acoustics

MDPI

Article

A Novel Scheme for Single-Channel Speech Dereverberation

Nikolaos Kilis † and Nikolaos Mitianoudis *,†

Electrical and Computer Engineering Department, Democritus University of Thrace, 67100 Xanthi, Greece
* Correspondence: nmitiano@ee.duth.gr; Tel.: +30-25410-79572
† These authors contributed equally to this work.

Received: 11 June 2019; Accepted: 29 August 2019; Published: 5 September 2019

Abstract: This paper presents a novel scheme for speech dereverberation. The core of our method is a two-stage single-channel speech enhancement scheme. Degraded speech obtains a sparser representation of the linear prediction residual in the first stage of our proposed scheme by applying orthogonal matching pursuit on overcomplete bases, trained by the K-SVD algorithm. Our method includes an estimation of reverberation and mixing time from a recorded hand clap or a simulated room impulse response, which are used to create a time-domain envelope. Late reverberation is suppressed at the second stage by estimating its energy from the previous envelope and removed with spectral subtraction. Further speech enhancement is applied on minimizing the background noise, based on optimal smoothing and minimum statistics. Experimental results indicate favorable quality, compared to two state-of-the-art methods, especially in real reverberant environments with increased reverberation and background noise.

Keywords: one-microphone dereverberation; reverberant-noisy speech enhancement; sparse coding; spectral subtraction

1. Introduction

In many everyday places, the presence of reverb may undermine our listening experience. Reverb is created by the sound waves' reflections on the walls of a closed room. Depending on the surfaces' sound absorbing properties, as well as the room's dimensions, reverb's effects can vary from pleasing to extremely annoying. The latter can occur in enclosed spaces that have not been explicitly designed for sound reproduction purposes. To control the effects of reverb, scientists can recommend a variety of strategies, including sound-absorbing panels, curtains, carpet floors, and sound traps. Nonetheless, it might not always be possible to improve the room's acoustics using sound-absorbing material and sound traps (e.g., historical buildings, churches) [1], as this may harm the room's interiors. In this case, we need to use signal processing algorithms to filter out the introduced reverb from the recordings. This whole procedure is often termed dereverberation. Moreover, additional speech degradation may occur during room recordings. In these cases, the other usual suspects apart from reverberation are echo and background noise. Echoes are distinct strong reflections from the room's surface that are distinctly audible by the human ears. The source of background noise may vary from simple air dissipation hiss inside the room to other sound sources outside the room (cars, rain, thunder, winds, etc.).

The problem of room dereverberation is known to the audio processing community. Hitherto, many methods have been proposed to improve or remove room reverberation. The idea of using multi-channel inputs and the delay-sum and filter-sum beamformer for dereverberation was introduced by Elko [2]. Gaubitch [3] showed that the performance of the beamforming-based approach depends only on the distance between the microphone array and the sound source, the distance

between the microphones, and the number of microphones. Thus, beamforming-based approaches can offer only medium improvement in room dereverberation. The concept of filtering and sparsifying the Linear-Prediction (LP) residual of reverberant speech was introduced in [4]. In [5], the authors sparsified the LP residual by kurtosis optimization working either on the time-domain signal or in a sub-band representation, as expressed by the Modulated Complex Lapped Transform (MCLT). The authors also introduced a multi-channel dereverberation variant. In [1], Tonelli et al proposed an equivalent maximum-likelihood sparsification method for the LP residual, assuming supergaussian priors. A multi-channel extension was also proposed. For a more complete review on previous approaches on dereverberation and speech enhancement, in general, the reader is referred to [6–9].

In this paper, the authors propose a single-channel dereverberation algorithm improving the approach of Wu and Wang [7], mainly by improving their smoothing function using an estimated real Room Impulse Response (RIR) envelope and sparsifying the LP residual using a sparse coding method, known as the K-SVD algorithm [10]. Last but not least, the proposed scheme addresses the problem of background noise suppression. Thus, the novel element of the approach is the use of the following elements in a single architecture: (a) sparsification of the residual via the K-SVD algorithm; (b) noise removal via minimum statistics; (c) a more flexible hand clap/RIR scheme; and (d) the time-envelope extracted from the hand clap/RIR signal that carries the energy from the mixing time to T_{60} and spectral subtraction of this energy.

2. Previous Work

In this section, we briefly outline some previous work on dereverberation that is important and constitutes the basis for our proposed method.

2.1. Wu and Wang Method

Wu and Wang (WW) [7] proposed a two-stage model to deal initially with coloration or the early reflections and the long-term reverberation. In the first stage, their model estimated an inverse filter of the room impulse response by maximizing the kurtosis of the linear prediction residual. When passing a room impulse response through the inverse filter and after multiple iterations, the reduction of coloration and the increase of the Signal to Reverberant Ratio (SRR) value was clearly demonstrated [7]. Replacing the room impulse response with reverberant data, the perceived acoustic result was similar to moving the sound source closer to the receiver. However, it is also clear that this stage does not improve the reverberation tail. To alleviate this, the inverse-filtered reverberant data go through a second stage. Wu and Wang [7] assumed that the power spectrum of late reverberation components is a smoothed and shifted version of the power spectrum of the inverse-filtered speech, i.e.,:

$$|S_{late}(\tau,k)|^2 = \gamma w(\tau-\rho) * |S_{inv}(\tau,k)|^2 \tag{1}$$

where $|S_{late}(\tau,k)|^2$ and $|S_{inv}(\tau,k)|^2$ are the short-term power spectra of the late reverberation components and the inverse-filtered speech, respectively. Indexes τ and k refer to the time frame and frequency bin, respectively. The symbol $*$ denotes the operation of convolution; γ is a scaling factor; $w(\tau)$ is a smoothing function; and ρ is the shift delay that indicates the relative delay of the late reflection components. A Rayleigh distribution was chosen as their smoothing function:

$$w(\tau) = \begin{cases} \frac{\tau+\alpha}{\alpha^2}\exp\left[\frac{-(\tau+\alpha)^2}{2\alpha^2}\right] & \text{, if } \tau > -\alpha \\ 0 & \text{, otherwise} \end{cases} \tag{2}$$

Then, they estimated the power spectrum of the anechoic speech by subtracting the power spectrum of the late reverberation components from the power spectrum of the inverse-filtered speech. More specifically,

$$|S_x(\tau,k)|^2 = |S_{inv}(\tau,k)|^2 \max\left[\frac{|S_{inv}(\tau,k)|^2 - \gamma w(\tau-\rho) * |S_{inv}(\tau,k)|^2}{|S_{inv}(\tau,k)|^2}, \epsilon\right] \quad (3)$$

where $\epsilon = 0.001$ is the energy floor and corresponds to the maximum attenuation of 30 dB. Finally, the silent gaps between speech utterances were detected and attenuated by 15 dB.

