SPEECH DEREVERBERATION BY CONSTRAINED AND REGULARIZED MULTI-CHANNEL SPECTRAL DECOMPOSITION: EVALUATED ON REVERB CHALLENGE

Meng Yu[1] *and Frank K. Soong*[2]

[1]Audience Inc.; [2]Microsoft Research Asia

myu@audience.com; frankkps@microsoft.com

ABSTRACT

We present our contribution to the REVERB Challenge in this paper. A multi-channel speech dereverberation system combines cross-channel cancellation and spectral decomposition. The reverberation is modeled as a convolution operation in the spectral domain. Using the generalized Kullback-Leibler (KL) divergence, we decompose the reverberant magnitude spectrum into clean magnitude spectrum convolved with a deconvolution filter. The magnitude spectrum is constrained and regularized by non-negativity and sparsity, respectively, while the deconvolution filter is constrained by non-negativity and cross-channel cancellation. Spectral decomposition of individual channels and cross-channel cancellation are jointly optimized by a multiplicative algorithm to achieve multi-channel speech dereverberation. Experimental evaluations on "speech enhancement task" are carried out according to the evaluation guidelines of the REVERB challenge, showing promising results. The objective metrics for measuring reverberation are investigated through the algorithm evaluation.

Keywords: REVERB challenge, Multichannel dereverberation, Spectral decomposition, Generalized KL divergence, Sparsity, Cross-channel cancellation.

1. INTRODUCTION

Reverberation is an acoustic phenomenon that happens when a sound wave is traveling in a physical enclosure and repeatedly reflected by the reflective surfaces of the enclosure. The multiple reflections cause the received sound (e.g. a distant microphone or a listener) to last even when original sound stops. The combinations of direct transmitted and reflected sound wave affect the intelligibility of speech or perception of the received acoustic wave. The objective comes to reduce reverberation and improve the quality of the signal. Substantial progress has been made in the field of reverberant speech signal processing, including both single- and multi-channel dereverberation techniques. Despite these studies, existing reverberant speech enhancement algorithms, however, do not reach a performance level demanded by many practical applications. Reverberation causes a noticeable change in speech quality. Berkley and Allen [1] identified that two physical variables, reverberation time T_{60} and the talker-listener distance, are important for the reverberant speech quality. The universally accepted set of objective quality measures has not been fully established for evaluating reverberant speech enhancement algorithms. The REVERB challenge is designed to evaluate state-of-the-art algorithms and direct the researchers to have comprehensive understanding of evaluation metrics for dereverberation algorithms.

Our contribution focuses on recovering the subband spectrum of an original speech signal from its reverberant version. The problem is formulated as a blind deconvolution problem with non-negative constraints, regularized by the sparse nature of speech magnitude spectra. However, single channel decomposition mathematically allows too much freedom, which possibly makes the solution deviate from the true solution. According to our paper [2], we constructed an effective cost function by combining multi-channel based cross-channel cancellation and spectral decomposition on individual channels to achieve multi-channel dereverberation. By investigating the criterion of decomposition, we proposed to incorporate generalized KL divergence as the decomposition metric. The outline of this paper is as follows. In Section 2 the related work is described and the subband deconvolution is verified to be equivalent to that in time domain. In Section 3 the main contribution of multichannel speech dereverberation method is presented. The experiment setup for REVERB challenge and evaluations are presented and discussed in Section 4. We conclude the algorithm in Section 5.

2. SUBBAND SPECTRAL DECONVOLUTION

The frequency domain blind dereverberation for reverberant speech mixtures has been extensively studied, because it can learn each frequency individually and selectively with much less computation under the assumption that convolution in the time domain can be represented as multiplication in the frequency domain. However, when we perform a frequency domain decomposition via short-time Fourier transform (STFT), we are often aware that the source separation or deconvolution filter should be longer enough than the conventional frame length (10 ms to 30 ms) for speech processing, because the reverberation time, typically 200-300 ms even in a small office environment, far exceeds the frame length. On the other hand, if we increase the frame length to make the filter length long enough under the same assumption, it results in decreasing the super-Guassianity of each frequency channel and consequently deteriorate the blind source separation or dereverberation performance. This fundamental limitation on frequency domain processing has been reported [3], yet still recognized as unavoidable limitation. The exact deconvolution operation in the STFT domain is shown below, demonstrating that it is similar to the time domain deconvolution in each frequency bin [4].

