ENHANCEMENT OF REVERBERANT AND NOISY SPEECH BY EXTENDING ITS COHERENCE

Scott Wisdom[⋆]*, Thomas Powers*[⋆]*, Les Atlas*[⋆]*, and James Pitton*[†⋆]

[⋆]Electrical Engineering Department, University of Washington, Seattle, USA
[†]Applied Physics Laboratory, University of Washington, Seattle, USA

ABSTRACT

We introduce a novel speech enhancement algorithm for removing reverberation and noise from recorded speech data. Our approach centers around using a single-channel minimum mean-square error log-spectral amplitude (MMSE-LSA) estimator, which applies gain coefficients in a time-frequency domain to suppress noise and reverberation. The main contribution of this paper is that the enhancement is done in a time-frequency domain that is coherent with speech signals over longer analysis durations than the short-time Fourier transform (STFT) domain. This extended coherence is gained by using a linear model of fundamental frequency variation over the analysis frame. In the multichannel case, we preprocess the data with either a minimum variance distortionless response (MVDR) beamformer, or a delay-and-sum beamformer (DSB). We evaluate our algorithm on the REVERB challenge dataset. Compared to the same processing done in the STFT domain, our approach achieves significant improvement on the REVERB challenge objective metrics, and according to informal listening tests, results in fewer artifacts in the enhanced speech.

Index Terms— Speech enhancement, speech dereverberation, time-warping, fan-chirp transform, adaptive basis, beamforming

1. INTRODUCTION AND PRIOR WORK

The enhancement of speech signals in the presence of reverberation and noise remains a challenging problem with many applications. Many methods are prone to generating artifacts in the enhanced speech, and must trade off noise reduction against speech distortion.

In this paper, we describe a new enhancement algorithm that suppresses both reverberation and background noise. We combine a statistically optimal single-channel enhancement algorithm that suppresses background noise and reverberation with an adaptive time-frequency transform domain that is coherent with speech signals over longer durations than the short-time Fourier transform (STFT). Thus, we are able to use longer analysis windows while still satisfying the assumptions of the optimal single-channel enhancement filter. Multichannel processing is made possible using a classic minimum variance distortionless response (MVDR) beamformer or, in the case of two-channel data, a delay-and-sum beamformer (DSB) preceding the single-channel enhancement.

First, we review the speech enhancement and dereverberation problem, as well as the enhancement algorithm we use proposed by Habets [1], which suppresses both noise and late reverberation based on a statistical model of reverberation. Then, we describe the fan-chirp transform, proposed by Weruaga and Kèpesi [2, 3] and improved upon by Cancela et al. [4], which provides an enhancement domain, the short-time fan-chirp transform (STFChT), that better matches time-varying frequency content of voiced speech. We discuss why performing the enhancement in the STFChT domain gives superior results compared to the STFT domain. Finally, we present our results on the REVERB challenge dataset [5], which shows that our new method achieves superior results versus conventional STFT-based processing in terms of objective measures.

Our basic multichannel architecture of single-channel enhancement preceded by beamforming is not unprecedented. Gannot and Cohen [6] used a similar architecture for noise reduction that consists of a generalized sidelobe cancellation (GSC) beamformer followed by a single-channel post-filter. Maas et al. [7] employed a similar single-channel enhancement algorithm for reverberation suppression and observed promising speech recognition performance in even highly reverberant environments.

There have been several dereverberation and enhancement approaches that estimate and leverage the time-varying fundamental frequency f_0 of speech. Nakatani et al. [8] proposed a dereverberation method using inverse filtering that exploits the harmonicity of speech to build an adaptive comb filter. Kawahara et al. [9] used adaptive spectral analysis and estimates of f_0 to perform manipulation of speech characteristics.

Droppo and Acero [10] observed how the fundamental frequency of speech can change within an analysis window, and proposed a new framework that could better predict the energy of voiced speech. Dunn and Quatieri [11] used the fan-chirp transform for sinusoidal analysis and synthesis of speech, and Dunn et al. [12] also examined the effect of various interpolation methods on reconstruction error. Pantazis et al. [13] proposed an analysis/synthesis domain that uses estimates of instantaneous frequency to decompose speech into quasi-harmonic AM-FM components. Degottex and Stylianou [14] proposed another analysis/synthesis scheme for speech using an adaptive harmonic model that they claim is more flexible than the fan-chirp, as it allows nonlinear frequency trajectories.

To our knowledge, single-channel enhancement has not been attempted in these new related transform domains. Here, we demonstrate improved performance using the STFChT instead of the STFT.

2. OPTIMAL SINGLE-CHANNEL SUPPRESSION OF NOISE AND LATE REVERBERATION

In this section, we review the speech enhancement problem and a popular statistical speech enhancement algorithm, the minimum mean-square error log-spectral amplitude (MMSE-LSA) estimator, which was originally proposed by Ephraim and Malah [15, 16] and later improved by Cohen [17]. We

This work is funded by ONR contract N00014-12-G-0078, delivery order 0013

review the application of MMSE-LSA to both noise reduction and joint dereverberation and noise reduction (the latter of which was proposed by Habets [1]).

2.1. Noise reduction using MMSE-LSA

A classic speech enhancement algorithm is the minimum mean-square error (MMSE) short-time spectral amplitude estimator proposed by Ephraim and Malah [15]. They later refined the estimator to minimize the MSE of the log-spectra [16]. We will refer to this algorithm as LSA (log-spectral amplitude). Minimizing the MSE of the log-spectra was found to provide better enhanced output because log-spectra are more perceptually meaningful. Cohen [17] suggested improvements to Ephraim and Malah's algorithm, which he referred to as "optimal modified log-spectral amplitude" (OM-LSA).