2.2. Spendred Algorithm

The Spendred algorithm was created by Doire et al. [8] in a completely different philosophy from the WW algorithm. Spendred is an online single-channel speech-enhancement method designed to improve the quality of speech, degraded by reverberation and noise. Based on an autoregressive model for the reverberation power and on a Hidden Markov model for clean speech production, a Bayesian filtering formulation of the problem was derived, and online joint estimation of the acoustic parameters and mean speech, reverberation, and noise powers were obtained in Mel-frequency bands. From these estimates, a real-valued spectral gain was derived, and spectral enhancement was applied in the Short-Time Fourier Transform (STFT) domain. More details about the algorithm description can be found in [8,9].

2.3. The K-SVD Algorithm

K-SVD is a machine learning algorithm that can learn a dictionary of atoms that can be used to create sparse representations of signals [10]. K-SVD serves as a generalization of the classic k-means method of clustering. The principle behind K-SVD is an iterative alternation between two procedures: (a) sparse coding of the input data using an estimated dictionary of atoms; (b) update the dictionary atoms in order to approximate the data more accurately.

The main concept is to express a signal $\mathbf{y} \in \mathcal{R}^N$ using a linear combination of a overcomplete set of dictionary atoms $\mathbf{d}_i$, where $i = 1, \ldots, K$ and $K >> N$. Assuming that $D = [\mathbf{d}_1; \ldots; \mathbf{d}_K]$ and $\mathbf{x} \in \mathcal{R}^K$ is a vector that contains all the coefficients of the aforementioned linear combination, this decomposition can be modeled via:

$$\mathbf{y} = D\mathbf{x} \quad (4)$$

The main objective of sparse coding and this decomposition is to estimate dictionary atoms that lead to a sparse vector $\mathbf{x}$, i.e., a vector that has only a few of its values active and the others are zero. Sparsity can be measured using the $L0$-norm of $\mathbf{x}$, i.e., $||\mathbf{x}||_0$. Hence, the problem of sparse coding is two-fold: (a) find dictionary atoms that lead to sparse representations; (b) estimate the sparse representation of $\mathbf{x}$ given the estimated dictionary D. The sparse coding problem is usually posed as the following optimization problem:

$$\min_{D,\mathbf{x}} ||\mathbf{x}||_0, \qquad \text{s.t. } \mathbf{y} = D\mathbf{x} \quad (5)$$

The K-SVD algorithm [10] slightly rephrases the above optimization problem and solves the following:

$$\min_{D,\mathbf{x}} ||\mathbf{y} - D\mathbf{x}||_F^2, \qquad \text{s.t. } ||\mathbf{x}||_0 < T_0 \quad (6)$$

The algorithm begins with a random initialization of the dictionary atoms. Using any pursuit algorithm, such as the Orthogonal Matching Pursuit (OMP) algorithm, and the dictionary algorithm, the first part of the algorithm estimates the atom coefficients x_i that constitute vector $\mathbf{x}$. The second part consists of updating the dictionary atoms. For every dictionary atom k, the algorithm

finds a set ψ_k of examples $\mathbf{x}_i$ from the dataset that use this atom. Then, the overall representation error matrix E_k is estimated via:

$$E_k = Y - \sum_{j \neq k} \mathbf{d}_j x_T^j \tag{7}$$

where x_T^j is the j^{th} row of a matrix X that contains all vectors $\mathbf{x}$. The matrix E_k is restricted by keeping only the columns that correspond to ψ_k, thus giving the matrix E_k^R. Singular-value decomposition is applied to $E_k^R = U\Delta V^T$. The first column of matrix U forms the new dictionary atom $\mathbf{d}_j$, while the first column of V, multiplied by $\Delta(1,1)$, forms the new coefficient x_R^k. The steps in the first and second part iterate until convergence of the dictionary atoms.

2.4. Denoising with Minimal Statistics

In [11], Martin proposed a robust algorithm to estimate the noise power density, in order to be used for spectral subtraction. A sliding window FFT was used to create the spectrogram of the input noisy signal $x(n)$. This creates the spectrogram $X(\tau, k)$, where τ and k refer to the time frame and frequency bin, respectively. The minimum statistics noise tracking method is based on the fundamental hypothesis that during a speech segment, there exist multiple pauses and silence periods between words. In theory, the signal energy in these pauses should be zero in the noiseless case. Nonetheless, in the noisy signal case, the power of these pausing periods should correspond to the power of the noise. Thus, this provides a method to track the noise power profile. Since the noise is traditionally modeled as additive, the estimated noise power profile can be subtracted from the signal power profile, thus performing denoising by spectral subtraction.

The extracted spectrogram is filtered in order to create a smoothed periodogram $P(\tau, k)$ via:

$$P(\tau, k) = \alpha P(\tau - 1, k) + (1 - \alpha)|Y(\tau, k)|^2 \tag{8}$$

A minimum power tracking algorithm traces the minimum power noise profile. An optimal value $\alpha = 0.85$ can be calculated by setting the variance of $P(\tau, k)$ equal to the variance of a moving average estimator with 0.2-s windows. The noise spectral estimate $P_N(\tau, k)$ is given by selecting the minimum value within a sliding window of 96 consecutive values of $P(\tau, k)$. This profile is updated regardless of the presence of speech or not.

Unfortunately, keeping α constant is not an optimal choice for efficient noise level estimation, because it tends to favor small noise levels, it cannot adapt quickly to noise level fluctuations, and noise levels tend to have large variance [11]. Thus, in [11], the constant α was replaced by $\alpha(\tau, k)$ in order to make it time and frequency dependent. To find an optimal value for α, we minimized the conditional mean squared error $\mathcal{E}\{(P(\tau, k) - P_N(\tau, k))^2 | P(\tau - 1, k)\}$. The optimal value for α is given by:

$$\alpha_{opt}(\tau, k) = \frac{1}{1 + (P(\tau, k)/P_N(\tau, k) - 1)^2} \tag{9}$$

To avoid possible shortcomings during the update of α_{opt}, an upper and a lower threshold were imposed, and it was also smoothed over time to give more stable versions of α_{opt}. Consequently, an efficient algorithm keeps track of the minimum noise level $P_N(\tau, k)$, which will be used during the spectral subtraction case. For more details on the method, one can always refer to [11].