Proposition 2.1 *Deconvolution in the time domain by a filter with longer length than a frame length used for subband decomposition is equivalent to the deconvolution in each subband again.*

$$Z(e^{j\omega_k}) = \sum_{l=0}^{N_H-1} H^l(e^{j\omega_k}) S^{\cdot-l}(e^{j\omega_k}), \tag{2.1}$$

where k is a frequency index, superscript $\cdot - l$ stands for the past frame, which is l frames before the current frame, and N_H represents the total number of frames to include the deconvolution filter length sufficiently, and the l^{th} tap subband domain convolution filter

$$H^l(e^{j\omega_k}) = \sum_{t=-\infty}^{\infty} h^l[t] e^{-j\omega_k t}, \tag{2.2}$$

where

$$h^l[t] = h[t]w[lR - t], \quad -\infty < t < \infty, \tag{2.3}$$

where $w[t]$ is a proper window function and from the overlap-add context

$$h[t] = \sum_{l=0}^{N_H-1} h^l[t]. \tag{2.4}$$

Proof. The equivalence between the subband domain deconvolution operation and the original time domain deconvolution can be demonstrated by taking an inverse Fourier transform on (2.1) and summing them over all possible frames, which is eventually shown to be equivalent to the results when we simply implement the deconvolution in the time domain with the original incoming signal and filter before applying subband decomposition. In the time domain, a deconvolution can be performed using a filter $h[t]$ with an incoming signal $s[t]$.

$$z[t] = \sum_{\tau=-\infty}^{\infty} s[\tau]h[t-\tau]. \tag{2.5}$$

From the perspective of subband deconvolution, the deconvolved signal can be obtained as following

$$\begin{aligned}
\hat{s}[t] &= \sum_{r=-\infty}^{\infty} \frac{1}{K} \sum_{k=0}^{K-1} \sum_{l=0}^{N_H-1} H^l(e^{j\omega_k}) S^{r-l}(e^{j\omega_k}) e^{j\omega_k t} \\
&= \sum_{r=-\infty}^{\infty} \frac{1}{K} \sum_{k=0}^{K-1} \sum_{l=0}^{N_H-1} H^l(e^{j\omega_k}) \\
&\quad \left[\sum_{\tau=-\infty}^{\infty} s[\tau] w[(r-l)R-\tau] e^{-j\omega_k \tau} \right] e^{j\omega_k t} \\
&= \sum_{\tau=-\infty}^{\infty} s[\tau] \sum_{l=0}^{N_H-1} \left[\frac{1}{K} \sum_{k=0}^{K-1} H^l(e^{j\omega_k}) e^{j\omega_k (t-\tau)} \right] \\
&\quad \sum_{r=-\infty}^{\infty} w[(r-l)R-\tau] \\
&= \sum_{\tau=-\infty}^{\infty} s[\tau] \sum_{l=0}^{N_H-1} h^l[t-\tau] \sum_{r=-\infty}^{\infty} w[(r-l)R-\tau] \\
&= \sum_{\tau=-\infty}^{\infty} s[\tau] h^l[t-\tau]
\end{aligned}$$

Note that with a careful choice of the window function, we can fulfill

$$\sum_{r=-\infty}^{\infty} w[(r-l)R-\tau] = 1 \tag{2.6}$$

Therefore, $s[t] = \hat{s}[t]$, and the subband domain deconvolution represented in (2.1) is a correct way of implementing a deconvolution in the subband domain.

The problem in [5, 6] is formulated as a single-channel subband blind deconvolution problem. The method in [6] differs from [5] in the domain of the model: instead of power spectrum in Fourier spectral domain, it works in Gammatone spectral domain, based on magnitude spectrum. The two methods try to estimate the (power) spectral magnitude of clean speech S through a decomposition of the reverberated speech (power) spectral magnitude X into its convolutive components S and H, where H is the (power) spectral magnitude of the room impulse response. The least-squares error criterion, i.e. l_2 norm, is formulated in [5, 6] to achieve the decomposition. In general, reverberation compensation algorithms should not require a priori knowledge of nature of the reverberation. The model in [5, 6] represents the reverberation effects as the filter H, whose parameters are not observed directly. Thus, the problem of decomposition is highly unconstrained. There exist infinitely many decompositions of X into S and H. To constrain the feasible space, two constraints are exploited in [5, 6]. One is that the (power) spectral magnitude are non-negative, *i.e.* all the elements in S and H are ≥ 0. The second assumption is the clean (power) spectral magnitude S is sparse.