Given samples of a noisy speech signal

$$y[n] = s[n] + v[n], \tag{1}$$

where $s[n]$ is the clean speech signal and $v[n]$ is additive noise, the goal of an enhancement algorithm is to estimate $s[n]$ from the noisy observations $y[n]$. The LSA estimator yields an estimate $\hat{A}(d,k)$ of the clean STFT magnitudes $|S(d,k)|$ (where $S(d,k)$ are assumed to be normally distributed) by applying a frequency-dependent gain $G_{\mathrm{LSA}}(d,k)$ to the noisy STFT magnitudes $|Y(d,k)|$:

$$\hat{A}(d,k) = G_{\mathrm{LSA}}(d,k)|Y(d,k)|. \tag{2}$$

Given these estimated magnitudes, the enhanced speech is reconstructed from STFT coefficients combining $\hat{A}(d,k)$ with noisy phase: $\hat{S}(d,k) = \hat{A}(d,k)e^{j\angle Y(d,k)}$. The LSA gains are computed as [16, (20)]:

$$G_{\mathrm{LSA}}(d,k) = \frac{\xi(d,k)}{1+\xi(d,k)} \exp\left\{\frac{1}{2}\int_{v(d,k)}^{\infty} \frac{e^{-t}}{t} dt\right\}, \tag{3}$$

where $\xi(d,k)$ is the *a priori* signal-to-noise ratio (SNR) for the kth frequency bin of the dth frame, and is defined to be $\xi(d,k) \triangleq \frac{\lambda_s(d,k)}{\lambda_v(d,k)}$, where $\lambda_s(d,k) = E\left\{|S(d,k)|^2\right\}$ is the variance of $S(d,k)$ and $\lambda_v(d,k) = E\left\{|V(d,k)|^2\right\}$ is the variance of $V(d,k)$. The variable $v(d,k) = \frac{\xi(d,k)}{1+\xi(d,k)}\gamma(d,k)$, where $\gamma(d,k)$ is the *a posteriori* SNR for the kth frequency bin of the dth frame, defined as $\gamma(d,k) \triangleq \frac{|Y(d,k)|^2}{\lambda_v(d,k)}$.

Cohen [17] refined Ephraim and Malah's approach to include a lower bound $G_{\min}$ for the gains as well as an *a priori* speech presence probability (SPP) estimator $p(d,k)$. Cohen's estimator is as follows [17, (8)]:

$$G_{\mathrm{OM-LSA}} = \{G_{\mathrm{LSA}}(d,k)\}^{p(d,k)} \cdot G_{\min}^{1-p(d,k)}. \tag{4}$$

Cohen also derived an efficient estimator for the SPP $p(d,k)$ [17] that exploits the strong interframe and interfrequency correlation of speech in the STFT domain.

2.2. Joint dereverberation and noise reduction

Habets [1] proposed a MMSE-LSA enhancement algorithm that uses a statistical model of reverberation to suppress both noise and late reverberation. The signal model he uses is

$$y[n] = s[n] * h[n] + v[n] = x_e[n] + x_\ell[n] + v[n], \tag{5}$$

where $s[n]$ is the clean speech signal, $h[n]$ is the room impulse response (RIR), and $v[n]$ is additive noise. The terms $x_e[n]$ and $x_\ell[n]$ correspond to the early and late reverberated speech signals, respectively. The partition between early and late reverberations is determined by a parameter n_e, which is a discrete sample index. All samples in the RIR before n_e are taken to cause early reflections, and all samples after n_e are taken to cause late reflections [1]. Thus,

$$h[n] = \begin{cases} 0, & \text{if } n < 0 \\ h_e[n], & \text{if } 0 \le n < n_e \\ h_\ell[n] & \text{if } n_e \le n. \end{cases} \tag{6}$$

Using these definitions, $x_e[n] = s[n] * h_e[n]$ and $x_\ell[n] = s[n] * h_\ell[n]$.

Habets proposed a generalized statistical model of reverberation that is valid both when the source-microphone distance is less than or greater than the critical distance [1]. This model divides the RIR $h[n]$ into a direct-path component $h_d[n]$ and reverberant component $h_r[n]$. Both direct-path and reverberant components are taken to be white, zero-mean, stationary Gaussian noise sequences $b_d[n]$ and $b_r[n]$ with variances σ_d^2 and σ_r^2 enveloped by an exponential decay,

$$h_d[n] = b_d[n]e^{-\bar{\zeta}n} \text{ and } h_r[n] = b_r[n]e^{-\bar{\zeta}n}, \tag{7}$$

where $\bar{\zeta}$ is related to the reverberation time T_{60} by [1]:

$$\bar{\zeta} = \frac{3\ln(10)}{T_{60} f_s}. \tag{8}$$

Using this model, the expected value of the energy envelope of $h[n]$ is

$$E\left[h^2[n]\right] = \begin{cases} \sigma_d^2 e^{-2\bar{\zeta}n}, & \text{for } 0 \le n < n_d \\ \sigma_r^2 e^{-2\bar{\zeta}n}, & \text{for } n \ge n_d \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

Under the assumptions that the speech signal is stationary over short analysis windows (i.e., duration much less than T_{60}), Habets proposed [1, (3.87)] the following model of the spectral variance of the reverberant component $x_r[n]$:

$$\begin{aligned} \lambda_{x_r}(d,k) =& e^{-2\bar{\zeta}(k)R}\lambda_{x_r}(d-1,k)... \\ &+ \frac{E_r}{E_d}\left(1 - e^{-2\bar{\zeta}(k)R}\right)\lambda_{x_d}(d-1,k), \end{aligned} \tag{10}$$

where R is the number of samples separating two adjacent analysis frames and E_r/E_d is the inverse of the direct-to-reverberant ratio (DRR). Thus, the spectral variance of the reverberant component in the current frame d is composed of scaled copies of the spectral variance of the reverberation and the spectral variance of the direct-path signal from the previous frame $d-1$.

Using this model, the variance of the late reverberant component can be expressed as [1, (3.85)]:

$$\lambda_{x_\ell}(d,k) = e^{-2\bar{\zeta}(k)(n_e - R)}\lambda_{x_r}\left(d - \frac{n_e}{R} + 1, k\right), \tag{11}$$

which is quite useful in practice, because the variance of the late-reverberant component can be computed from the variance of the total reverberant component.

To suppress both noise and late reverberation, the *a priori* and *a posteriori* SNRs $\xi(d,k)$ and $\gamma(d,k)$ from the previous section become *a priori* and *a posteriori* signal-to-interference ratios (SIRs), given by [1, (3.25), (3.26)]:

$$\xi(d,k) = \frac{\lambda_{x_e}(d,k)}{\lambda_{x_\ell}(d,k) + \lambda_v(d,k)} \tag{12}$$

and

$$\gamma(d,k) = \frac{|Y(d,k)|^2}{\lambda_{x_\ell}(d,k) + \lambda_v(d,k)}. \tag{13}$$

The gains are computed by plugging the SIRs in (12) and (13) into (3) and (4). Habets suggested an additional change to (4), which makes $G_{\min}$ time- and frequency-dependent. This is done because the interference of both noise and late reverberation is time-varying. The modification is [1, (3.29)]

$$G_{\min}(d,k) = \frac{G_{\min,x_\ell}\hat{\lambda}_{x_\ell}(d,k) + G_{\min,v}\hat{\lambda}_v(d,k)}{\hat{\lambda}_{x_\ell}(d,k) + \hat{\lambda}_v(d,k)}. \tag{14}$$

Notice that two parameters in (8) and (10) are not known *a priori*; namely, T_{60} and the DRR. These parameters must be blindly estimated from the data. For T_{60} estimation, Löllmann et al. [18] propose an algorithm, which we found to be effective. As for the DRR, Habets suggested an online adaptive procedure [1, §3.7.2].

