Article

Interaural Coherence Estimation for Speech Processing in Reverberant Environment

Seong-Hu Kim and Yong-Hwa Park *

Department of Mechanical Engineering, Korea Advanced Institute of Science and Technology, 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Korea; seonghu.kim@kaist.ac.kr

* Correspondence: yhpark@kaist.ac.kr; Tel.: +82-42-350-3235

Received: 1 December 2019; Accepted: 20 January 2020; Published: 22 January 2020

Abstract: Interaural coherence is used to quantify the effects of reverberation on speech, and previous studies applied the conventional method using all previous time data in the form of an infinite impulse response filter to estimate interaural coherence. To consider a characteristic of speech that continuously changes over time, this paper proposes a new method of estimating interaural coherence using time data within a finite length of speech, which is called the quasi-steady interval. The length of the quasi-steady interval is determined with various frequency bands, reverberation times, and short-time Fourier transform (STFT) variables through numerical experiment, and it decreased as reverberation time decreased and the frequency increased. In this interval, a diffuse speech, which is an infinite sum of reflected speeches of different propagating paths, is uncorrelated between two microphones apart from each other; thus, the coherence is close to zero. However, a direct speech measured at the two microphones has steady amplitude and phase difference in this interval; thus, the coherence is close to one. Moreover, the new method is the form of a finite impulse response filter that has a linear phase delay or zero phase delay with respect to speech to frequency; thus, the same or zero time delay for each frequency is applied to the power spectral density. Therefore, the coherence estimation of the new method is closer to the ideal value than the conventional one, and the coherence is accurately estimated at the time–frequency bins of direct speech, which is time-varying according to speech variation.

Keywords: reverberation; interaural coherence; speech processing; power spectral density estimation

1. Introduction

In a real situation, influences of the surrounding environment, such as multiple speakers, external noise sources, reverberation, etc., distort target speech information such that the performance of speech recognition methods get worse [1]. Among the causes of speech distortion, multiple speakers and external noise sources are additive noises, which have a low correlation with the target speech, making it easy to extract information about the target speech. Reverberation, on the other hand, is convolutive noise caused by the sound waves reflected by the surrounding walls or objects. In reverberant speech, both direct speech and reverberation, which is attenuated direct speech with a time delay, are involved. The reverberant speech has a high correlation with target speech; hence, it is difficult to extract the target information. Thus, the performance of speech separation decreases because the reverberation changes the amplitude and phase of time–frequency bins of direct speech [2,3]; moreover, speech recognition performance is also degraded.

On the other hand, humans recognize speech accurately even in real reverberant environments. In the human hearing system, a direct source and early reflections are emphasized, and late reflections are suppressed, which is called the precedence effect [4–6]. By implementing the precedence effect with a computational algorithm based on the relationship between reverberation and interaural coherence

(IC), previous studies attempted to solve reverberation problems. The ideal coherence between signals obtained from two microphones for the diffuse source reflected off the walls can be represented in the form of a square value of the sinc function in the frequency domain [7,8]. If the two microphones are apart far enough, the IC appears close to zero for diffuse sources and close to one for direct sources at most frequencies. Based on these characteristics, the performance is improved by applying IC to the direction of arrival (DoA) estimation of the speaker [9], speech, or source separation [10] and dereverberation [11–13] in a reverberation environment.

To get the best performance of the various speech preprocessing algorithms that use coherence, the estimated ICs of reverberant speech should match the ideal IC. Most of the algorithms [9–13] use the IC in the form of an infinite impulse response (IIR) filter in the calculation of power spectral densities. Due to recursive nature of the IIR filter, all of the past data have influence on the IC estimation, which may give an adversary effect on non-stationary speech data. Since the utterance changes continuously in speech, using all the past data will make coherence estimation different from the ideal coherence. Therefore, we propose a new IC estimation method of speech using a finite number of data in the form of a finite impulse response (FIR) filter to accurately estimate it close to the ideal IC. The optimal number of data required for estimating IC varies according to gender, frequency, and reverberation time; hence, the optimal IC estimation parameters should be determined by considering these factors. The performance of the proposed method is compared with that of the conventional IC estimation method through a numerical experiment using real speech data and measured room impulse responses.

The structure of this paper is organized as follows: Section 2 presents a new IC estimation method and theoretical validation of the proposed method. In Section 3, the optimal IC estimation parameters are determined based on gender, frequency, and reverberation time, and compared to the conventional IC estimation method for the estimation accuracy of IC. Sections 4 and 5 present the discussion and conclusions, respectively.

2. Interaural Coherence Estimation for Reverberant Speech

2.1. Interaural Coherence

We assume that the speech at a specific speaker location is recorded using two microphones apart from each other, and each recording signal is $x_1(t)$ and $x_2(t)$. To analyze the speech signal in both the time and the frequency domain, short-time Fourier transform (STFT) using overlap is applied, and the time–frequency bin values of the two microphones for the n -th time frame with frequency f can be expressed as $X_1(n, f)$ and $X_2(n, f)$, respectively. The auto- and cross-power spectral density (PSD) of the two microphone signals is represented by $\Phi_{ij}(n, f)$ ($i, j = 1, 2$). In this case, this value and the IC $\Gamma^2(n, f)$ are defined as

$$\Gamma^2(n, f) = \frac{|\Phi_{12}(n, f)|^2}{\Phi_{11}(n, f)\Phi_{22}(n, f)}, \quad (1)$$

$$\Phi_{ij}(n, f) = E[X_i^*(n, f)X_j(n, f)] \quad (i, j = 1, 2). \quad (2)$$

It is hard to calculate accurate auto- and cross-PSD (i.e., $\Phi_{ij}(n, f)$) using Equation (2) with finite-length $X_1(n, f)$ and $X_2(n, f)$. In previous studies [9–11], the PSD was estimated by multiplying exponentially decaying weight and summing continuous time–frequency bins over time as

$$\Phi_{ij}(n, f) = \alpha X_i^*(n, f)X_j(n, f) + (1 - \alpha)\Phi_{ij}(n - 1, f). \quad (3)$$