By considering the mathematical formulation, it is assumed that actual observation sequence is $Z[n,k]$. $Z[n,k] \approx X[n,k] = S[n,k] * H[n,k]$, where $*$ is convolution, n denotes index of time frame and k denotes the frequency bin. Since the length of impulse response filter in time domain might be longer than the window for short time Fourier transform (STFT), H in magnitude spectral domain remains to be convolved with the signal S. The difference between Z and X can result from observation noise or from the error in decomposing Z into convolutional components S and H. The objective is to minimize the mean-squared error between Z and X. After imposing the non-negativity and sparsity constraints, the objective function is defined in each frequency bin as

$$\begin{aligned} Min.E(S,H) = &\sum_i (Z[i,k] - \sum_m S[m,k]H[i-m,k])^2 \\ &+ \lambda \sum_i S[i,k] \qquad (2.7) \\ s.t.\ \ S[n,k] \geq 0, &H[n,k] \geq 0, \sum_n H[n,k] = 1 \end{aligned}$$

where $H[n,k]$ is constrained to sum to 1 to avoid scaling problems. l_1 norm is selected to apply sparsity regularization on S.

The above model is an approximation and will in general incur an approximation error e as follows

$$X[n,k] = \hat{X}[n,k] + e[n,k] = S[n,k] * H[n,k] + e[n,k] \qquad (2.8)$$

It is empirically observed in [6] that the approximation error e is lower in the magnitude spectral domain than in the power spectral domain. Thus, working in the magnitude spectral domain incurs lower error. Experimentally they found that the power of e is usually about 13 dB below the power of $\hat{X}$ in the power spectral domain. In contrast, in the magnitude spectral domain, they observed an approximation error attenuation of 17dB. Thus, working in the magnitude spectral domain incurs lower error.

3. MULTICHANNEL FORMULATION

Instead of inferring the filter H parameters through the observed single channel data X, we attempt to jointly estimate multi-channel (without loss of generality, we use dual-channel as an example) filters $H_i, i = 1,2$, and clean magnitude spectrum S by the reverberant speech magnitude spectrum $X_i, i = 1,2$. This problem is however highly unconstrained, which renders infinitely many decompositions of X_i into S and H_i. By inheriting the constraints from above single channel model, extra new constraints are incorporated for building the multi-channel dereverberation model.

- Different channels estimate the same magnitude spectrum of clean speech S.
- Cross-channel cancellation enforces the filters $H_i, i = 1,2$, to resolve the spatial difference between channels. The cross-channel cancellation error is to be minimized.

Back to time domain, the two microphones capture the reverberant speech as

$$x_i[n] = s[n] * h_i[n], \ \ i = 1,2 \qquad (3.9)$$

Suppose $h_i[n], i = 1,2$ can be successfully resolved, then by performing cross-convolution and subtraction, we have

$$\begin{aligned} &x_1[n] * h_2[n] - x_2[n] * h_1[n] \\ =&s[n] * h_1[n] * h_2[n] - s[n] * h_2[n] * h_1[n] = 0 \quad (3.10) \end{aligned}$$

The cross-channel cancellation in spectral domain becomes

$$\begin{aligned} X_1[n,k]*H_2[n,k] - X_2[n,k] * H_1[n,k] = S[n,k] * H_1[n,k] \\ * H_2[n,k] - S[n,k] * H_2[n,k] * H_1[n,k] = 0 \\ (3.11) \end{aligned}$$

After imposing the constraints, the objective function in each frequency bin becomes

$$\begin{aligned} E(S,H_1,H_2) = &\sum_{j=1}^{2} \sum_i (Z_j[i,k] - \sum_m S[m,k]H_j[i-m,k])^2 \\ &+ \beta \sum_i (\sum_m Z_1[m,k]H_2[i-m,k] \\ &- \sum_m Z_2[m,k]H_1[i-m,k])^2 \\ &+ \lambda \sum_i S[i,k] \\ s.t.\ \ S[n,k] \geq 0, &H_j[n,k] \geq 0, \sum_n H_j[n,k] = 1, j = 1,2 \\ &(3.12) \end{aligned}$$