3. The Proposed Method

The basic structure of the proposed dereverberation scheme includes the following stages: (a) sparsification; (b) RIR envelope estimation; and (c) spectral subtraction. Figure 1 depicts the proposed dereverberation system in full.

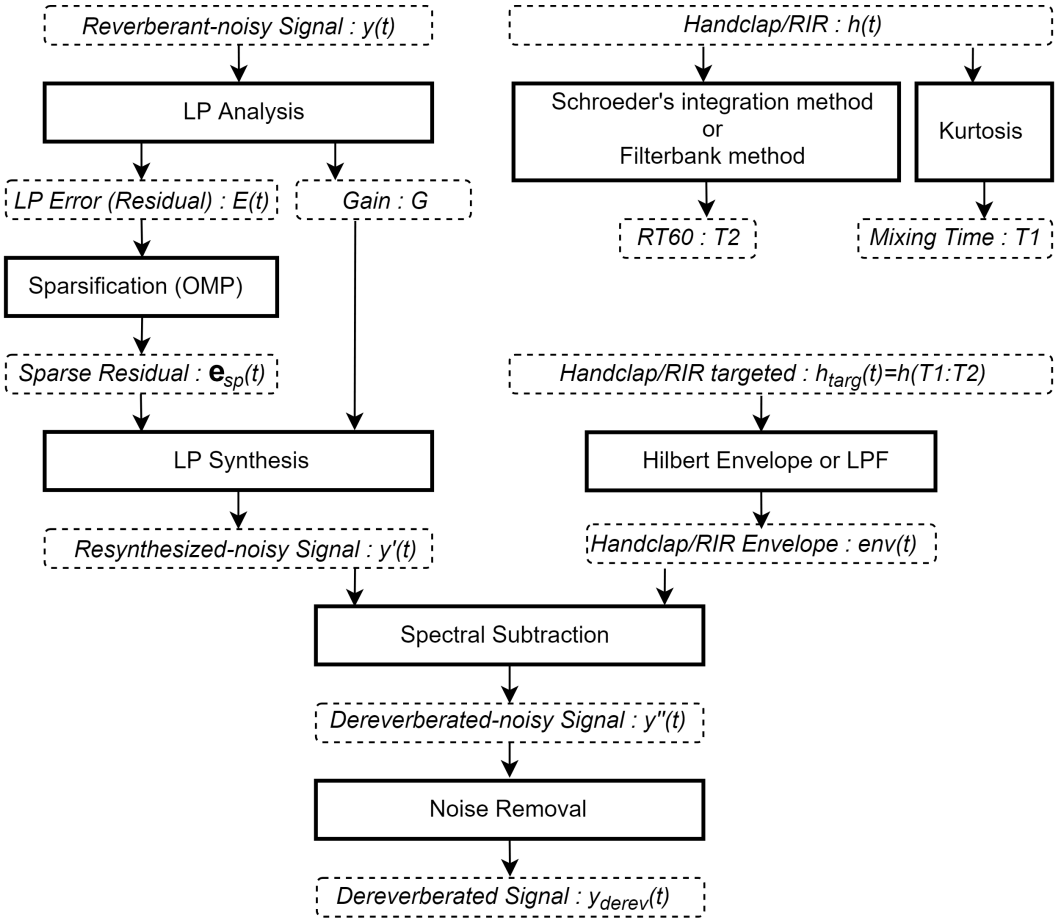


Figure 1. The proposed dereverberation system. LP: Linear-Prediction; OMP: Orthogonal Matching Pursuit; RIR: Room Impulse Response.

3.1. Sparsification

In this section, we attempt to sparsify the LP residual of the reverberant signal. The LP residual of clean speech has higher kurtosis than that of reverberant speech. Bearing this in mind, we can use a sparse coding algorithm on the LP residual to make it more sparse, thus increasing its kurtosis, i.e., removing the effects of reverberation on the audio signal. Signal sparsification is performed by estimating an overcomplete dictionary of audio bases using a sparse coding algorithm, such as the K-SVD algorithm, which was mentioned earlier.

More specifically, we used about an hour of audio samples, containing mainly speech. The samples were divided into about 11.6-ms frames, as commonly considered stationary for speech signals. We used the K-SVD algorithm to estimate an overcomplete set of audio bases. The number of estimated bases was selected as double the frame size. For 44.1-kHz audio, the frame size is typically 512 samples, and the number of extracted bases is 1024. This set of bases D was estimated only once and was then stored to be used for the sparsification of the reverberant speech LP residual.

We used a quite long LP analysis of 31 coefficients, in order to capture the signal's structure in detail. The reverberant signal was segmented into 11.6 ms frames (512 samples at a 44.1-kHz sampling rate) with 50% overlap. The LP residual $e(n)$ was estimated using the estimated 31 LP coefficients. A sparse representation of the LP residual was estimated using the overcomplete bases set and the technique, called Orthogonal Matching Pursuit (OMP). This technique is an approximation method that attempts to find the best bases from the bases' set that approximate better the signal to be analyzed. OMP also requires an error threshold to be defined, after which OMP terminates. Here, we can set this error threshold slightly higher than normal in order to produce a sparser representation $\mathbf{E}_{sp}$ of the LP residual. A typical value used in our experiments was 5×10^{-3}. Since the input signal was

normalized to unit variance, this threshold was independent of the signal's variance. The estimated sparse representation $\mathbf{E}_{sp}$ was used to resynthesize the sparsified LP residual via:

$$\mathbf{e}_{sp} = D\mathbf{E}_{sp} \tag{10}$$

The sparsified LP residual signals $\mathbf{e}_{sp}$ along with the estimated LP gain G were used to resynthesize the audio signal $y'(t)$, which now featured less reverb, due to the sparsification and filtering out of the LP residual. In Figure 2, the effect of LP residual sparsification is depicted for a speech signal recorded in a real university lecture room. It is more than evident that the algorithm achieved the desired sparsification effect, but this can also be verified by kurtosis measurements. The original LP residual had a kurtosis value of 9.23, while the sparsified LP residual had a kurtosis value of 14.68.