In this case, α has a range of $0 \leq \alpha \leq 1$. As a recursion formula, all the time–frequency bin information for the previous time affects the estimation of the PSD for the time frame n . However, since the speech changes continuously over time, the equation using all previous data can estimate the inaccurate PSD. An incorrect IC may be provided, which can cause performance problems. In addition,

the proposed method is calculated recursively using all speech bins, which can take more time. To solve these problems, this paper proposes a new IC estimation method using finite time–frequency bin data, and it is compared with the conventional method in terms of the estimation.

Interaural Coherence Estimation Using Finite Time–Frequency Bins

In digital signal processing, the PSD of a stationary time-domain signal $x(t)$ is derived by ensemble averaging of $X(n, f)$ based on the assumption of the ergodic process, and this is called Welch's method [14]. Based on this method, the estimated PSD for a specific time n and frequency f is calculated as follows by averaging $L_1 + L_2 + 1$ data for time ($L_1, L_2 \geq 0$):

$$\Phi_{ij}(n, f) = \frac{1}{L_T} \sum_{k=-L_1}^{L_2} X_i^*(n+k, f) X_j(n+k, f). \quad (4)$$

Here, L_T is the total length of the interval to apply averaging, and it is equal to $L_1 + L_2 + 1$. In the case of speech, it does not satisfy the ergodic process because it is an unsteady state that continually changes with time. Nevertheless, a speech signal within short time intervals can be almost steady, and it is called a quasi-steady-state signal in this paper. The number of time frames L_T is a quasi-steady-state interval of the speech signal, and it is applied to Equation (4) to estimate the correct PSD. Also, in the case of using finite data L_T , the uniform estimation method is applied to most time–frequency bins and provides accurate IC. Before estimating the PSD for a particular time frame n , it is more quasi-steady for the interval close to n for time; thus, we apply L_1 and L_2 as L and estimate the accurate PSD using the bin data in the interval of $2L + 1$.

By applying the IC based on the proposed PSD estimation from an anechoic to an echoic room, the theoretical validity is assessed. Firstly, we suppose that two microphones are used to record speech in an anechoic environment. Here, X_1 and X_2 have a phase difference of $e^{j2\pi f\tau_0}$ for all n due to time delay τ_0 according to the azimuth of the speaker, since only the direct speech exists. The estimated auto- and cross-PSD is expressed as

$$\Phi_{11}(n, f) = \frac{1}{L_T} \sum_{k=-L}^L |X_1(n+k, f)|^2, \quad (5)$$

$$\Phi_{22}(n, f) = \frac{1}{L_T} \sum_{k=-L}^L |X_2(n+k, f)|^2, \quad (6)$$

$$\Phi_{12}(n, f) = \frac{1}{L_T} \sum_{k=-L}^L |X_1(n+k, f)| |X_2(n+k, f)| e^{j2\pi f\tau_0}. \quad (7)$$

If $x_i(t)$ is steady within the ensemble average interval L_T , $|X_i(n-k, f)|$ is constant for $-L \leq k \leq L$, and the IC appears as one using Equations (5)–(7). Conversely, if $x_i(t)$ is unsteady, we get k with different values of $|X_i(n-k, f)|$ for $-L \leq k \leq L$, and the following inequality is satisfied:

$$\left(\sum_{k=-L}^L |X_1(n+k, f)| |X_2(n+k, f)| \right)^2 < \left(\sum_{k=-L}^L |X_1(n+k, f)|^2 \right) \left(\sum_{k=-L}^L |X_2(n+k, f)|^2 \right). \quad (8)$$

The IC calculated based on Equation (8) appears to be less than 1. In the case of speech, it can be seen that the IC is very close to one when the ensemble averaging interval L_T is determined to be the same length as the quasi-steady interval.

Secondly, we suppose that two microphones are used to record speech with static locations of a speaker and two microphones in an echoic environment. When $s(t)$ is a speech signal coming into

the first microphone without reflection, the two microphone signals $x_1(t)$ and $x_2(t)$ are expressed as follows based on the image source method [15]:

$$x_1(t) = s(t) + \sum_{k=1}^{\infty} a_{1k}s(t - \tau_{1k}), \quad (9)$$

$$x_2(t) = a_0s(t - \tau_0) + \sum_{k=1}^{\infty} a_{2k}s(t - \tau_{2k}). \quad (10)$$

a_0 and τ_0 are the attenuation and time delay due to the distance and azimuth between the speaker and microphones, and a_{ik} and τ_{ik} ($i = 1, 2$) are the attenuation and time delay for the speech signal, respectively, reflected from the wall. Diffuse speech, the sum of the reflected speech except for the direct speech $s(t)$ and $a_0s(t - \tau_0)$, represents noise for the direct speech. To see only the effect of diffuse speech, we assume that $s(t)$ and $a_0s(t - \tau_0)$ for the interval L_T are quasi-steady. Diffuse speech is an infinite sum of $a_{ik}s(t - \tau_{ik})$; thus, it also satisfies the quasi-steady state. However, diffuse speech is a random signal compared to direct speech because the speech continues to change over time. Consequently, the coherence between two microphones is far from one due to the effect of reverberation, and it deteriorates as the energy of diffuse speech is greater than that of direct speech. This shows that the IC in an echoic environment is lower than that in an anechoic environment.