3. ANALYSIS AND SYNTHESIS USING THE FAN-CHIRP TRANSFORM

In this section, we review the forward short-time fan-chirp transform (STFChT) and describe a method of inverting the STFChT.

3.1. The forward fan-chirp transform

We adopt the fan-chirp transform formulation used by Cancela et al. [4]. The forward fan-chirp transform is defined as

$$X(f,\alpha) = \int x(t)\phi'_\alpha(t)e^{-j2\pi f\phi_\alpha(t)}dt \tag{15}$$

where $\phi_\alpha(t) = \left(1 + \frac{1}{2}\alpha t\right)t$ and $\phi'_\alpha(t) = 1+\alpha t$. The variable α is an analysis chirp rate. Using a change of variable $\tau \leftarrow \phi_\alpha(t)$, (15) can be written as the Fourier transform of a time-warped signal:

$$X(f,\alpha) = \int_{-\infty}^{\infty} x(\phi_\alpha^{-1}(\tau))e^{-j2\pi f\tau}d\tau. \tag{16}$$

The short-time fan-chirp transform (STFChT) of $x(t)$ is defined as the fan-chirp transform of the dth short frame of $x(t)$:

$$X_d(f,\hat{\alpha}_d) = \int_{-T_w/2}^{T_w/2} w(\tau)x_d(\phi_{\hat{\alpha}_d}^{-1}(\tau))e^{-j2\pi f\tau}d\tau \tag{17}$$

where $w(t)$ is an analysis window, $\hat{\alpha}_d$ is the analysis chirp rate for the dth frame given by (21), and $x_d(t)$ is the dth short frame of the input signal of duration T:

$$x_d(t) = \begin{cases} x(t - dT_{hop}), & -T/2 \le t \le T/2 \\ 0, & \text{otherwise.} \end{cases} \tag{18}$$

T is the duration of the pre-warped short-time duration, T_{hop} is the frame hop, T_w is the post-warped short-time duration, and $w(t)$ is a T_w-long analysis window. The analysis window is applied after time-warping so as to avoid warping of the window, which can cause unpredictable smearing of the Fourier transform.

Implementing the fan-chirp transform as a time-warping followed by a Fourier transform allows efficient implementation, consisting simply as an interpolation of the signal followed by an FFT. In the implementation provided by Cancela et al. [4], the interpolation used in the forward fan-chirp transform is linear.

Kèpesi and Weruaga [2] provide a method for determination of the analysis chirp rate α using the gathered log spectrum (GLogS). The GLogS is defined as follows:

$$\rho(f_0,\alpha) = \frac{1}{N_h}\sum_{k=1}^{N_h} \ln|X(kf_0,\alpha)| \tag{19}$$

where N_h is the maximum number of harmonics that fit within the analysis bandwidth. That is,

$$N_h = \left\lfloor \frac{f_s}{2f_0\left(1 + \frac{1}{2}|\alpha|T_w\right)} \right\rfloor. \tag{20}$$

Cancela et al. [4] proposed several enhancements to the GLogS. First, they observed improved results by replacing $\ln|\cdot|$ with $\ln(1+\gamma|\cdot|)$. Cancela et al. note that this expression approximates a p-norm, with $0 < p < 1$, where lower values of γ with $\gamma \ge 1$ approach the 1-norm, while higher values approaches the 0-norm. Cancela et al. note that $\gamma = 10$ gave good results for their application.

Additionally, Cancela et al. propose modifications that suppress multiples and submultiples of the current f_0. Also, they propose normalizing the GLogS such that is has zero mean and unit variance. This is necessary because the variance of the GLogS increases with increasing fundamental frequency. For mean and variances measured over all frames in a database, a polynomial fit is determined and the GLogS are compensated using these polynomial fits.

Let $\bar{\rho}_d(f_0,\alpha)$ be the GLogS of the dth frame with these enhancements applied. For practical implementation, finite sets $\mathcal{A}$ of candidate chirp rates and $\mathcal{F}_0$ of candidate fundamental frequencies are used, and the GLogS is exhaustively computed for every chirp rate in $\mathcal{A}$ and fundamental frequency in $\mathcal{F}_0$. The analysis chirp rate $\hat{\alpha}_d$ for the dth frame is thus found by

$$\hat{\alpha}_d = \underset{\alpha\in\mathcal{A}}{\operatorname{argmax}} \max_{f_0\in\mathcal{F}_0} \bar{\rho}_d(f_0,\alpha). \tag{21}$$

3.2. The inverse fan-chirp transform

Inverting the fan-chirp transform is a matter of reversing the steps used in the forward transform. Thus, the inverse fan-chirp transform for a short-time frame consists of an inverse Fourier transform, removal of the analysis window, and an inverse time-warping. The removal of the analysis window $w(t)$ from the T_w-long warped signal limits the choice of analysis windows to positive functions only, such as a Hamming window, so the window can be divided out. Also, since the warping is nonuniform, it is possible that the sampling interval between points may exceed the Nyquist sampling interval. To combat this, the data should be oversampled before time-warping, which means the data must be downsampled after undoing the time-warping.

The choice of post-warped duration T_w and the method of interpolation used in the inverse time-warping affect the reconstruction error of the inverse fan-chirp transform. There is a trade-off between reconstruction performance and computational complexity, because interpolation error decreases as interpolation order increases. Kèpesi and Weruaga [19] analyzed fan-chirp reconstruction error with respect to order of the time-warping interpolation and oversampling factor, and found that for cubic Hermite splines and an oversampling factor of 2, a signal-to-error ratio of over 30dB can be achieved. For our application, we choose an oversampling factor of 8 and cubic-spline interpolation.

4. MMSE-LSA IN THE FAN-CHIRP DOMAIN

In this section we analyze performing joint dereverberation and noise reduction using MMSE-LSA in the STFChT domain and provide an example for why the STFChT (17), which is a domain that is more coherent with speech signals, is a more appropriate enhancement domain than the STFT.

The MMSE-LSA framework for joint dereverberation and noise reduction implicitly assumes that the the frequency content of speech does not change very much over the analysis duration. Such an assumption relies on the local stationarity of speech signals within the analysis frame. For voiced speech, this is essentially equivalent to the fundamental frequency f_0 being constant over the analysis frame, and the frequency variation of voiced speech limits analysis durations to 10-30ms.