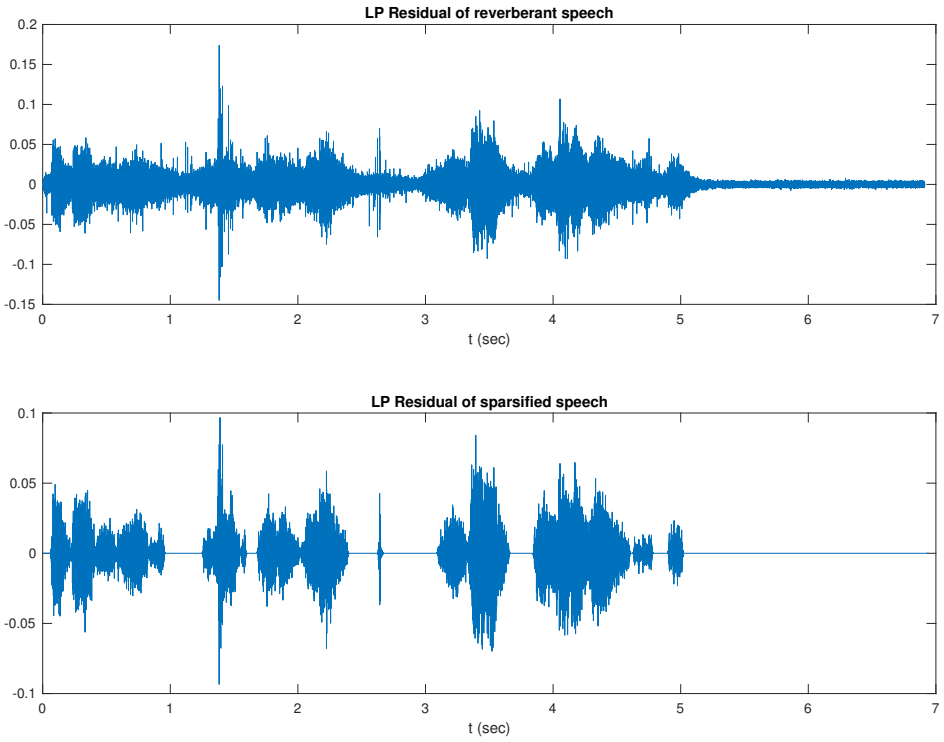


Figure 2. The effect of sparsification on the reverberant speech residual.

3.2. Rir Envelope Estimation

To estimate several properties of the RIR, we used a hand clap $h(t)$ that was recorded in the same room. The algorithm separated the hand clap into small frames (11.5 ms) with no overlap and calculated the Normalized Kurtosis (NK) of each frame, using the formula:

$$NK = \frac{\mathcal{E}\{(h - m_h)^4\}}{\sigma_h^4} - 3 \tag{11}$$

where $\mathcal{E}\{\cdot\}$ is the expectation operator, μ_h is the mean, and σ_h is the standard deviation of $h(t)$. When the calculated NK was around zero, we assumed that the signal followed a Gaussian distribution. The mixing time was obtained from the time difference between the highest value of normalized kurtosis and the hand clap's starting point. An example of mixing time estimation and a hand clap recorded in a room is shown in Figure 3. The red line in Figure 3 depicts the normalized kurtosis for

each frame. The normalized kurtosis is a descriptor of the shape of a probability distribution and more specifically how far/close from the Gaussian distribution it is. We can deduct from Figure 3 that the silent and the low energy frames followed a Gaussian-like distribution, resulting in low normalized kurtosis values. On the other hand, as soon as the acoustic energy rises, we observed a peak in the value and consequently an abrupt reduction. Based on this observation, we were able to get an estimate for the mixing time. This method was initially proposed by Stewart and Sandler [12].

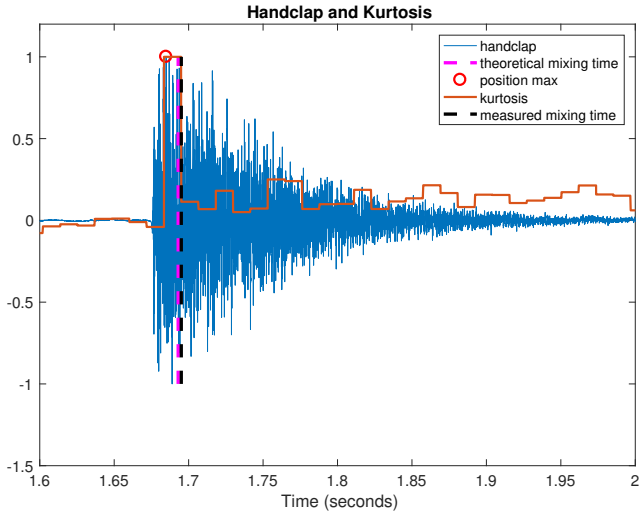


Figure 3. An example of mixing time estimation using the normalized kurtosis method and a hand clap for a frame size of 11.6 ms. The estimated mixing time and the theoretical mixing time are very close.

Another important issue is the choice of the frame size for the calculation of the NK. In general, choosing a large frame size leads to smoother kurtosis variability, but due to the large frame size, we had decreased accuracy in the estimation of the mixing time. On the other hand, choosing smaller frame sizes, we obtained better precision at estimating the mixing time, However, the variability of kurtosis became more erratic and may end up destroying the aforementioned kurtosis pattern at the beginning of the RIR. This can render the estimation of the mixing time totally impossible. In Figure 4, we can see the kurtosis fluctuations for frame sizes of 2.9, 5.8, 11.6, 23.2, and 46.4 ms, respectively. We can see that for larger frame sizes, kurtosis had less variability. The mixing time was estimated via the time lapse between the time when the normalized RIR rose >0.05 and the time kurtosis dropped from its peak value. Thus, for the frame sizes of 11.6, 23.2, and 46.4 ms, we obtained a mixing time estimate of 8.798, 8.798, and 3.2 ms, whereas the theoretical value was 6.7 ms. For the smaller frame sizes of 2.9 and 5.8 ms, the normalized kurtosis became far more noisy and erratic to keep track of significant changes. The 5.8 ms frame gave a similar 8.798 ms mixing time estimate, whereas the 2.9 ms frame yielded an erroneous mixing time of 0.9 ms. This implied that the three middle frame sizes were giving better estimates of the mixing time. Thus, we kept the middle frame size of 11.6 ms for the rest of the analysis to keep in line with the framing in other parts of the proposed algorithm.