The criterion metric used for spectral decomposition is l_2 norm in (3.12). However, it can be replaced by any appropriate metric D. $D(x|y)$ for generalized KL divergence is defined as $D(x|y) = x \log \frac{x}{y} - x + y$, where x corresponds to Z as the observation, while y corresponds to $S * H_i$ ($*$ is convolution) as the underlying model. The optimization process for spectral decomposition can be understood as maximizing the probability that the observation Z is generated by the underlying model with parameter $S * H_i$. l_2 norm in (2.7) and (3.12) indicates Gaussian distribution of the maximum likelihood function, that is equivalent to the least mean squares estimation between Z and $S * H_i$. However, despite the fact that Z is a spectrum distribution density, it is desirable that the likelihood function is only defined on the non-negative axis. By an appropriate normalization, the Poisson distribution is a representative example of such a probability density function. On the other hand, the generalized KL divergence is asymmetric, giving more penalty to positive errors, and thus emphasizes the goodness of fitting between spectral peaks [7]. By plugging in the exact form of divergence, the objective function is

$$\begin{aligned} Min.E = &\sum_{j=1}^{2}\sum_{i}(Z_j[i,k]\log\frac{Z_j[i,k]}{\sum_m S[m,k]H_j[i-m,k]} \\ &- Z_j[i,k] + \sum_m S[m,k]H_j[i-m,k]) \\ &+ \beta\sum_i(\sum_m Z_1[m,k]H_2[i-m,k] \\ &- \sum_m Z_2[m,k]H_1[i-m,k])^2 + \lambda\sum_i S[i,k] \\ s.t.\ \ S[n,k] &\geq 0, H_j[n,k] \geq 0, \sum_n H_j[n,k] = 1, j = 1,2 \end{aligned} \tag{3.13}$$

Next, the goal is to derive an iterative estimation algorithm formally equivalent to the EM algorithm without making use of Bayes' rule. Guided by the idea of NMF [8] and EM algorithm, we derive an efficient iterative algorithm that ensures a monotonic decrease (convergence to a stationary point) in the objective function and simultaneously, the non-negativity of the parameters. The objective function E is a function of variable S and $H_j, j = 1,2$. Fix two of the three variables, E is a function of one variable, denoted as $E(x)$. We need to find an auxiliary function $G(x,x')$ for $E(x)$ such that $G(x,x') \geq E(x)$, and $G(x,x) = E(x)$. It is obvious that $E(x)$ is non-increasing under the update

$$x^{t+1} = \arg\min_x G(x,x^t) \tag{3.14}$$

Since $E(x^{t+1}) \leq G(x^{t+1},x^t) \leq G(x^t,x^t) = E(x^t)$, the above statement holds. By using Jensen's inequality on the convex logarithm function, we construct the auxiliary functions $G(S,S^t)$, $G(H_j,H_j^t)$ for $E(S)$ and $E(H_j)$, respectively.

$$\begin{aligned} G(S,S^t) = &\sum_{j=1}^{2}\sum_i\sum_m S[m,k]H_j[i-m,k] \\ &-\sum_{j=1}^{2}\sum_i\sum_m Z_j[i,k]\frac{S^t[m,k]H_j[i-m,k]}{\sum_{m'}S^t[m',k]H_j[i-m',k]} \\ &(\log S[m,k]H_j[i-m,k] - \log\frac{S^t[m,k]H_j[i-m,k]}{\sum_{m'}S^t[m',k]H_j[i-m',k]}) \\ &+\frac{\lambda}{2}\sum_i(\frac{S[i,k]^2}{S^t[i,k]} + S^t[i,k]) \end{aligned} \tag{3.15}$$

The auxiliary function for $E(H_j)$ can be derived similarly. In the auxiliary function above, we omit the terms without the corresponding variable, since those terms vanish while taking derivative w.r.t. the desired variable. The iterative algorithm for solving (3.13) is shown in (3.16). By tuning the trade-off parameters β and λ, we can achieve good quality for the estimated speech signal.

```
**Initialize** $S = Z_j, j = 1\ or\ 2$
**For** $Iter = 1 : MaxIter$
  (i) $X_j[i,k] = S[i,k] * H_j[i,k], \quad j = 1,2$
  (ii) $W_1[i,k] = Z_1[i,k] * H_2[i,k],\ W_2[i,k] = Z_2[i,k] * H_1[i,k]$
  (iii) $S[n,k] \leftarrow S[n,k]\frac{\sum_{j=1}^{2}\sum_i H_j[i-n,k]\frac{Z_j[i,k]}{X_j[i,k]}}{\sum_{j=1}^{2}\sum_i H_j[i-n,k] + \lambda}$
  (iv) $H_1[n,k] \leftarrow H_1[n,k]\frac{\sum_i S[i-n,k]\frac{Z_1[i,k]}{X_1[i,k]} + \beta\sum_i Z_2[i-n,k]W_1[i,k]}{\sum_i S[i-n,k] + \beta\sum_i Z_2[i-n,k]W_2[i,k]}$
  (v) $H_2[n,k] \leftarrow H_2[n,k]\frac{\sum_i S[i-n,k]\frac{Z_2[i,k]}{X_2[i,k]} + \beta\sum_i Z_1[i-n,k]W_2[i,k]}{\sum_i S[i-n,k] + \beta\sum_i Z_1[i-n,k]W_1[i,k]}$
  (vi) $H_2[n,k] \leftarrow \frac{H_2[n,k]}{\sum_i H_1[i,k]},\ H_1[n,k] \leftarrow \frac{H_1[n,k]}{\sum_i H_1[i,k]}$
**end For**                                                          (3.16)
```