In summary, if direct speech is dominant, it is similar to an anechoic environment; thus, the IC is close to one. If there is only diffuse speech without direct speech, the IC appears near zero. Also, direct speech should be quasi-steady and diffuse speech is an almost random signal within the averaging interval L_T ; thus, the IC can quantify the effects of reverberation. However, contrary to the theory, the quasi-steady interval cannot be perfectly steady in real speech. To compensate for this, an exponentially decreasing weight away from the n -th time frame is applied, and it estimates the PSD rather than giving the same weight in the interval as in Equation (4). The final proposed PSD estimation equation is as follows:

$$\Phi_{ij}(n, f) = \sum_{k=-L}^L A\beta^{|k|} X_i^*(n+k, f) X_j(n+k, f), \quad (11a)$$

$$\sum_{k=-L}^L A\beta^{|k|} = 1. \quad (11b)$$

Here, β is the decaying ratio, and it has a range of $0 < \beta \leq 1$. Also, A is a value to make the sum of weights equal to one; therefore, we can derive A according to L and β from Equation (11b). Thus, it satisfies the following equation:

$$A = \{1 + 2\beta(1 - \beta^L)(1 - \beta)^{-1}\}^{-1} \quad (12)$$

It is necessary to determine L (or L_T) and β corresponding to the quasi-steady interval to estimate the accurate IC using Equation (11a). Based on the numerical experiment, the optimal L_T and β are determined according to gender, frequency, and reverberation time. The proposed method is compared with the conventional IC estimation method to validate it.

3. Numerical Experiment and Results

As mentioned in the previous section, accurate estimated IC means that it is close to one for direct speech and zero for diffuse speech. Therefore, estimation accuracy is judged based on the IC values of the time–frequency bin with direct speech and the bins without direct speech for reverberation speech. The reverberant speech data used for verification were generated by convolution of the binaural room impulse response to utterances recorded in an anechoic environment. The anechoic utterance data

included 4380 male and 1920 female utterances from the Texas Instruments/Massachusetts Institute of Technology (TIMIT) database spoken by 630 native American English speakers [16]. The dataset was divided into a training set with 4620 utterances and a test set with 1680 utterances. The training set was used for determining the optimal L_T and β , and the test set was used for performance comparison between the conventional and proposed methods. The binaural room impulse response was based on the Aachen Impulse Response (AIR) database [17], which recorded impulse according to various reverberation environments and azimuth using a KEMAR dummy head. By selecting the binaural impulse response recorded in an office, stairway, and lecture room, reverberant speech with a sampling rate of 16 kHz was generated, and simulations were performed.

3.1. Evaluation Metric

For reverberant speech, the original speech is used to distinguish between bins with direct speech plus a relatively small amount of diffuse speech and bins with diffuse speech only. The time when the impulse starts from the impulse response corresponding to the specific reverberation environment is applied as a time delay to the original speech to align the start time of direct speech. STFT is applied to the time-aligned original speech, and a binary mask M_{direct} that suppresses the time–frequency bins lower than a threshold to reduce speech energy by 0.01 dB is determined. M_{direct} is a binary mask to suppress diffuse speech bins of reverberant speech. In contrast, a binary mask M_{diffuse} suppresses the time–frequency bins higher than the threshold. The perceptual evaluation of speech quality (PESQ) [18] between the original speech and M_{direct} applied original speech is about 4.0 compared to the maximum value of 4.5, which means that it extracts only direct speech. Thus, the IC values of direct plus a relatively small amount of diffuse speech and diffuse speech only are extracted by applying M_{direct} and M_{diffuse} to the IC of reverberant speech.

To quantify the estimation performance, the mean accuracy of coherence (MAC) is defined as the average value of similarity between the ideal coherence (one for direct speech only and zero for diffuse speech only) and the estimated IC. When the IC of reverberant speech is $\Gamma^2(n, f)$, the coherence accuracy of direct speech bin is set to Γ^2 , and the coherence accuracy of diffuse speech bin is set to $1 - \Gamma^2$. The MAC is calculated as

$$\text{MAC}_{\text{direct}} = \frac{\sum_{n,f} \Gamma^2(n, f) M_{\text{direct}}(n, f)}{\sum_{n,f} M_{\text{direct}}(n, f)}, \quad (13)$$

$$\text{MAC}_{\text{diffuse}} = \frac{\sum_{n,f} \{1 - \Gamma^2(n, f)\} M_{\text{diffuse}}(n, f)}{\sum_{n,f} M_{\text{diffuse}}(n, f)}. \quad (14)$$

Because the IC ranges from 0–1, the MAC also ranges from 0–1, and the closer it is to one, the better the estimation performance is. L_T and β , where M_{direct} and M_{diffuse} have minimum values, are optimal parameters. They are calculated by multiplying a binary mask for the frequency interval to determine the parameters according to frequency. $\text{MAC}_{\text{total}}$ is defined as

$$\text{MAC}_{\text{total}} = \text{MAC}_{\text{direct}} + \text{MAC}_{\text{diffuse}}, \quad (15)$$

which was used to compare the overall performance and determine L_T and β according to gender and frequency. The determined parameters were applied to the proposed method, and its performance was compared with that of the conventional estimation method.