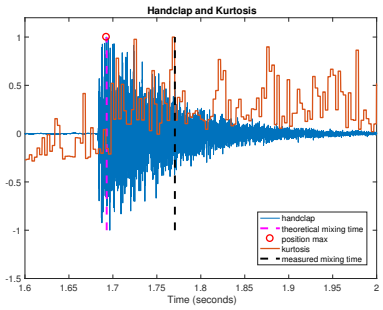
Older studies [13,14] showed that a hand clap differed from a room impulse response on three points:

- in the low frequency region
- there was a form of coloration in the frequency spectrum
- the overall spectral energy was different

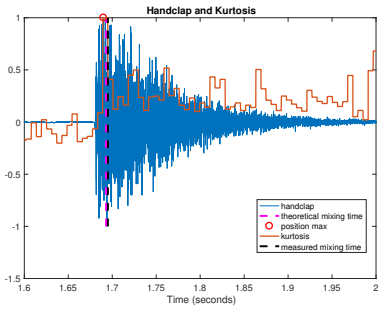
Late reverberation followed a Gaussian-like distribution, and its spectrum was almost similar to the spectrum of white noise [15]. It is, therefore, reasonable to assume that the same can be applied for the late room response part, due to a hand clap. Moreover, speech signals do not carry significant

energy in the low-frequency range, in which the hand clap differed. The experimental verification of the above assumptions has been shown in previous studies [16,17].

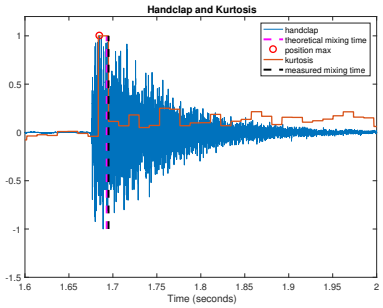
Given an estimate of the RIR or a hand clap, the user has a choice between two methods for calculating the reverberation time T_{60}: (a) the Schroeder backward integration [18] and (b) the filter bank method. The filter bank method filters the hand clap/RIR, according to the critical bands of hearing (Bark scale) and by using the same principle of Schroeder's backward integration estimates the T_{60} for each band. The next step uses the hand clap or the RIR to find the mixing time of the room.



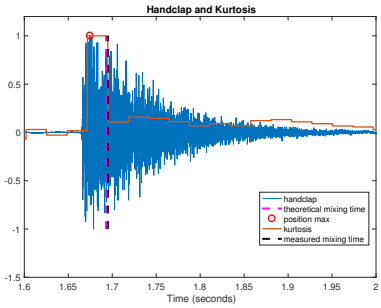
(**a**) Frame size 2.9 ms



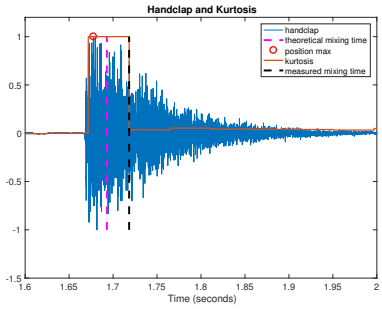
(**b**) Frame size 5.8 ms



(**c**) Frame size 11.6 ms



(**d**) Frame size 23.2 ms



(**e**) Frame size 46.4 ms

Figure 4. Mixing time estimation for different frame sizes.

Having the hand clap, the mixing time, and T_{60}, we formed a time-domain envelope $env(t)$, which will be used as a smoothing function in the next stage. Firstly, from the hand clap signal, we kept only the values between the mixing time and T_{60}. Then, we applied to the remaining values two simple envelope followers. The first used the analytical signal via the Hilbert transform, while the second used a simple low-pass filtering scheme. We also implemented the Rayleigh envelope, which was used in Wu and Wang's method [7], for comparison reasons. The user of the algorithm can choose between these three envelopes. An example of the three envelopes is shown in Figure 5.

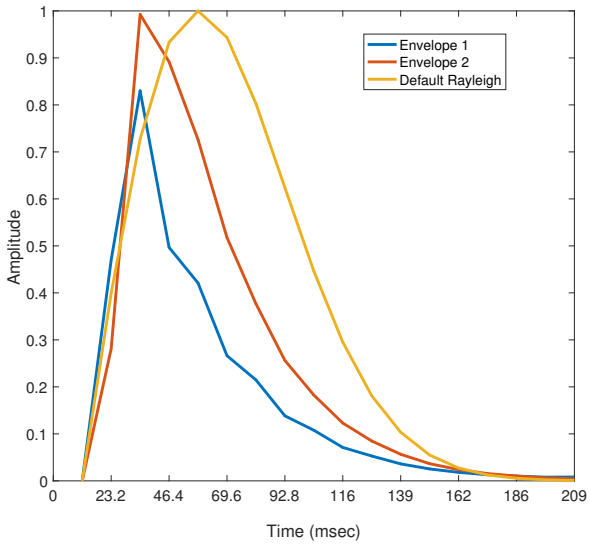


Figure 5. RIR envelope $env(t)$ estimation for a sample room. Envelope 1 uses the analytical signal via the Hilbert Transform, while Envelope 2 uses a simple low-pass filtering scheme.

3.3. Spectral Subtraction

After this stage, we further enhance the sparsified audio signal by suppressing the late reverberation through spectral subtraction. It is well known that the smearing effects of late reverberation force the speech signals to have a smoother spectrum in the time domain [7]. We also assumed that the power spectrum of late reverberation components was a smooth and shifted version of the power spectrum of the sparsified speech, i.e., (1). The difference lies in the shape of the smoothing function, which in our case was the envelope we calculated in the previous step. As a last step, we estimated the power spectrum of anechoic speech by subtracting the power spectrum of the late reverberation components from that of the sparsified speech, i.e., (3), which is now transformed to the following form:

$$|S_{y''}(\tau,k)|^2 = |S_{y'}(\tau,k)|^2 \max\left[\frac{|S_{y'}(\tau,k)|^2 - \gamma env(\tau-\rho) * |S_{y'}(\tau,k)|^2}{|S_{y'}(\tau,k)|^2}, \epsilon\right] \tag{12}$$

where $S_{y'}$ is the spectrogram of the sparsified signal $y'(t)$, $S_{y''}$ is the spectrogram of the deverberated signal $y''(t)$, $env(t)$ is the envelope derived from the previous stage, $\gamma = 0.2$ is a scaling constant, and $\epsilon = 0.001$ is the energy floor and corresponds to the maximum attenuation of 30 dB. Finally, the silent gaps between speech utterances were detected and attenuated by 15 dB.

As a last step that is most essential in recordings, where background noise is very noticeable, we decided to include a state-of-the-art method for removing noise in our scheme [11]. Here, we estimated the power spectral density of nonstationary noise, not by using a Voice Activity Detector (VAD), but by minimizing a conditional mean squared estimation error criterion, using the algorithm in [11]. Recursively, an optimal smoothing parameter of the power spectral density was formed. Based on the optimally-smoothed power spectral density estimate and the analysis of the statistics of spectral minima, an unbiased noise estimator was developed. The traditional SNR-based voice activity detectors are difficult to tune, and their application to low SNR speech often results in clipped speech. This soft-decision scheme is capable of updating the noise power spectral density (psd) during speech activity. It has been confirmed [19] that the minimum statistics algorithm [20] performs well in nonstationary noise. For the noise removal stage, we used no special settings, just the algorithm as proposed in [11].