By extending the dual-channel model to any M-channel model, the optimization formula (3.13) is modified as

$$\begin{aligned} Min.E = &\sum_{j=1}^{M}\sum_i D(Z_j[i,k]|\sum_m S[m,k]H_j[i-m,k]) \\ &+ \beta\sum_{1\leq p<q\leq M}\sum_i(\sum_m Z_p[m,k]H_q[i-m,k] \\ &- \sum_m Z_q[m,k]H_p[i-m,k])^2 + \lambda\sum_i S[i,k] \end{aligned} \tag{3.17}$$

$$s.t.\ \ S[n,k] \geq 0, H_j[n,k] \geq 0, \sum_n H_j[n,k] = 1, j = 1,2,...,M$$

The schematic diagram of the proposed multi-channel dereverberation algorithm is shown in Fig. 1.

4. EVALUATION

Experimental results shown in this paper are all carried out according to the guidelines of the REVERB challenge [9]. We contribute on the speech enhancement challenge task of enhancing noisy & reverberant speech with multi-channel technique and evaluating the enhanced utterances in terms of objective evaluation metrics in this paper. The operation environment is Matlab 2013a in Windows 7 with CPU 3.30GHz (8 cores), 2GB RAM.

4.1. Metrics

According to REVERB challenge, cepstrum distance (CD), log likelihood ratio (LLR), frequency-weighted segmental

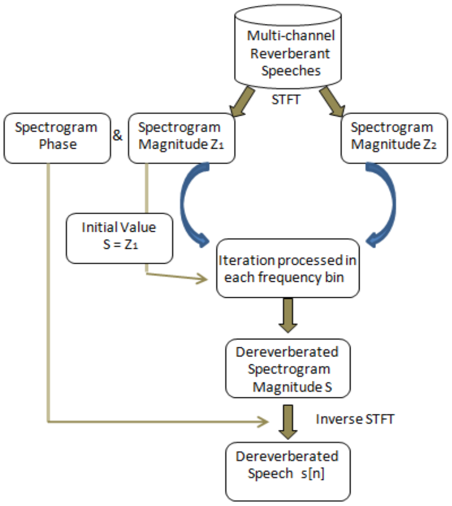
Multi-channel Reverberant Speeches
STFT
Spectrogram Phase
&
Spectrogram Magnitude Z₁
Spectrogram Magnitude Z₂
Initial Value S = Z₁
Iteration processed in each frequency bin
Dereverberated Spectrogram Magnitude S
Inverse STFT
Dereverberated Speech s[n]

Figure 1: *The schematic diagram of constrained multi-channel (dual channel as an example) speech dereverberation based on sparse and non-negative spectral decomposition and cross-channel cancellation*

SNR (FWSNR) [10], speech-to-reverberation modulation energy ratio (SRMR) [11], computational cost (wall clock time in sec., WCT), and perceptual evaluation of speech quality (PESQ) are incorporated for the system evaluation.

4.2. Dataset

All the reverberant utterances are provided as 1-channel, 2-channel, and 8-channel recordings for development test set and evaluation test set, respectively [9]. The whole dataset for development and evaluation contains SimData and RealData. SimData are utterances from the WSJCAM0 corpus [12], which are convolved by room impulse responses (RIRs) measured in different rooms. Recorded background noise is added to the reverberant test data at a fixed signal-to-noise ratio (SNR). It simulates 6 different reverberation conditions: 3 rooms with different volumes (small, medium and large size), 2 types of distances between a speaker and a microphone array (near=50cm and far=200cm). RealData are utterances from the MC-WSJ-AV corpus [13], which consists of utterances recorded in a noisy and reverberant room. It contains 2 reverberation conditions: 1 room (large size), 2 types of distances between a speaker and a microphone array (near= 100cm and far= 250cm).Recordings are measured with an array (8-ch circular array with diameter of 20 cm, uniformly distributed omni-directional microphones) that has the same array geometry as the ones used for SimData. For the SimData, noise is added to the reverberant speech with SNR of 20dB. Rereverberation time (T60) of small, medium, and large-size rooms are about 0.25s, 0.5s, 0.7s, respectively. On the other hand, a meeting room used for the RealData recording has reverberation time of 0.7s. For the SimData and RealData, it can be assumed that the speakers stay still within an utterance.