3.2. Optimal Parameter Determination

Using the proposed IC estimation methods with $\text{MAC}_{\text{total}}$, L_T and β were analyzed according to gender and frequency in various reverberation environments with RT_{60} . Numerical simulations were conducted to see the tendency of parameters according to gender, frequency, and reverberation time. In total, 3260 male and 1360 female speech samples of the training set were generated assuming

a speaking situation in front of the dummy head in three different reverberation environments with 0.37 s, 0.69 s, and 0.79 s of RT_{60} , as shown in Table 1. STFT with a 25-ms Hamming window, 10-ms hop length, and 512-point fast Fourier transform (FFT), which are commonly used for general speech processing, was applied to the generated reverberant speech.

Table 1. Properties of the various rooms.

Room	d_{LM}	RT_{60}
Office	1.00 m	0.37 s
Stairway	2.00 m	0.69 s
Lecture room	5.56 m	0.79 s

First of all, to determine the L_T and β for the male speech, MAC_{total} was calculated at every 1-kHz interval for each STFT-applied 3260 reverberant male speech samples of the training set, and the parameters with the largest MAC_{total} were determined to be optimal. Since the range of MAC_{total} was different for each reverberant speech sample, to reduce the effect of different utterances, we calculated a histogram of the optimal parameters for each frequency interval. As an example, the histogram for the 3–4-kHz range of 3260 male speech samples in the reverberation environment of 0.37 s is shown in Figure 1. The optimal L_T and β values for each speech sample are distributed around the maximum point in the histogram. To determine the same IC estimation equation for various speech, we tried to determine the optimal parameters corresponding to the maximum point of the histogram. $L_T = 7$ and $\beta = 0.75$ for the maximum point could be estimated to be the optimal parameters. Around the optimum value of $L_T (= 7)$ and $\beta (= 0.7–0.8)$, the MAC_{total} was insensitive with variation less than or equal to 0.0001, which is small.

Figure 1. Histogram of 3260 reverberant male speech samples for L_T and β at 3 kHz to 4 kHz in $RT_{60} = 0.37$ s based on mean accuracy of coherence (MAC_{total}).

In the same way, the optimum L_T and β of the 3260 male speech samples for a total of eight frequency intervals in three different reverberation environments were determined, and the results are shown in Figure 2. With less reverberation and with 0.37 s of RT_{60} , it is appropriate to use nine time–frequency bins at 0–1 kHz and seven bins at the other intervals, and β decreased from about 0.8 to 0.5 as the frequency increased. As the effects of reverberation increased to 0.69 s and 0.79 s of RT_{60} , L_T was nine or 11 bins and β was 0.9 or 1.0. L_T for all frequency intervals increased as the reverberation time increased, but it can be seen that L_T had a higher value from 0 kHz to 4 kHz regardless of the reverberation. β had high and steady values with all frequencies in a high reverberation time environment, but it decreased as the frequency increased in 0.37 s and 0.69 s of RT_{60} . This tendency

according to frequency and reverberation time can be used to estimate the parameters of male speech in other reverberation environments.

Figure 2. Optimal L_T and β of male speech according to frequency intervals in (a) 0.37 s, (b) 0.69 s, and (c) 0.79 s of RT_{60} with a 25-ms Hamming window, 10-ms hop length, and 512-point FFT.

Secondly, the same method based on the histogram was used to determine the parameters for the 1360 female speech samples. The results for optimal L_T and β according to frequency and reverberation time are shown Figure 3. For each of the three reverberation environments, the optimal parameter values for frequency were similar than those of the male samples, but the difference was that L_T appeared low and β appeared high in the range of 0–4 kHz with 0.69 s of RT_{60} . Since L_T represents the length of the quasi-steady interval and β indicates how close the interval is to the steady state, this difference was not considered significant. Likewise, the parameters increased, and β was steady for frequency with increasing RT_{60} , while L_T tended to be relatively high at low frequencies. Therefore, it can be seen that there was no significant difference in the parameters according to gender, and we newly determined the optimal L_T and β according to the reverberation environment and frequency using 4620 speech samples of the training set, as shown in Table 2. It can be seen that the parameters using all speech samples had a consistent tendency and similar values for each gender. These optimal L_T and β values were applied to the proposed method in this paper, and the performance was compared with that of the conventional IC estimation method.

Figure 3. The optimal L_T and β of female speech according to frequency intervals in (a) 0.37 s, (b) 0.69 s, and (c) 0.79 s of RT_{60} with a 25-ms Hamming window, 10-ms hop length, and 512-point FFT.

The parameters of the proposed method according to the frequency and reverberation environment were determined to estimate accurate IC in time–frequency bins with the STFT variables mainly used in speech processing. However, the time–frequency bins of the speech may change depending on the STFT variables (window, hop, and FFT length), and it should be checked that the proposed IC estimation method can achieve sufficient estimation performance even when various STFT variables are applied.