4. Quality and Intelligibility Metrics

There are two aspects to consider in speech enhancement. The first one is the overall perceived quality, which relates to how clear or natural the enhanced speech sounds. The second aspect is speech intelligibility, which is the perception accuracy of the enhanced speech. One way to measure intelligibility is the percentage of the correct words we can distinguish, relative to the total words that exist in a sentence. Although there is a correlation between quality and intelligibility, in some cases, with low quality, we detected high intelligibility and vice versa [21–23].

In order to evaluate speech quality, we can use subjective and/or objective quality measures. In subjective quality measures, we had a group of listeners to hear the anechoic and the processed speech and asked them to rank the quality according to a predetermined scale subjectively. By averaging the results, we obtained a subjective quality score, which is known as the Mean Opinion Score (MOS). Objective speech quality measures, on the other hand, use physical measurements, like acoustic pressure and some mathematically-calculated values from these measurements. Most objective measures are highly correlated with subjective measures. In this paper, we used only objective measures for benchmarking, such as the Frequency-weighted segmental Signal-to-Noise Ratio (fwSNRseg) [24] and Perceptual Evaluation of Speech Quality (PESQ) [25]. Studies have shown that fwSNRseg shows significantly higher correlation with subjective quality than the classical SNR or the segmental SNR [26–28]. PESQ gives a prediction of the perceived quality that would be given to the signal by subjects in a subjective test. ITU-TRec.P.862 provides raw scores in the range -0.5–4.5.

To estimate and compare speech intelligibility in this paper, we utilized two methods: (a) Normalized Sub-band Envelope Correlation using a Gammatone filter bank (NSECGT) [29] and (b) Short-Time Objective Intelligibility (STOI) [30,31]. The first is based on the similarity between the time-varying spectral envelopes of target speech and enhanced speech, as measured by correlation. The second method is based on an intermediate intelligibility measure for short-time (approximately 400 ms) Time-Frequency (TF) regions and uses a simple DFT-based TF-decomposition. Furthermore, in order to compare the amount of reverberant energy in our signals, we used the metric segmental Signal-to-Reverberation Ratio (SRRseg). Last but not least, an adaptive non-intrusive quality and intelligibility measurement termed Speech-to-Reverberation Modulation energy Ratio (SRMR) was employed for a more general comparison [32].

5. Performance Evaluation

5.1. Experimental Setup

We used two datasets to compare the three algorithms: (a) the simulated reverberant dataset and (b) the recorded reverberant dataset. Relative to the simulated dataset we selected, we used the same corpus of speech utterances from eight speakers (four female and four male) from the TIMITdatabase. The reverberant signals were produced by convolving the anechoic signals and the room impulse response function that had a T_{60} of 0.3 s. This constituted the "simulated" dataset. In order to test the algorithms in more realistic and noisy situations, we created the recorded reverberant dataset. The dataset is available from https://github.com/NickKilis/Dereverb. Anechoic audio was reproduced from one loudspeaker and recorded from a microphone, which was connected to a laptop via an external audio card at 24 bits and 44.1 kHz. This process took place in three rooms that had different volumes and T_{60}. The first room was a small domestic room, while the other two were a medium-sized classroom and a large amphitheater at our university. The small room had a T_{60} around 0.5 s and dimensions of 5.25 m $\times$ 2.8 m $\times$ 2.8 m. The classroom had T_{60} around 1.3 s, and its dimensions were 9 m $\times$ 8.2 m $\times$ 2.8 m. The exact dimensions of the amphitheater were not calculated because of its shape complexity. The amphitheater's T_{60} was estimated at around 1.8 s. This constituted the "recorded" dataset. Table 1 outlines the reverberation times of the rooms used in our experiments. The first five seconds of the anechoic audio were speech by a female speaker, and the rest of the 1.93 s were silence. In these extra 1.93 s, we could see the reverberant tail, whilst

the sound energy decayed. This enabled us to compare it visually with the suppressed reverberant tail of the processed signals.

Table 1. Reverberation times RT_{60} for the rooms used in our experiments.

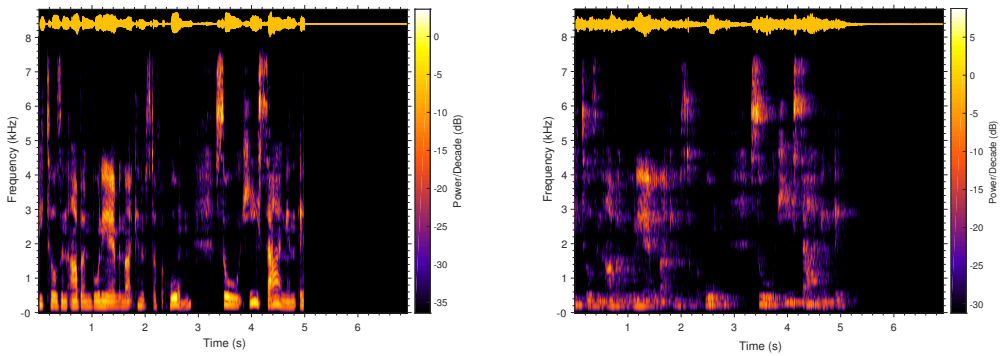
Room Type	RT_{60} **(s)**
Small room	0.5
Classroom	1.3
Amphitheater	1.8
Simulation	0.3

5.2. Evaluation

We compared the new method with the method of Wu and Wang [7] and the Spendred algorithm [8] for the two datasets. All methods were implemented in MATLAB. The Spendred algorithm is publicly available by the VOICEBOX toolbox http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html. The comparison is outlined in Table 2, where values in bold denote the best performance. Time-domain signals aligned with the corresponding spectrogram were compared for the three methods in the case of the small classroom, as depicted in Figure 6.