Using shorter analysis frames means only a finite amount of approximately stationary data is available at any specific time, and this finite amount of data limits the performance of statistical estimators. To improve this situation, we propose to increase the analysis duration by changing the time base of the analysis such that there is less frequency variation within the frame, which makes the data more stationary. This time base modification is performed using the fan-chirp, which uses a linear model of frequency variation within the frame.

To give intuition about the benefits of the fan-chirp in the presence of reverberation, we present a simple example in figure 1. Consider two successive Gaussian-enveloped harmonic chirps with duration 200 ms and spaced 100 ms apart. Let the f_0 of the first harmonic chirp start at 200 Hz and rise to 233 Hz, and let the f_0 of the second harmonic chirp have a range from 250 Hz falling to 200 Hz. Both chirps have 20 harmonics. This sequence of harmonic chirps has parameters that are typical of two successive voiced vowels (here we do not consider the spectral shape imposed by a vocal tract filter, for simplicity). Now, let us apply reverberation to this signal (we use the first channel of the `MediumRoom2_far_AnglA` RIR provided in the REVERB challenge development set [5]), and examine the clean and reverberated versions in both the STFT and STFChT domains. The result is shown in figure 1. STFT and STFChT parameters are exactly matched, with a sampling rate of $f_s = 16$kHz, a Hamming window of duration 2048 samples, a frame hop of 128 samples, and a 3262-length FFT.

Notice that at higher frequencies in the STFT of the clean signal (top left panel of figure 1), the harmonics become broader and more smeared across frequency. In contrast, the STFChT of the clean signal (center left panel of figure 1) exhibits narrow lines at all frequencies. When reverberation is applied, the STFT of the reverberated signal (top right panel of figure 1) exhibits smears that become wider with increasing frequency, and adjacent harmonics even become smeared together. In contrast, in the STFChT of the reverberated signal (center right panel of figure 1), the direct path signal shows up as narrow lines at all frequencies, while some

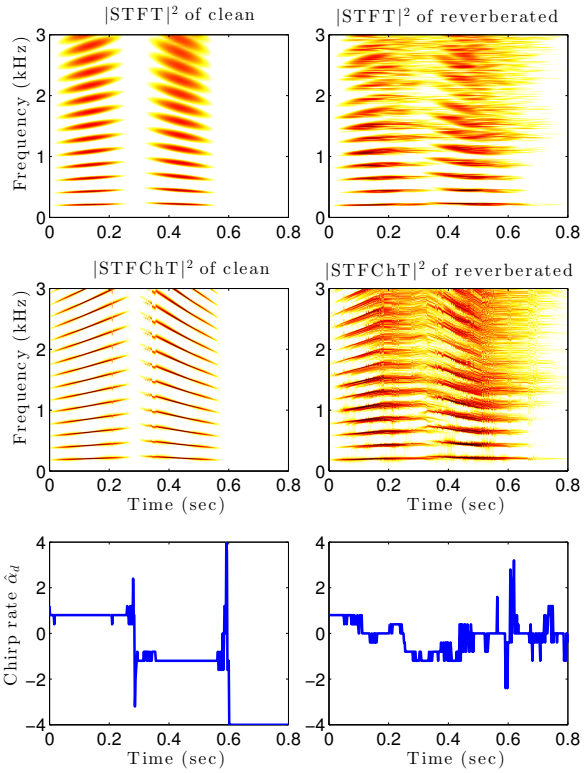

Fig. 1: Simple example showing benefits of fan-chirp both for narrower harmonics across frequency and for better coherence with direct path signal in the presence of reverberation. Test signal is two consecutive synthetic speech-like harmonic stacks. All colormaps are identical with a dynamic range of 40dB. Left plots are the representations of the clean signal and right plots are the representations of the reverberated signal. The chosen analysis chirp rates $\hat{\alpha}_d$ are shown in the bottom plots.

smearing results in frames that contain reverberant energy. One cause of the smearing during reverberation-dominated frames seems to be errors in the estimation of $\hat{\alpha}_d$, the analysis chirp rate, caused by the additional reverberant components, which are shown in the bottom panels of figure 1. Despite these estimation errors, the STFChT still seems to give a better representation of the signal compared to the STFT, because the STFChT reduces smearing of higher-frequency components and achieves better coherence with the direct-path signal (i.e., direct-path signals show up as more narrow lines).

5. IMPLEMENTATION

Our algorithms are implemented in MATLAB, and we use utterance-based processing. The algorithm starts by using the utterance data to estimate the T_{60} time of the room using the blind algorithm proposed by Löllmann et al. [18]. Multichannel utterance input data is concatenated into a long vector, and as recommended by Löllmann et al., noise reduction is performed beforehand. We use Loizou's implementation [20] of Ephraim and Malah's LSA [16] for this pre-enhancement. Figure 2 shows histograms of the T_{60} estimation performance using this approach.

For multichannel data, we estimate the direction of arrival (DOA) by cross-correlating oversampled data between chan-

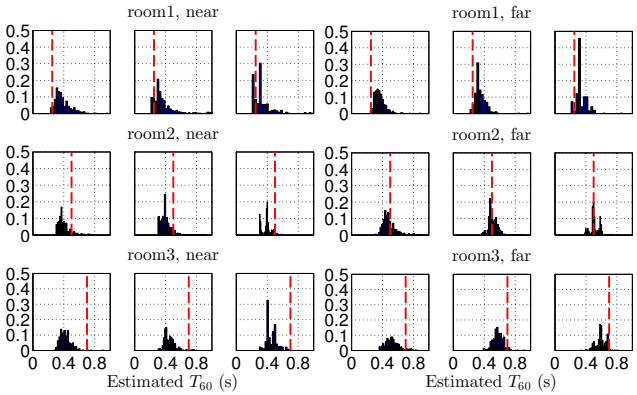

Fig. 2: Histograms of estimated T_{60} time measured on SimData evaluation dataset (these results were not used to tune the algorithm). For each condition, left plot is for 1-channel data, center plot is for 2-channel data, and right plot is for 8-channel data. These plots show that T_{60} estimation [18] precision generally improved with increasing amounts of data (i.e., with more channels), although for some conditions T_{60} estimates were inaccurate. Dotted lines indicate approximate T_{60} times given by REVERB organizers [5].

nels. That is, we compute a N_{ch}-length vector of time delays $\mathbf{d}$ with $d_1 = 0$ and d_i, i=2,...,N_{ch} given by

$$d_i = \underset{k}{\operatorname{argmax}} \frac{r_{1i}[k]}{U f_s}, \tag{22}$$

where $r_{1i}[k] = \sum_n x_1[n]x_i[n-k]$, U is the oversampling factor, and $c = 340$ meters per second, the approximate speed of sound in air.