Thus, the time–frequency bins with a 64-ms Hamming window, a 16-ms hop length, and 1024-point FFT for speech were used in the three different reverberant environments. L_T and β are only related to the quasi-steady interval based on theoretical analysis, and they were not affected by gender in the previous numerical experiment. In this case, we rechecked that there was also no influence of gender; however, the details of rechecking results are skipped in this paper. The optimal parameters for each reverberation environment and frequency using the longer window, hop, and FFT length, as shown in Table 3, were determined based on the averaged MAC_{total} of the 4620 speech samples of the training set through the same process used for the previous numerical experiment. In the environment with 0.37 s of RT_{60} , L_T was relatively high in the low frequency range. However, L_T showed a steady value of seven according to frequency in other reverberation environments, and β also appeared constant according to frequency in the case of the largest reverberation. In comparison to the previous numerical experiment with short STFT variables, L_T was relatively small, and β was largely changed by L_T because of the long window and hop lengths. Based on these optimal L_T and β values, the estimation performance of the conventional method and the proposed method with the longer STFT length was also compared in this paper.

Table 2. Optimal L_T and β according to the frequency intervals and reverberation environment using a 25-ms Hamming window, 10-ms hop length, and 512-point FFT.

RT ₆₀ .		Frequency							
		0–1 kHz	1–2 kHz	2–3 kHz	3–4 kHz	4–5 kHz	5–6 kHz	6–7 kHz	7–8 kHz
0.37 s	L _T	9	7	7	7	7	7	7	7
	β	0.70	0.85	0.70	0.75	0.70	0.75	0.60	0.65
0.69 s	L _T	9	9	9	9	7	9	9	9
	β	0.90	0.95	0.95	0.95	1.00	0.85	0.80	0.80
0.79 s	L _T	11	11	11	9	9	9	9	9
	β	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 3. Optimal L_T and β according to the frequency intervals and reverberation environment using a 64-ms Hamming window, 16-ms hop length, and 1024-point FFT.

RT ₆₀ .		Frequency							
		0–1 kHz	1–2 kHz	2–3 kHz	3–4 kHz	4–5 kHz	5–6 kHz	6–7 kHz	7–8 kHz
0.37 s	L _T	7	7	5	5	5	5	5	5
	β	0.55	0.55	0.65	0.70	0.60	0.70	0.55	0.60
0.69 s	L _T	7	7	7	7	5	7	7	7
	β	0.85	0.85	0.80	0.85	1.00	0.75	0.70	0.70
0.79 s	L _T	7	7	7	7	7	7	7	7
	β	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.75

3.3. IC Estimation Performance Comparison with Conventional Method

By comparing the proposed method with the conventional method, we tried to confirm the validity of the algorithm and assumption based on the quasi-steady interval proposed in this paper. Before applying the conventional method, the IC estimation performance depends on α ; thus, we compared the performance of conventional method according to α with that of the proposed method using the specific parameters determined from previous simulations. In this numerical experiment, MAC_{total} was also used to accurately compare the performance of estimation.

MAC_{total} was calculated by applying the proposed method to 1680 speech samples of the test set using STFT with a 25-ms Hamming window, 10-ms hop length, and 512-point FFT according to the reverberant environment. The MAC_{total} for speech was averaged to compare the performance of the

two methods numerically. By applying the proposed method in the same way as the conventional method, the averaged MAC_{total} was calculated according to α , and its results are represented by the solid lines in Figure 4. The results of applying the previously determined L_T and β to the proposed method are represented by the dotted lines. The averaged MAC_{total} was a low value when α was close to zero and one, which means that the IC of reverberant speech was inaccurately estimated. A large MAC_{total} means a high IC estimation performance; therefore, α for the maximum value of averaged MAC_{total} , which is an indicator of overall performance used in determining L_T and β , was the optimal value. It was 0.60 in 0.37 s of RT_{60} , 0.50 in 0.69 s of RT_{60} , and 0.25 in 0.79 s of RT_{60} . In each reverberation environment, the maximum values of the averaged MAC_{total} of the conventional method were 0.0008, 0.0255, and 0.0524 lower than the averaged MAC_{total} of the proposed method with the optimal parameters. Thus, the estimation performance of the proposed method was better in each reverberation environment compared to the conventional method with optimal α .

Figure 4. Averaged MAC_{total} using short-time Fourier transform (STFT) with a 25-ms Hamming window, 10-ms hop length, and 512-point FFT for 1680 speech samples of conventional and proposed methods in (a) 0.37 s, (b) 0.69 s, and (c) 0.79 s of RT_{60} .

In the previous numerical experiment of determining parameters, L_T and β with STFT using a longer window and longer hop lengths were determined. Based on these parameters, the performance of the proposed method was also compared with that of the conventional method using STFT with a 64-ms Hamming window, 16-ms hop length, and 1024-point FFT. The average MAC_{total} according to α was calculated for the same 1680 speech samples, and the results obtained for the conventional and proposed methods are shown in Figure 5. They are similar to the results obtained using short STFT variables, and α for the maximum of the averaged MAC_{total} was the optimal value for the conventional method. It was 0.65 in 0.37 s of RT_{60} , 0.40 in 0.69 s of RT_{60} , and 0.25 in 0.79 s of RT_{60} . In each reverberation environment, the maximum values of the averaged MAC_{total} of the conventional method were 0.0236, 0.0494, and 0.0017 higher than the averaged MAC_{total} of the proposed method with the optimal parameters. Therefore, when large STFT variables were applied, the conventional method achieved better performance for the environment with reverberation.