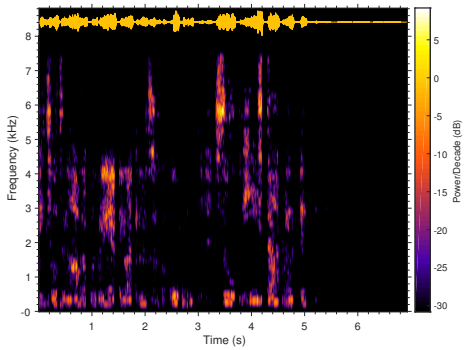
Table 2. Average performance comparison between the proposed algorithm, WW, and the Spendred algorithm (Values in bold indicate the best performance). fwSNRseg, Frequency-weighted segmental Signal-to-Noise Ratio; PESQ, Perceptual Evaluation of Speech Quality; NSECGT, Normalized Sub-band Envelope Correlation using a Gammatone filter bank; STOI, Short-Time Objective Intelligibility; SRRseg, segmental Signal-to-Reverberation Ratio; SRMR, Speech-to-Reverberation Modulation energy Ratio.

	Reference	**WW** [7]	**Spendred** [8]	**Proposed**
		Simulated Dataset		
fwSNRseg (dB)	−4.01	4.01	**4.07**	4.02
PESQ	2	2.19	2.3	**2.36**
NSECGT	0.92	0.81	0.88	**0.89**
STOI	0.79	0.5	0.75	**0.76**
SRRseg (dB)	−20.24	−29.22	**−15**	−18.37
SRMR (dB)	5.55	6.12	7.33	**7.65**
		Recorded Dataset		
fwSNRseg (dB)	−29.07	27.1	13.11	**29.07**
PESQ	1.4	1.18	1.6	**1.62**
NSECGT	0.78	0.71	0.79	**0.81**
STOI	0.49	0.33	0.48	**0.481**
SRRseg (dB)	−27.55	−28.8	**−24.54**	−25.85
SRMR (dB)	3.49	**8.4**	5.7	3.63

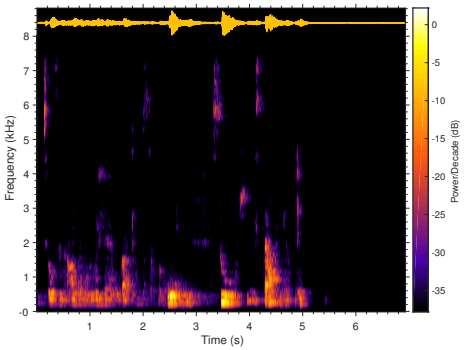


(**a**) Anechoic Signal

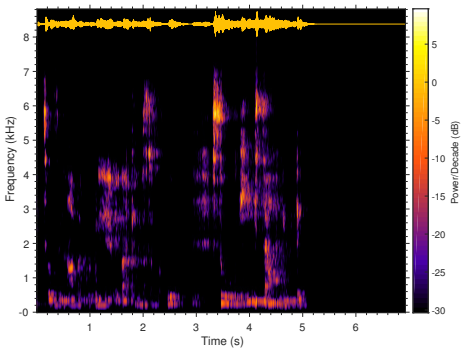
(**b**) Reverberant Signal



(**c**) WW method



(**d**) Spendred



(**e**) Proposed

Figure 6. Time-domain and frequency-domain comparison between the original speech data, the reverberated speech in the "classroom" and the dereverberated output from the three approaches: (**c**) The Wu and Wang (WW) method; (**d**) Spendred; and (**e**) the proposed method.

Our findings on the fwSNRseg metric showed similar average quality performance for the three algorithms in the case of the simulated dataset, but if we also considered the average performance of the recorded dataset, we observed a small lead for our proposed method. It should be mentioned that in the cases of "classroom" and "amphitheater" of the recorded dataset, the Spendred algorithm performed poorly, giving 14.9 dB and 2.6 dB respectively, compared to 27.4 dB and 29.63 dB, achieved by the proposed method. This can be also observed, listening to the outputs of the three algorithms, especially in the case of the noisy recorded signal in the amphitheater. The overall quality performance shown in PESQ was similar for the three methods, but favored slightly the proposed algorithm. Moving on to the overall average intelligibility performance shown in NSECGT and STOI, we can gather that the proposed algorithm had a small lead. The results in segmental SRR demonstrated

two things. First, in both datasets, we observed that the values of this metric were higher than expected, as far as the Wu and Wang method was concerned. Second, the Spendred algorithm performed better, especially in the simulated dataset. The final metric, called SMRM, clearly showed an improvement for all algorithms. The results of this experimental section can be listened to online https://github.com/NickKilis/Dereverb. Overall, in most metrics, the proposed algorithm seemed to outperform the other two approaches.

The main difference in our work is the sparsification stage performed by the K-SVD that may justify the superior performance. This was the novel element that none of the other approaches had incorporated and performed more intelligent sparsification of the LP residual, compared to previous approaches. In addition, we also included some of the interesting elements of the WW method and Martin [11], which offered an intelligent and adaptive method to estimate the noise frequency profile. The effect of Martin's method [11] was only audible when there was increased noise in the room, such as in the amphitheater experiment. Only in these cases, one could hear substantial improvement, which was offered by [11]. Furthermore, in this case, the proposed method offered an advantage compared to the WW method and Spendred, as is highlighted by the results of Table 2. In other cases, where noise was minimal, the offering of [11] was also minimal. However, we chose to incorporate it in our approach to offer a complete solution, in order to cater to the possible presence of noise.

6. Conclusions

In this paper, we presented a new method, based on the Wu and Wang algorithm [7], which can be considered an evolution of their method. Here, the inverse filtering of Wu and Wang's method was replaced by an LP-residual sparsification overcomplete method based on the K-SVD algorithm. The proposed method seemed to outperform Wu and Wang's method and performed favorably with the Spendred method. In addition, the proposed method was flexible, compared to Wu and Wang's method, because of the fact that a hand clap can be easily recorded in any room. Thus, the estimated RIR envelope was more accurate than the Rayleigh envelope used by Wu and Wang. The method's computational complexity seemed to be mainly due to the OMP algorithm. This can be decreased using more computationally-friendly approaches to the OMP algorithm used in the sparsification process [33].

Author Contributions: N.K. was responsible for the investigation, methodology, software, validation, and writing, original draft preparation. N.M. was responsible for the supervision, methodology, software, and writing, review and editing.

Funding: This research received no external funding.