Given a time delay vector $\mathbf{d}$, the DOA estimate is given by the solution to $\mathbf{P}\hat{\mathbf{a}} = \frac{1}{c}\mathbf{d}$, where $\hat{\mathbf{a}}$ is a 3×1 unit vector representing the estimated DOA of the speech signal and $\mathbf{P}$ is a $N_{ch} \times 3$ matrix containing the Cartesian (x, y, z) coordinates of the array elements. For example, for an eight-element uniform circular array, $P_{i1} = x_i = r\cos(i\pi/4)$, $P_{i2} = y_i = r\sin(i\pi/4)$, and $P_{i3} = z_i = 0$ for $i = 0, 1, ..., 7$, where r is the array radius.

For the 8-channel case, the estimated DOA is used to form the steering vector $\mathbf{v}^H(f)$ for a frequency-domain MVDR beamformer applied to the multichannel signal. The weights $\mathbf{w}^H(d, f)$ for the MVDR are [21, (6.14-15)]

$$\mathbf{w}^H(d, f) = \frac{\mathbf{v}^H(f)\mathbf{S}_{\mathbf{yy}}^{-1}(d, f)}{\mathbf{v}^H(d, f)\mathbf{S}_{\mathbf{yy}}^{-1}(d, f)\mathbf{v}(d, f)}, \tag{23}$$

where $\mathbf{S}_{\mathbf{yy}}(d, f)$ is the spatial covariance matrix at frequency f and frame d estimated using N snapshots $Y(d - n, f)$ for $-N/2 \leq n < N/2$ and $\mathbf{v}$ is given by

$$\mathbf{v}(f) = \exp\left(j\frac{2\pi f}{c}\mathbf{P}\hat{\mathbf{a}}\right). \tag{24}$$

The MVDR uses a 512-sample long Hamming window with 25% overlap, a 512-point FFT, and $N = 24$ snapshots for spatial covariance estimates. For 2-channel data, we use a delay-and-sum beamformer to enhance the signal with the delay given by the DOA estimate. Single-channel data is enhanced directly by the single-channel MMSE-LSA algorithm. A block diagram of these three cases is shown in figure 3.

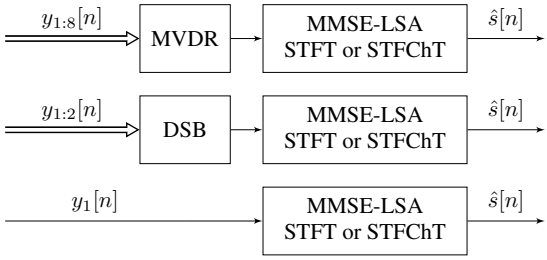

Fig. 3: Block diagrams of processing for 8-channel data using a minimum variance distortionless response (MVDR) beamformer (top), 2-channel data using a delay-and-sum beamformer (DSB, middle), and 1-channel data (bottom).

We tried three analysis/synthesis domains for the MMSE-LSA enhancement algorithm: the STFT with a short window, the STFT with a long window, and the STFChT. The STFT with a short window uses 512-sample long ($T = 32$ms) Hamming windows, a frame hop of 128 samples, and an FFT length of 512. Short-window STFT processing is chosen to match conventional speech processing window lengths. The STFT with a long window uses 2048-sample long ($T = 128$ms) Hamming windows, a frame hop of 128 samples, and an FFT length of 3262. Long-window STFT processing is intended to match the parameters of STFChT processing for a direct comparison. STFChT processing uses an analysis duration of 2048 samples, a Hamming analysis window, a frame hop of 128 samples, an FFT length of 3262, oversampling factor of 8, and a set of possible analysis chirp rates $\mathcal{A}$ consisting of 21 equally spaced αs from -4 to 4.

The forward STFChT, given by (17), proceeds frame-by-frame, estimating the optimal analysis chirp rate $\hat{\alpha}_d$ using (21), oversampling in time, warping, applying an analysis window, and taking the FFT. Then MMSE-LSA weights are estimated frame-by-frame and applied in the STFChT domain, and the enhanced speech signal is reconstructed using the inverse STFChT.

For all methods, noise estimation is performed with a decision-directed method and simple online updating of the noise variance. Voice activity detection to determine if a frame is noise-only or speech-plus-noise is done using Loizou's method, which compares the following quantity to a threshold η_{thresh}:

$$\eta(d) = \sum_k \ln\gamma(d, k)\frac{\xi(d, k)}{1 + \xi(d, k)} - \ln(1 + \xi(d, k)). \tag{25}$$

If $\eta(d) < \eta_{\text{thresh}}$, the frame is determined to be noise-only and the noise variance is updated as $\lambda_v(d, k) = \mu_v\lambda_v(d - 1, k) + (1-\mu_v)|Y(d, k)|^2$, with $\mu_v = 0.98$ and $\eta_{\text{thresh}} = 0.15$.

For our implementation of Habets's joint dereverberation and noise reduction algorithm, we used Loizou's implementation [20] of Ephraim and Malah's LSA `logmmse` MATLAB implementation as a foundation. The forward STFChT code was written by Cancela et al. [4]. We wrote our own MATLAB implementation of the inverse STFChT.

Computation times for processing REVERB evaluation data are shown in figure 7. We measured reference wall clock times of 265.43s and 39.62s, respectively, for SimData and RealData. For 8-channel data, the MVDR and the STFChT require the most computation. For 1-channel and 2-channel data, the STFChT requires the most computation. For the STFChT, much of the computation is used to compute the GLogS for estimation of the analysis chirp rate $\hat{\alpha_d}$ (21) for each frame. Note that this computation could be easily parallelized in hardware.