Although the proposed method with the optimal L_T and β had higher MAC_{total} than the conventional method using STFT with 512-point FFT, the MAC_{total} only represents the overall estimation performance of the IC. By comparing the ICs for each time–frequency bin of reverberation speech, we tried to identify the characteristics of the bins where each method accurately estimated the coherence. The IC estimation method was applied to an utterance with 0.79 s of RT_{60} environment where the difference of the MAC_{total} was largest.

Spectrograms of the clean speech and the reverberant speech using STFT with a 25-ms Hamming window, 10-ms hop length, and 512-point FFT are represented as (a) and (b) in Figure 6. Due to the effect of reverberation, the time–frequency bins of the direct speech were not exactly known in the spectrogram of the reverberation speech. In this case, the clean speech was known; thus, the bins of the direct speech could be determined using it, and the binary mask of the direct speech is represented as (c) in Figure 6. For the reverberant speech, the ICs were calculated by applying the proposed method

and the conventional method, and each result was as shown in (d) and (e) of Figure 6. It is difficult to visually compare the difference between the estimated coherences. To solve this problem, the difference in the coherence accuracy (Γ^2 in the direct speech bins and $1 - \Gamma^2$ in the diffuse speech bins) of each method was calculated as shown in (f) of Figure 6. If the difference of the coherence accuracy is greater than zero, the proposed method estimated the IC more accurately; otherwise, the conventional method estimated the IC more accurately. It can be seen that the light-colored bins, which were larger than zero, almost matched the direct speech mask in (c). Among the figures, the difference in coherence accuracy was larger at the boundary between direct and diffuse speech bins. In most of the diffuse speech bins, the conventional method estimated the IC accurately. As a result, the proposed method accurately estimated the IC of the direct speech bins including the boundaries, but the conventional method accurately estimated the diffuse speech bins.

Figure 5. Averaged MAC_{total} using STFT with a 64-ms Hamming window, 16-ms hop length, and 1024-point FFT for 1680 speech samples of conventional and proposed methods in (a) 0.37 s, (b) 0.69 s, and (c) 0.79 s of RT₆₀.

Figure 6. The spectrogram of (a) clean speech and (b) reverberant speech with 0.79 s of RT₆₀, calculated by STFT with a 25-ms Hamming window, 10-ms hop length, and 512-point FFT. From the clean speech, (c) the binary mask of direct speech was determined. The interaural coherence (IC) of time–frequency bins was estimated by (d) the proposed method and (e) the conventional method with optimal parameters. (f) The difference in coherence accuracy between each method.

3.4. Computation Time

Various speech preprocessing and speech recognition techniques are performed in real time; hence, a low computation time of IC estimation is required to add a coherence-based technique. The computation time was verified for 16-kHz speech samples with various lengths from 0 to 8 s and compared for the two IC estimation methods. The IC of the conventional method was calculated sequentially over time. Based on MATLAB, the average computation time was calculated 100 times for each speech length, and the results are shown in Figure 7.

Figure 7. Calculation time according to speech length of reference and proposed calculation methods.

Increasing the length of the applied speech increased the computation time, and the proposed method always showed less computation time. In this numerical experiment, the proposed method was found to be superior to the conventional method in terms of computation time, and as the speech became longer, the computation time difference also increased.

4. General Discussion

Using MAC_{total} based on time–frequency bins with a 25-ms Hamming window, 10-ms hop length, and 512-point FFT, the optimal L_T and β according to frequency were determined for each reverberation environment. Based on theoretical analysis with static locations of microphones and speaker, the direct speech has a steady value within the quasi-steady interval; thus, the IC of direct speech is close to one. The diffuse speech is an infinite sum of the reflected speech having a different propagating path; therefore, the diffuse speech between two microphones is uncorrelated in the quasi-steady interval, and the IC of diffuse speech appears close to zero. In a real environment, the speaker may move, or the position of the microphone may change. Equations (9) and (10) are expressions of reverberant speech in the interval of quasi-steady state, which was represented by 0.085–0.135 s in the simulation. Thus, when the motion of microphones and speaker is slow enough to be regarded as static in the quasi-state time scale, it is based on a proper assumption. However, the room impulse responses in the simulation were measured in the static environment, which are limited when showing the performance in the dynamic environment in this paper. To estimate the IC similar to the ideal coherence, the optimal L_T was the average quasi-steady interval of speech, and the IC estimation was inaccurate for the section beyond the quasi-steady state of speech. The optimal β was the weight to compensate for the steady state in the quasi-steady interval; thus, an increase in the optimal β can also be seen as an increase in the quasi-steady interval length. Since the fundamental frequency differs according to gender, we assumed that the optimal parameters according to gender would differ, but this was not shown in the numerical experiment. This means that the quasi-steady interval is independent of gender, and gender was not considered when determining the optimal parameters.