4.3. Parameter setup

The STFT is computed using a Hamming window that is 64 ms long with a 48 ms overlap. λ is set to be the proportionate weights $\frac{\sum_{i=1}^{N} Z[i,k]\times 10^{-3}}{N}$ as the processing doesn't starts until the whole sentence is read in. N is the number of time frames that are taken for averaging, and set to be the total number of frames within a sentence. β is set at 10. S is initially equal to Z. H_j is initially set to be an exponentially decaying envelope, with the length 12 time frames, which is approximately 240 ms in time domain.

4.4. Evaluation results and discussion

The proposed dereverberation algorithm processes utterance streams one by one. The buffer size is simply the sentence size, though this could be reduced to tens of ms, i.e. a few time frames to speed up the computation. Meanwhile, as our algorithm runs independently in each subband, it could be parallel executed on multi-thread. The computation time can be further reduced as much as $\frac{1}{N}$ of single-thread processing, where N is the number of the threads. A noise suppression post processing (optimally-modified log-spectral amplitude speech estimator [14]) is applied to the dereverberated signal to suppress the background noise as an option. Table 1 lists the performance measurements of the proposed algorithm system for the REVERB challenge. The evaluation were carried out based on 4-channel dereverberation plus post denoise processing. The 4 microphones correspond to channel 1 to 4 of the provided SimData and RealData, while room 1, 2 and 3 are the rooms with small, medium and large size for SimData, respectively (Room 1 is large room regarding RealData). The reference wall clock time is calculated based on a two-microphone beamforming algorithm in single thread (provided by the REVERB challenge). As the cost of our algorithm varies while using different number of threads, for a fair and clear illustration, we evaluate our dereverberation algorithm by single-thread processing as well. According to Table 1, the real time factor of our 4-channel and single-thread based algorithm is around 1.95, compared to 0.02 of reference processing, while this number for 2-channel and single-thread based our algorithm is around 1.23 in the experiment done in Fig. 3 below. Besides WCT, all other metrics for speech enhancement task are used to evaluate the proposed algorithm system. From room1 to room3 of SimData, the enhancement becomes more and more significant as T_{60} increases from small room to large room. As introduced in Section 1, besides reverberation time T_{60}, another factor

Table 1: *Performance measurements of speech enhancement task, comparing original reverberant & noisy streams with enhanced streams*

		SimData							RealData		
		Room 1		Room 2		Room 3		Ave.	Room 1		Ave.
Config	Metrics	Near	Far	Near	Far	Near	Far	-	Near	Far	-
4 ch dereverberation + denoise Length of deconvolution filter H is 12 frames	CD org	1.99	2.67	4.63	5.21	4.38	4.96	3.97	-	-	-
	CD enh	3.39	3.50	3.58	4.16	3.60	4.21	3.74	-	-	-
	LLR org	0.35	0.38	0.49	0.75	0.65	0.84	0.58	-	-	-
	LLR enh	0.56	0.51	0.67	0.82	0.72	0.79	0.68	-	-	-
	FWSnr org	8.12	6.68	3.35	1.04	2.27	0.24	3.62	-	-	-
	FWSnr enh	8.18	8.55	8.06	6.84	7.15	6.00	7.46	-	-	-
	SRMR org	4.50	4.58	3.74	2.97	3.57	2.73	3.68	3.17	3.19	3.18
	SRMR enh	4.79	5.04	4.65	4.36	4.66	4.09	4.60	5.95	6.02	5.98
	PESQ org	2.14	1.61	1.40	1.19	1.37	1.17	1.48	-	-	-
	PESQ enh	1.92	1.86	1.68	1.45	1.66	1.36	1.66	-	-	-
	WCT ref	62.12	61.97	64.05	63.64	61.96	61.50	62.54	28.76	26.64	27.70
	WCT	5578.9	5534.1	5994.4	5829.8	5645.5	5631.5	5702.4	2469.5	2229.9	2349.7