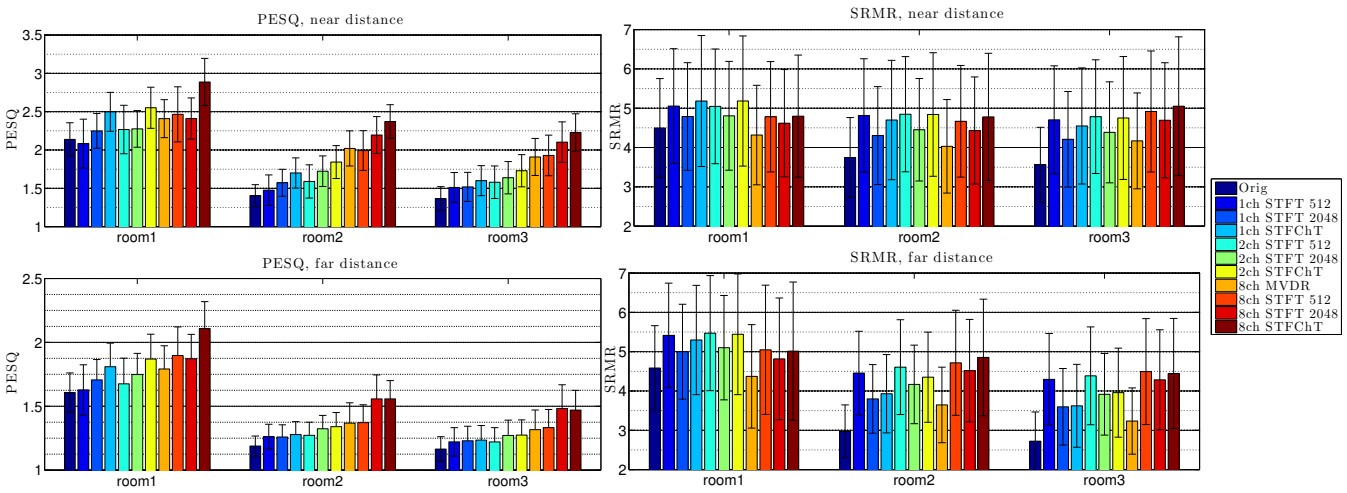

Fig. 4: PESQ and SRMR results for SimData evaluation set. Upper plots are near distance condition, lower plots are far distance condition. Left plots are PESQ, right plots are SRMR.

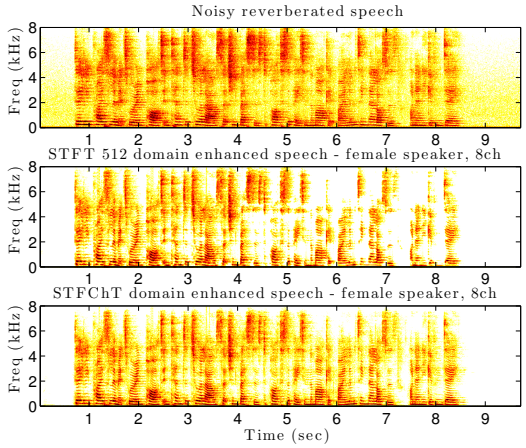

Fig. 5: Spectrogram comparisons for one 8-channel far-distance utterance, `c3bc020q`, from SimData evaluation set.

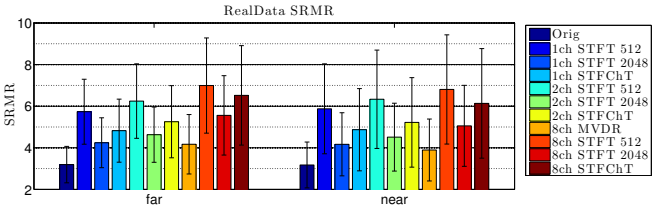

Fig. 6: SRMR results for RealData evaluation set.

6. RESULTS AND DISCUSSION

Our results on REVERB evaluation data are shown in figures 4, 6, and 7. For the challenge, we submitted results using STFChT-based processing. We choose to display PESQ (Perceptual Evaluation of Speech Quality) [22] and SRMR (source-to-reverberation modulation energy ratio) [23] more prominently because the former is the ITU-T standard for voice quality testing [24] and the latter is both a measure of dereverberation and the only non-intrusive measure that can be run on RealData (for which the clean speech is not available).

For SimData, STFChT-based enhancement always performs better in terms of PESQ than STFT-based enhancement using either a short (512-sample) window or a long (2048-sample) window, for the 8-, 2-, and 1-channel cases (except for 8-channel, far-distance data in room 3). Informal listening tests revealed an oversuppression of speech and some musical noise artifacts in STFT processing, while STFChT processing did not exhibit oversuppression or musical noise artifacts. The oversuppression of direct-path speech by STFT processing can be seen in the spectrogram comparisons shown in figure 5. In terms of SRMR, STFChT processing yields equivalent or slightly worse SRMR scores than long-window STFT processing for the 8-, 2-, and 1-channel cases (except for 8-channel, near-distance data, where STFChT processing does slightly better). One issue with these SRMR comparisons, however, is that the variance of the SRMR scores is quite high. Thus, for SimData, STFChT processing achieves better perceptual audio quality while still achieving almost equivalent dereverberation compared to STFT processing.

For RealData, we achieved SRMR improvements of over 3, as shown in figure 6. Short-window STFT processing achieved higher scores than STFChT processing (especially for 1- and 2-channel data), but informal listening tests revealed an oversuppression of speech and some musical noise artifacts in STFT processing, while little oversuppression and fewer artifacts were perceived in STFChT processing. Informal listening tests also indicated that the STFChT processing suppresses reverberation slightly less as compared to STFT processing, which concurs with lower SRMR scores for STFChT processing. Thus, though STFT processing achieves better dereverberation on RealData, better dereverberation performance seems to come at the cost of oversuppression of direct-path speech and addition of artifacts. STFChT processing, on the other hand, achieves slightly less dereverberation on RealData, but the enhanced speech does not seem to suffer from oversuppression or artifacts.

7. CONCLUSION AND FUTURE WORK

In this paper we combined an optimal MMSE log-spectral amplitude estimator for joint dereverberation and noise reduction with a recently-developed adaptive time-frequency transform that is coherent with speech signals over longer durations. Our approach yielded improved results on the RE-