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

References

1. Tonelli, M.; Mitianoudis, N.; Davies, M. A Maximum Likelihood approach to blind Audio de-reverberation. In Proceedings of the 7th International Conference on Digital Audio Effects (DAFX-04), Naples, Italy, 5–8 October 2004.
2. Elko, G.W. Microphone array systems for hands-free telecommunication. *Speech Commun.* **1996**, *20*, 229–240. [CrossRef]
3. Gaubitch, N.D. Blind Identification of Acoustic Systems and Enhancement of Reverberant Speech. Ph.D. Thesis, Imperial College, University of London, London, UK, 2006.
4. Yegnanarayana, B.; Murthy, P.S. Enhancement of reverberant speech using LP residual signal. *IEEE Trans. Speech Audio Process.* **2000**, *8*, 267–281. [CrossRef]
5. Gillespie, B.W.; Florencio, D.A.F.; Malvar, H.S. Speech de-reverberation via maximum-kurtosis sub-band adaptive filtering. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City, UT, USA, 7–11 May 2001; pp. 3701–3704.
6. Tonelli, M. Blind Speech Dereverberation. *Prog. Vib. Acoust.* **2014**, *2*, 1–37.

7. Wu, M.; Wang, D. A two-stage algorithm for one-microphone reverberant speech enhancement. *IEEE Trans. Audio Speech Lang. Process.* **2006**, *14*, 774–784.
8. Doire, C.S.J.; Brookes, M.; Naylor, P.A.; Hicks, C.M.; Betts, D.; Dmour, M.A.; Holdt-Jensen, S. Single-Channel Online Enhancement of Speech Corrupted by Reverberation and Noise. *IEEE Trans. Audio Speech Lang. Process.* **2017**, *25*, 572–587. [CrossRef]
9. Doire, C.S.J. Single-Channel Enhancement of Speech Corrupted by Reverberation and Noise. Ph.D. Thesis, Imperial College, London, UK, 2016.
10. Aharon, M.; Elad, M.; Bruckstein, A.M. The K-SVD: An Algorithm for Designing of Overcomplete Dictionaries for Sparse Representation. *IEEE Trans. Signal Process.* **2006**, *54*, 4311–4322. [CrossRef]
11. Martin, R. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Trans. Speech Audio Process.* **2001**, *9*, 504–512. [CrossRef]
12. Stewart, R.; Sandler, M. Statistical measures of early reflections of room impulse responses. In Proceedings of the 10th International Conference on Digital Audio Effects (DAFx-07), Bordeaux, France, 10–15 December 2007; pp. 59–62.
13. Sumarac-Pavlovic, D.; Mijic, M.; Kurtovic, H. A simple impulse sound source for measurements in room acoustics. *Appl. Acoust.* **2008**, *69*, pp. 378–383. [CrossRef]
14. Repp, B.H. The sound of two hands clapping: An exploratory study. *J. Acoust. Soc. Am.* **1987**, *81*, 1100–1109. [CrossRef] [PubMed]
15. Blesser, B. An interdisciplinary synthesis of reverberation viewpoints. *J. Audio Eng. Soc.* **2001**, *49*, 867–903.
16. Georganti, E.; Mourjopoulos, J.; Jacobsen, F. Analysis of room transfer function and reverberant signal statistics. *Anal. Room Transf. Funct. Reverberant Signal Stat.* **2008**, *123*, 376. [CrossRef]
17. Georganti, E.; Zarouchas, T.; Mourjopoulos, J. Reverberation analysis via response and signal statistics. In Proceedings of the 128th AES Convention, London, UK, 23–25 May 2010.
18. Schroeder, M.R. New Method of Measuring Reverberation Time. *J. Acoust. Soc. Am.* **1965**, *37*, 409–412, . [CrossRef]
19. Meyer, J.; Simmer, K.U.; Kammeyer, K.D. Comparison of one-and two-channel noise-estimation techniques. In Proceedings of the 5th International Workshop Acoustic Echo Control Noise Reduction, London, UK, 11–12 September 1997; pp. 17–20.
20. Martin, R. Spectral subtraction based on minimum statistics. In Proceedings of the European Signal Processing Conference, Edinburgh, UK, 13–16 September 1994; pp. 1182–1185.
21. Hu, Y.; Loizou, P.C. A comparative intelligibility study of speech enhancement algorithms. In Proceedings of the ICASSP, Honolulu, HI, USA, 15–20 April 2007.
22. Shannon, R.; Zeng, F.-G.; Kamath, V.; Wygonski, J.; Ekelid, M. Speech recognition with primarily temporal cues. *Science* **1995**, *270*, 303–304. [CrossRef] [PubMed]
23. Wang, D.; Kjems, U.; Pedersen, M.S.; Boldt, J.B.; Lunner, T. Speech perception of noise with binary gains. *J. Acoust. Soc. Am.* **2008**, *124*, 2303–2307. [CrossRef] [PubMed]
24. Tribolet, J.M.; Noll, P.; McDermott, B.J. A study of complexity and quality of speech waveform coders. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Tulsa, OK, USA, 10–12 April 1978.
25. *ITU-T: Recommendation P.862: Perceptual Evaluation of Quality (PESQ) : An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs*; ITU: Geneva, Switzerland, 2001.
26. Quackenbush, S.R.; Clements, M.A. *Objective Measures of Speech Quality*; Prentice-Hall: Englewood Cliffs, NJ, USA, 1988.
27. Hu, Y.; Loizou, P.C. Evaluation of objective quality measures for speech enhancement. *IEEE Trans. Audio Speech Lang. Process.* **2008**, *16*, 229–238. [CrossRef]
28. Ma, J.; Hu, Y.; Loizou, P.C. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. *J. Acoust. Soc. Am.* **2009**, *125*, 3387–3405. [CrossRef] [PubMed]
29. Boldt, J.B.; Ellis, D.P.W. A Simple Correlation-Based Model of Intelligibility for Nonlinear Speech Enhancement and Separation. In Proceedings of the 17th European Signal Processing Conference (EUSIPCO), Glasgow, UK, 24–28 August 2009.
30. Taal, C.H.; Hendriks, R.C.; Heusdens, R.; Jensen, J. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In Proceedings of the ICASSP, Dallas, TX, USA, 14–19 March 2010.

31. Taal, C.H.; Hendriks, R.C.; Heusdens, R.; Jensen, J. An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech. *IEEE Trans. Audio Speech Lang. Process.* **2011**, *19*, 2125–2136. [CrossRef]
32. Falk, T.H.; Zheng, C.; Chan, W.-Y. A Non-Intrusive Quality and Intelligibility Measure of Reverberant and Dereverberated Speech. *IEEE Trans. Audio Speech Lang. Process.* **2010**, *18*, 1766–1774. [CrossRef]
33. Blumensath, T.; Davies, M.E. Stagewise Weak Gradient Pursuits. *IEEE Trans. Signal Process.* **2009**, *57*, 4333–4346. [CrossRef]