [13] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *Signal Processing Letters*, Vol. 9, No. 1, pp. 12-15, Jan. 2002.
- [14] Y. Ephraim and D. Malah, "Speech enhancement using a minimum meanSquare error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109-1121, 1984.
- [15] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," *Proc. ICASSP2010*, pp. 4266-4269, 2010.
- [16] I. Cohen, "On the decision-directed estimation approach of Ephraim and Malah," *Proc. ICASSP2004*, pp. 293-296, 2004.