SimData summary

Ch.	Method	Comp. Time (s)	Mean CD	Median CD	SRMR	Mean LLR	Median LLR	Mean FWSegSNR	Median FWSegSNR	PESQ
	Orig	—	*3.97*	*3.68*	*3.68*	*0.57*	*0.51*	*3.62*	*5.39*	*1.48*
8	STFT 512/2048	7447.86 / 7622.38	3.56 / 3.18	3.23 / 2.83	4.77 / 4.56	0.61 / 0.43	0.50 / 0.38	8.06 / 6.79	8.47 / 9.31	1.83 / 1.94
8	STFChT	17132.60	2.97	2.49	4.82	0.43	0.37	9.21	10.63	2.10
2	STFT 512/2048	1955.89 / 2022.04	3.80 / 3.57	3.42 / 3.22	4.86 / 4.47	0.65 / 0.49	0.55 / 0.44	7.26 / 5.46	7.93 / 7.86	1.60 / 1.66
2	STFChT	8248.49	3.33	2.83	4.75	0.51	0.45	7.68	9.19	1.77
1	STFT 512/2048	1003.54 / 1074.93	3.87 / 3.84	3.48 / 3.51	4.79 / 4.28	0.68 / 0.54	0.58 / 0.47	6.72 / 4.65	7.62 / 6.71	1.53 / 1.59
1	STFChT	7454.59	3.57	3.07	4.55	0.57	0.49	7.07	8.60	1.69
SimData, far distance, room 1										
	Orig	—	*2.67*	*2.38*	*4.58*	*0.38*	*0.35*	*6.68*	*9.24*	*1.61*
8	STFT 512/2048	7337.43 / 7477.34	3.16 / 2.28	2.88 / 2.05	5.05 / 4.82	0.52 / 0.37	0.43 / 0.34	8.50 / 9.16	8.73 / 10.57	1.90 / 1.87
8	STFChT	16730.48	2.47	2.04	5.01	0.34	0.30	10.16	11.26	2.11
2	STFT 512/2048	1872.86 / 1936.2	3.32 / 2.41	2.98 / 2.14	5.47 / 5.10	0.50 / 0.38	0.41 / 0.35	8.06 / 8.30	8.48 / 10.14	1.68 / 1.75
2	STFChT	7175.92	2.66	2.18	5.44	0.35	0.31	8.99	10.14	1.87
1	STFT 512/2048	975.37 / 1057.11	3.34 / 2.60	3.00 / 2.32	5.42 / 5.00	0.51 / 0.37	0.42 / 0.34	8.09 / 7.82	8.59 / 9.97	1.63 / 1.71
1	STFChT	6378.58	2.83	2.34	5.30	0.37	0.32	8.86	10.12	1.81
SimData, near distance, room 1										
	Orig	—	*1.99*	*1.68*	*4.50*	*0.35*	*0.33*	*8.12*	*10.72*	*2.14*
8	STFT 512/2048	7230.86 / 7500.12	2.92 / 1.97	2.71 / 1.74	4.78 / 4.62	0.46 / 0.34	0.38 / 0.32	8.72 / 9.73	8.83 / 10.56	2.47 / 2.41
8	STFChT	16807.41	2.12	1.68	4.80	0.28	0.25	10.83	11.62	2.89
2	STFT 512/2048	1842.15 / 1904.45	2.90 / 1.90	2.64 / 1.64	5.05 / 4.81	0.44 / 0.36	0.37 / 0.34	8.83 / 9.36	9.08 / 10.83	2.27 / 2.28
2	STFChT	7104.62	2.18	1.72	5.18	0.30	0.27	10.23	11.13	2.55
1	STFT 512/2048	954.31 / 1032.19	3.02 / 2.11	2.75 / 1.83	5.05 / 4.79	0.48 / 0.36	0.41 / 0.34	8.66 / 8.83	8.88 / 10.66	2.08 / 2.25
1	STFChT	6389.66	2.29	1.84	5.18	0.31	0.28	10.07	11.01	2.50
SimData, far distance, room 2										
	Orig	—	*5.21*	*5.04*	*2.97*	*0.75*	*0.63*	*1.04*	*1.77*	*1.19*
8	STFT 512/2048	7824.78 / 7954.58	4.31 / 4.25	3.85 / 3.87	4.72 / 4.52	0.72 / 0.47	0.60 / 0.40	7.43 / 4.98	8.29 / 8.28	1.37 / 1.56
8	STFChT	17652.06	3.82	3.29	4.85	0.58	0.49	7.84	9.71	1.56
2	STFT 512/2048	2064.93 / 2134.76	4.59 / 4.75	4.05 / 4.46	4.61 / 4.17	0.78 / 0.58	0.64 / 0.50	6.24 / 3.33	7.49 / 5.72	1.27 / 1.32
2	STFChT	9119.6	4.29	3.76	4.35	0.71	0.61	5.93	7.74	1.34
1	STFT 512/2048	1020.71 / 1075.23	4.59 / 4.98	4.08 / 4.75	4.46 / 3.80	0.82 / 0.68	0.69 / 0.57	5.26 / 2.20	6.70 / 3.71	1.26 / 1.26
1	STFChT	8177.94	4.53	4.01	3.93	0.79	0.68	5.01	6.68	1.28
SimData, near distance, room 2										
	Orig	—	*4.63*	*4.24*	*3.74*	*0.49*	*0.40*	*3.35*	*5.52*	*1.40*
8	STFT 512/2048	7545.94 / 7782.53	3.31 / 3.33	3.03 / 2.84	4.67 / 4.43	0.51 / 0.28	0.41 / 0.20	9.82 / 7.68	9.93 / 11.63	1.99 / 2.20
8	STFChT	18103.93	2.78	2.32	4.78	0.33	0.26	11.54	13.18	2.37
2	STFT 512/2048	1944.39 / 2010.15	3.69 / 4.07	3.41 / 3.55	4.85 / 4.45	0.60 / 0.36	0.51 / 0.28	8.69 / 5.80	8.94 / 9.53	1.59 / 1.72
2	STFChT	9007.05	3.30	2.84	4.84	0.46	0.38	9.42	11.26	1.84
1	STFT 512/2048	988.66 / 1052.96	3.95 / 4.48	3.66 / 4.00	4.82 / 4.30	0.67 / 0.44	0.57 / 0.35	7.59 / 4.55	8.23 / 7.57	1.48 / 1.57
1	STFChT	8267.96	3.64	3.17	4.70	0.54	0.45	8.25	10.08	1.70
SimData, far distance, room 3										
	Orig	—	*4.95*	*4.72*	*2.72*	*0.83*	*0.76*	*0.24*	*0.88*	*1.16*
8	STFT 512/2048	7526.24 / 7576.84	4.29 / 4.06	3.83 / 3.71	4.49 / 4.28	0.80 / 0.62	0.68 / 0.57	5.94 / 3.32	6.99 / 5.83	1.33 / 1.48
8	STFChT	16566.95	3.82	3.27	4.45	0.63	0.55	5.96	7.72	1.47
2	STFT 512/2048	2037.28 / 2106.18	4.56 / 4.45	4.03 / 4.14	4.39 / 3.92	0.85 / 0.71	0.74 / 0.66	4.68 / 2.00	6.10 / 4.02	1.22 / 1.27
2	STFChT	8555.67	4.23	3.67	3.96	0.73	0.66	4.33	6.02	1.27
1	STFT 512/2048	1055.52 / 1117.79	4.49 / 4.68	3.96 / 4.40	4.30 / 3.60	0.86 / 0.76	0.75 / 0.69	4.19 / 1.27	5.84 / 2.62	1.22 / 1.23
1	STFChT	7759.75	4.47	3.92	3.62	0.79	0.70	3.74	5.45	1.23
SimData, near distance, room 3										
	Orig	—	*4.37*	*4.03*	*3.56*	*0.65*	*0.58*	*2.27*	*4.20*	*1.37*
8	STFT 512/2048	7221.89 / 7442.87	3.39 / 3.18	3.10 / 2.78	4.92 / 4.69	0.65 / 0.47	0.53 / 0.42	7.96 / 5.85	8.07 / 9.02	1.93 / 2.10
8	STFChT	16934.76	2.79	2.33	5.05	0.43	0.36	8.92	10.28	2.23
2	STFT 512/2048	1973.75 / 2040.5	3.72 / 3.82	3.39 / 3.41	4.78 / 4.39	0.72 / 0.56	0.62 / 0.49	7.07 / 3.98	7.50 / 6.90	1.58 / 1.64
2	STFChT	8528.08	3.32	2.82	4.75	0.53	0.46	7.18	8.83	1.73
1	STFT 512/2048	1026.68 / 1114.32	3.80 / 4.18	3.44 / 3.79	4.70 / 4.21	0.74 / 0.61	0.64 / 0.53	6.57 / 3.21	7.51 / 5.76	1.51 / 1.52
1	STFChT	7753.63	3.64	3.12	4.55	0.60	0.52	6.49	8.24	1.60