In contrast, the variation in optimum parameters according to frequency was shown. The optimal L_T and β decreased as the frequency increased, and this tendency of the optimal β was not shown in the case of high RT_{60} . In real human speech, unvoiced sound is mainly distributed in high frequency, voiced sound is distributed in low frequency, and unvoiced sound is usually shorter than voiced sound. The quasi-steady state for unvoiced sound is shorter than that for voiced sound; hence, the quasi-steady state appears shorter at high frequency, which is the same as the tendency of the optimum parameters. This result is consistent with the theoretical analysis that the IC estimation should be applied within the quasi-steady interval. In addition, we assumed that the length of the quasi-steady state was constant according to speech regardless of the reverberation environment, but the optimal L_T and β increased, and β was close to one as the RT_{60} increased in the numerical experiment. This means that the quasi-steady interval became longer as RT_{60} increased, and the quasi-steady interval included not only direct speech but also reflections, which had a small time delay with direct speech and similar energy to that of direct speech. Moreover, it was difficult to accurately distinguish between direct and diffuse speech; thus, the IC became inaccurate, and the overall MAC_{total} increased as RT_{60} increased from 0.37 s to 0.79 s using the optimal parameters, as in the numerical experiment. To sum up, for the proposed IC estimation method, it is appropriate to calculate along the quasi-steady interval; therefore, L_T and β should decrease as the frequency increases, and too large a value of RT_{60} causes incorrect estimation of the IC.

The length of the quasi-steady interval was constant in the same reverberation environment, but the optimal L_T became smaller when a longer window and a longer hop length were applied. In the quasi-steady interval for 0.79 s of RT_{60} with the same β as one, the time interval length of seven spectrogram bins with a 25-ms window was 85 ms, but the time interval of nine spectrogram bins with a 64-ms window was 192 ms; thus, the IC was estimated for the longer time interval. The estimation performance was relatively low when larger STFT variables were applied in the numerical experiment, which means that the calculated time interval was longer than the quasi-steady interval, and it provided incorrect IC values. The reason for the different quasi-steady intervals is that it was difficult to represent time–frequency bins of the same length as the quasi-steady interval for speech as the STFT variables became larger. On the contrary, the conventional method was affected only by the hop length because all the previous data were used; hence, the optimal α appeared constant regardless of the STFT variable based on the short difference in the hop length of 6 ms. Therefore, it is very important to apply the IC estimation method proposed in this study within the time–frequency bins that match the accurate quasi-steady interval length, whereas the conventional method is not affected by window length.

The performance of the conventional method and the proposed method was determined by MAC_{total} , and the proposed method had 0.0008, 0.0255, and 0.0524 higher MAC_{total} than the conventional method with the optimal α when the STFT variables, which are mainly used in speech processing, were applied. The reason for this result can be elicited from the PSD estimation equation. The conventional estimation equation, which uses all the previous time data, is similar to a type of IIR filter that shows a nonlinear phase delay with respect to the frequency and distortion in comparison with the original signal. Thus, the conventional method also estimates PSD by applying an IIR filter to $X_i^* X_j$, resulting in distortion in comparison to reverberation speech. The proposed method in this paper is an FIR filter type that shows a linear phase delay or zero phase delay with respect to the frequency and no distortion. The estimation of the PSD using the proposed method does not lead to distortion of the PSD; thus, it provides more accurate IC.

However, in the previous numerical experiment, the proposed method accurately estimated the IC of the direct speech, but it had worse IC estimation performance of diffuse speech compared to the conventional method. This means that applying the calculation within a finite interval, as in the proposed method, was suitable for direct speech but not for diffuse speech. The FIR filter form of the proposed method with quasi-steady interval compensated for the problem of speech that is continuously varying over time; thus, it was possible to accurately estimate the IC of the direct speech. Also, it quickly reflected changes of speech; thus, the boundary bins between the direct and diffuse speech with

the proposed method had more accurate IC than the conventional method. In a high-RT₆₀ environment, the energy of diffuse speech decreased slowly with respect to time; therefore, diffuse speech was also quasi-steady within the quasi-steady interval of direct speech. Since the IC of diffuse speech can appear close to one when the proposed method recognizes diffuse speech as direct speech, the estimation performance for diffuse speech was lower than that of the conventional method, as shown in the test experiment. Nevertheless, the IC of direct speech was more accurately estimated, and the overall performance was improved. Moreover, the computation time between the proposed and conventional methods was compared according to speech length. In the calculation of proposed coherence, all the computation was done in a parallel manner simply by batch multiplication of all recorded speech data with filter weights. On the other hand, in the conventional method using the IIR filter in Equation (3), both speech data and past PSD were required in a recursive fashion. Hence, the computation was sequential, which took more time compared to the proposed method.

5. Conclusions

This paper proposed an IC estimation method that applies the exponential weight with β to L_T time–frequency bins of the same length as the quasi-steady interval of speech and mean accuracy of coherence (MAC) to quantify estimation performance. The optimal L_T and β according to gender, frequency, and reverberation time were determined, but the results showed that gender does not affect optimal parameters. The speech characteristics change according to age and language; thus, the effects of age and language especially in correlation with their spectrograms need to be investigated. The optimal L_T and β values were determined as the maximum point of the histogram with MAC_{total} in specific reverberation environments with 0.37 s, 0.69 s, and 0.79 s of RT₆₀. The optimal parameters for only specific reverberation environments were proposed in this paper. However, based on the theoretical analysis and the tendency of the numerical experiment results, the optimal parameters in any reverberation environment can be determined as L_T and β with small values at high frequency and low reverberation time. The IC estimation of reverberant speech with STFT variables, which are mainly used for speech processing, using the proposed method with optimal parameters showed 0.0008, 0.0255, and 0.0524 larger MAC_{total} than the conventional method; thus, the proposed method achieved better estimation performance. While the conventional method is in the form of an IIR filter, the proposed method is in the form of an FIR filter; therefore, it provides the accurate IC with less distortion of the estimated PSD and less computation time. However, the FIR filter needs more coefficients and, therefore, it may require more computations than the IIR filter. In the simulations, the difference in IC estimation performance was small, but intensive improvement was found in the vicinities of the border between direct and diffused speech, which implies that the proposed method has advantage in the rapid change of speech data because past data do not affect current PSDs in the FIR filter. On the other hand, the estimation performance for diffuse speech was degraded compared to the conventional method. Some diffuse speech bins have lower energy than direct speech bins. Since the importance of the coherence estimation accuracy for these bins is small, the proposed method that estimates direct speech more accurately is more suitable.