degrading speech quality is the speaker-to-microphone distance. This could be seen from the table by comparing performance between *Near* and *Far*. It is also known that the dereverberation and denoise algorithms make speech attenuated and distorted more significantly in the low reverberation condition. Regression can be found in *Room1 Near*, such as PESQ. Constant improvement of the proposed algorithm is proved by the metrics, such as SRMR, FWSNR and PESQ. Table 2 investigates the correlation between the five objective metrics for measuring dereverberation performance. The correlation is measured by the sequences of enhancement δ's in 6 SimData conditions presented in Table 1 ($\delta = enh - org$ for FWSNR, SRMR and PESQ as larger *enh* indicates good performance, while $\delta = org - enh$ for CD and LLR as smaller *enh* indicates good signal quality). In Table 2, those numbers in bold indicate high correlation. It shows that FWSNR, SRMR and PESQ behaves similarly for measuring the presented dereverberation algorithm, which are likely more appropriate for reverberant speech quality measurement than CD and LLR.

Table 2: *Correlation between metrics in the evaluation.*

	CD	LLR	FWSNR	SRMR	PESQ
CD	-	0.15	0.66	0.58	0.69
LLR	0.15	-	-0.45	-0.49	-0.33
FWSNR	0.66	-0.45	-	**0.99**	**0.87**
SRMR	0.58	-0.49	**0.99**	-	**0.82**
PESQ	0.69	-0.33	**0.87**	**0.82**	-

We compared the 4-channel based dereverberation (deconvolution filter H length is 12 frames) plus post denoise processing with other two variants, 4-channel based dereverberation (deconvolution filter H length is 36 frames) plus post denoise processing and 4-channel based dereverberation (deconvolution filter H length is 12 frames), respectively. The algorithm configuration in Table 1 performs best among the three variants. It shows that post denoise processing improves the speech quality in most metrics except for LLR by comparing blue and red curves. Compared with 36 frames, 12 frames for deconvolution filter H is good enough in all metrics but SRMR, and shorter filters reduce the computation and memory significantly.

Fig. 3 illustrates the performance difference between 2-channel and 4-channel based algorithms. Both of them carry out dereverberation plus denoise processing with the length of deconvolution filter H as 12 time frames. Apparently, 4-channel algorithm wins in all metrics for almost all the acoustic conditions. However, we should realize that real time factor is lifted up to 1.95 from 1.23 in single-thread computing.

5. CONCLUSION

We have presented the constrained multi-channel speech dereverberation method based on spectral decomposition under generalized KL divergence and cross-channel cancellation. An iterative algorithm is presented for the optimization. The proposed multi-channel speech dereverberation system could substantially improve the speech quality on both simulated data and real data of REVERB challenge. Various metrics are investigated based on the presented algorithm, among which FWSNR, SRMR and PESQ are highly

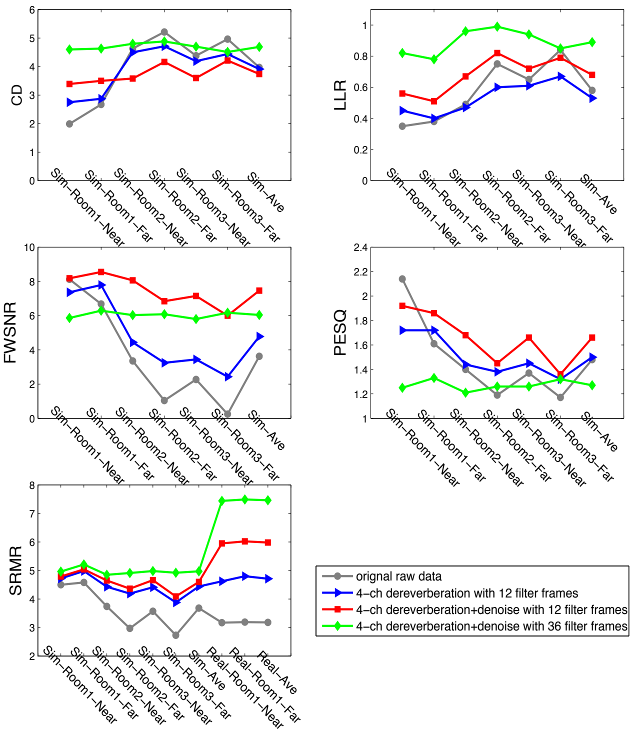

Figure 2: *Performance comparison with/without post processing, and shorter/longer deconvolution filter length. For simplicity, sim-room1-far indicates SimData in room 1(small size) with far speaker-to-microphone distance.*