RealData summary

Ch.	Method	Comp. Time (s)	SRMR
	Orig	—	*3.18*
8	STFT 512/2048	3080.52 / 3152.88	6.90 / 5.31
8	STFChT	5236.74	6.33
2	STFT 512/2048	852.84 / 934.18	6.29 / 4.57
2	STFChT	3036.49	5.24
1	STFT 512/2048	610.26 / 682.97	5.80 / 4.21
1	STFChT	2753.87	4.85

RealData, far distance

Ch.	Method	Comp. Time (s)	SRMR
	Orig	—	*3.19*
8	STFT 512/2048	2908.92 / 2977.25	6.99 / 5.56
8	STFChT	4962.74	6.52
2	STFT 512/2048	810.37 / 943.14	6.25 / 4.63
2	STFChT	2922.66	5.25
1	STFT 512/2048	754.06 / 870.11	5.73 / 4.24
1	STFChT	2624.72	4.82

RealData, near distance

Ch.	Method	Comp. Time (s)	SRMR
	Orig	—	*3.18*
8	STFT 512/2048	3252.12 / 3328.51	6.81 / 5.05
8	STFChT	5510.74	6.13
2	STFT 512/2048	895.3 / 925.23	6.33 / 4.51
2	STFChT	3150.32	5.22
1	STFT 512/2048	466.47 / 495.82	5.87 / 4.17
1	STFChT	2883.02	4.87

Fig. 7: Results for SimData and RealData evaluation sets.

VERB challenge dataset versus standard STFT processing. Processing in the STFChT domain resulted in less reverberation at the output without introducing artifacts, which concurs with substantial increase in the PESQ scores, the ITU-T standard for voice quality. We also provided insight as to why enhancement performance improves using the STFChT domain. The improvement gained by STFChT-based processing is an interesting result, and warrants further investigation. Further exploration may be fruitful, as combination of the fan-chirp or other coherent transforms with other methods for dereverberation and/or noise reduction may yield improved results.

8. REFERENCES

[1] E. A. P. Habets, "Speech dereverberation using statistical reverberation models," in *Speech Dereverberation*, Patrick A. Naylor and Nikolay D. Gaubitch, Eds. Springer, July 2010.

[2] M. Képesi and L. Weruaga, "Adaptive chirp-based time–frequency analysis of speech signals," *Speech Communication*, vol. 48, no. 5, pp. 474–492, May 2006.

[3] L. Weruaga and M. Képesi, "The fan-chirp transform for non-stationary harmonic signals," *Signal Processing*, vol. 87, no. 6, pp. 1504–1522, June 2007.

[4] P. Cancela, E. López, and M. Rocamora, "Fan chirp transform for music representation," in *Proc. International Conference On Digital Audio Effects (DAFx)*, Graz, Austria, 2010, p. 1–8.

[5] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, and R. Maas, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, 2013.

[6] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function GSC and postfiltering," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 6, pp. 561–571, 2004.

[7] R. Maas, E. A. P. Habets, A. Sehr, and W. Kellermann, "On the application of reverberation suppression to robust speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Japan, 2012, pp. 297–300.

[8] T. Nakatani, K. Kinoshita, and M. Miyoshi, "Harmonicity-based blind dereverberation for single-channel speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 80–95, 2007.

[9] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3, pp. 187–207, 1999.

[10] J. Droppo and A. Acero, "A fine pitch model for speech.," in *Proc. Interspeech*, Antwerp, Belgium, 2007, p. 2757–2760.

[11] R. Dunn and T. Quatieri, "Sinewave analysis/synthesis based on the fan-chirp tranform," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, 2007, pp. 247–250.

[12] R. Dunn, T. Quatieri, and N. Malyska, "Sinewave parameter estimation using the fast fan-chirp transform," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, 2009, pp. 349–352.

[13] Y. Pantazis, O. Rosec, and Y. Stylianou, "Adaptive AM-FM signal decomposition with application to speech analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 290–300, 2011.

[14] G. Degottex and Y. Stylianou, "Analysis and synthesis of speech using an adaptive full-band harmonic model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9-10, pp. 2085–2095, 2013.

[15] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[16] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[17] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Processing Letters*, vol. 9, no. 4, pp. 113–116, 2002.

[18] H. Löllmann, E. Yilmaz, M. Jeub, and P. Vary, "An improved algorithm for blind reverberation time estimation," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Tel Aviv, Israel, 2010, p. 1–4.

[19] L. Weruaga and M. Képesi, "Speech analysis with the fast chirp transform," in *Proc. European Signal Processing Conference (EUSIPCO)*, Vienna, Austria, 2004, p. 1011–1014.

[20] P. C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, June 2007.

[21] H. L. Van Trees, *Optimum Array Processing. Part IV of Detection, Estimation, and Modulation Theory*, Wiley-Interscience, New York, 2002.

[22] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Salt Lake City, UT, 2001, vol. 2, p. 749–752.

[23] T. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.

[24] ITU-T P.862.2, "Wideband extension to rec. p.862 for the assessment of wideband telephone networks and speech codecs," 2007.