Author Contributions: Conceptualization, S.-H.K. and Y.-H.P.; methodology, S.-H.K.; software, S.-H.K.; validation, S.-H.K.; formal analysis, S.-H.K.; investigation, S.-H.K.; resources, S.-H.K.; data curation, S.-H.K.; writing—original draft preparation, S.-H.K.; writing—review and editing, S.-H.K. and Y.-H.P.; visualization, S.-H.K.; supervision, Y.-H.P.; project administration, Y.-H.P.; funding acquisition, Y.-H.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: This research was supported by the “Human Resources Program in Energy Technology” of the Korea Institute of Energy Technology Evaluation and Planning (KETEP), the Ministry of Trade, Industry & Energy, Korea (No. 20184030202000), and the “Research Project for Railway Technology” of the Korea Agency for Infrastructure Technology Advancement (KAIA).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Vincent, E.; Watanabe, S.; Nugraha, A.A.; Barker, J.; Marxer, R. An analysis of environment, microphone and data simulation mismatches in robust speech recognition. *Comput. Speech Lang.* **2017**, *46*, 535–557. [\[CrossRef\]](#)
2. Jourjine, A.; Rickard, S.; Yilmaz, O. Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures. In Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing, Istanbul, Turkey, 5–9 June 2000; Proceedings (Cat. No. 00CH37100). IEEE: New York, NY, USA.
3. Mandel, M.I.; Weiss, R.J.; Ellis, D.P. Model-based expectation-maximization source separation and localization. *IEEE Trans. Audio Speech Lang. Process.* **2009**, *18*, 382–394. [\[CrossRef\]](#)
4. Zurek, P.M. The Precedence Effect. In *Directional Hearing*; Springer: Berlin/Heidelberg, Germany, 1987; pp. 85–105.
5. Litovsky, R.Y.; Colburn, H.S.; Yost, W.A.; Guzman, S.J. The precedence effect. *J. Acoust. Soc. Am.* **1999**, *106*, 1633–1654. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Wang, D.; Brown, G.J. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*; Wiley-IEEE Press: New York, NY, USA, 2006.
7. Jacobsen, F.; Roisin, T. The coherence of reverberant sound fields. *J. Acoust. Soc. Am.* **2000**, *108*, 204–210. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Kuttruff, H. *Room Acoustics*; Crc Press: Boca Raton, FL, USA, 2016.
9. Faller, C.; Merimaa, J. Source localization in complex listening situations: Selection of binaural cues based on interaural coherence. *J. Acoust. Soc. Am.* **2004**, *116*, 3075–3089. [\[CrossRef\]](#) [\[PubMed\]](#)
10. Alinaghi, A.; Wang, W.; Jackson, P.J. Spatial and coherence cues based time-frequency masking for binaural reverberant speech separation. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; IEEE: New York, NY, USA.
11. Jeub, M.; Schafer, M.; Esch, T.; Vary, P. Model-based dereverberation preserving binaural cues. *IEEE Trans. Audio Speech Lang. Process.* **2010**, *18*, 1732–1745. [\[CrossRef\]](#)
12. Jeub, M.; Vary, P. Binaural dereverberation based on a dual-channel wiener filter with optimized noise field coherence. In Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, 14–19 March 2010; IEEE: New York, NY, USA.
13. Schwarz, A.; Brendel, A.; Kellermann, W. Coherence-based dereverberation for automatic speech recognition. In *Processings DAGA*; Citeseer: Forest Grove, OR, USA, 2014.
14. Welch, P. The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Trans. Audio Electroacoust.* **1967**, *15*, 70–73. [\[CrossRef\]](#)
15. Allen, J.B.; Berkley, D.A. Image method for efficiently simulating small-room acoustics. *J. Acoust. Soc. Am.* **1979**, *65*, 943–950. [\[CrossRef\]](#)
16. Garofolo, J.S.; Lamel, L.F.; Fisher, W.M.; Fiscus, J.G.; Pallett, D.S. DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon Tech. Rep. N* **1993**, 93.
17. Jeub, M.; Schafer, M.; Vary, P. A binaural room impulse response database for the evaluation of dereverberation algorithms. In Proceedings of the 2009 16th International Conference on Digital Signal Processing, Santorini, Greece, 5–7 July 2009; IEEE: New York, NY, USA.
18. Rix, A.W.; Beerends, J.G.; Hollier, M.P.; Hekstra, A.P. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City, UT, USA, 7–11 May 2001; Proceedings (Cat. No. 01CH37221). IEEE: New York, NY, USA.