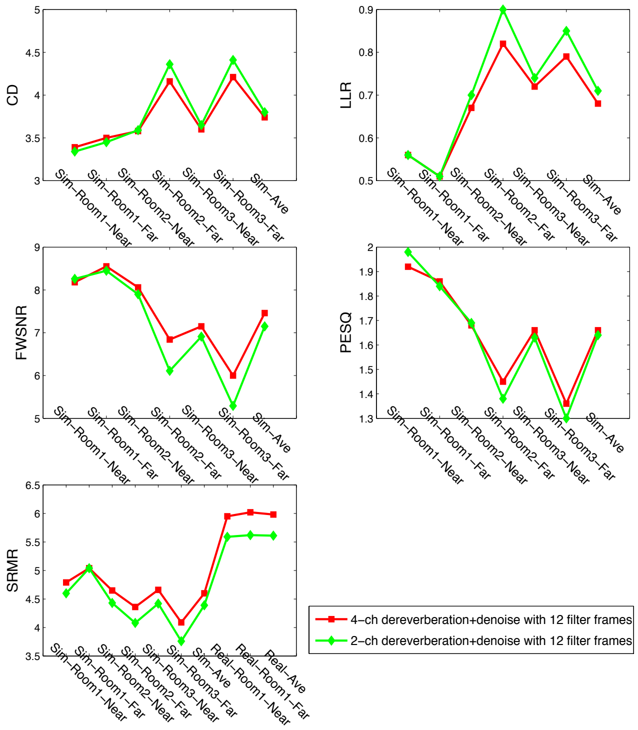

Figure 3: *Performance comparison between 2-channel and 4-channel processings based on dereverberation+denoise post processing with 12 frames' deconvolution filter H.*

correlated and proper for reverberation measurement. Future work should extend the current evaluation of speech enhancement to ASR evaluation.

6. REFERENCES

[1] D. A. Berkley and J. B. Allen, "Normal listening in typical rooms: the physical and psychophysical correlates of reverberation," *in Acoustical Factors Affecting Hearing Aid Performance, 2nd ed, G. A. Studebaker and I. Hochberg, Eds. Needham Heights, MA: Allyn and Bacon*, pp. 3-14, 1993.

[2] M. Yu, F. K. Soong, "Constrained multichannel speech dereverberation", *Interspeech*, 2012.

[3] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 109-116, 2005.

[4] L-H Kim and M Hasegawa-Johnson, "Toward Overcoming Fundamental Limitation in Frequency-Domain Blind Source Separation for Reverberant Speech Mixtures", the 44th Asilomar Conference on Signals, Systems and Computers.

[5] H. Kameoka, T. Nakatani, T. Yoshioka, "Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms", *in Proc. IEEE ICASSP*, pp. 45-48, 2009.

[6] K. Kumar, R. Singh, B. Raj, R. Stern, "Gammatone sub-band magnitude-domain dereverberation for ASR", *in Proc. IEEE ICASSP*, pp. 4604-4607, 2011.

[7] H. Kameoka, "Statistical Approach to Multipitch Analysis," Ph.D. Thesis, University of Tokyo, 2007.

[8] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization", *NIPS* 2000.

[9] K. Kinoshita; M. Delcroix; T. Yoshioka; T. Nakatani; E. Habets; R. Haeb-Umbach; V. Leutnant; A. Sehr; W. Kellermann; R. Maas; S. Gannot; and B. Raj, "The REVERB Challenge: A Common Evaluation Framework for Dereverberation and Recognition of Rever-

berant Speech," *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA-13)*, 2013.

[10] Hu and Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE T-ASLP*, 16(1), pp. 229-238, 2008

[11] Falk, et al., "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE T-ASLP*, 18(7), pp. 1766-1774, 2010

[12] T. Robinson, J. Fransen, D. Pye and J. Foote and S. Renals, "Wsjcam0: A British English Speech Corpus For Large Vocabulary Continuous Speech Recognition", *In Proc. ICASSP*, pp.81-84, 1995

[13] M. Lincoln, I. McCowan, J. Vepa and H.K. Maganti, "The multi-channel Wall Street Journal audio visual corpus (MC-WSJ-AV): Specification and initial experiments", *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2005

[14] I. Cohen, "Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging", *IEEE Trans. Speech and Audio Processing*, Vol. 11, pp. 466-475, 2003.