TOWARDS SPEECH RECOGNITION ORIENTED DEREVERBERATION

Pamornpol Jinachitra *

CCRMA
Stanford University
pj97@stanford.edu

Ramon E.Prieto

Toyota InfoTechnology Center U.S.A.
Human Machine Interaction Group
rprieto@us.toyota-itc.com

ABSTRACT

We show the effect of reverberation on the speech recognition performance in a far-field microphone. Given a reference of close-talk microphone signal, an improvement is shown using asymmetric non-causal inverse filter both in a synthetic and real room environment. Its variants in time and frequency domain are also presented and compared with other existing techniques. We argue for the approaches which specifically consider recognition performance as a goal in deriving the dereverberating schemes, with evaluation on real room recording recognition, as a future direction for solving reverberation problem in speech recognition.

1. INTRODUCTION

The performance of a speech recognizer on utterances from a far-away microphone suffers greatly from reverberation effect as well as low signal-to-noise ratio. This prevents an effective use of speech recognition engine in hands-free environment, for example, in a car cockpit. A number of approaches to tackle this problem have been proposed over the last few decades, including matched training and the use of reverberation-robust features. A number of algorithms which try to undo the effect of reverberation have also been proposed and studied extensively. The problem of non-minimum phase system inversion makes inverse filtering difficult. Each algorithm in this area deals with this problem differently but most only direct the result towards equalizing room or loud-speaker response for listening purpose [1] [2]. Though such perceptual goal usually relates to better recognition result, this is not necessarily the case. Algorithms which specifically try to improve recognition performance directly, with attention to a speech recognizer's characteristics and behaviors, might be desired. Few examples in this direction include [3] [4] and [5]. Few also show results in real speech recognition tests. While the current trend in solving reverberation problem seems to be in the direction of using multiple microphones [6] [7], a single microphone situation is still worthy of consideration. Besides lower cost, some blind source separation algorithms which try to avoid source whitening will need single channel dereverberation post processing.

Given the above motivations, we investigate the problem of single channel dereverberation with the goal of improving speech recognition result as a primary objective. To better understand what happens, we look at dereverberation given a reference speech signal from a close-talk microphone. This is more practical in many situations than using non-speech signal as a probe signal. Though it is not quite a simple system identification problem due to speech characteristics such as sparseness and non-stationarity.

2. SPEECH DEREVERBERATION WITH REFERENCE

In this paper, we use a speech recognizer engine version 8.0 from Nuance with no additional "out-of-box" modification. The test set utterances are limited vocabulary isolated words recorded by a close-talk and far-away microphone in an office room environment. Isolated word utterances were chosen to avoid variation of acoustic path for the purpose of our study. The room dimension is 3.4 x 3.7 x 2.65 m while the microphones separation is 40 cm. The total number of utterances is 1380 from 17 subjects. Care was taken to make sure they are long enough for reliable filter estimation. Only the speech segment is used in estimation for high SNR samples.

As a baseline reference, we performed a recognition test on close-talk and far-field speech signals. The results of recognition error rate are shown in Table 1. The last column shows the result of the close-talk speech signal mixed with more noise to give the same SNR as the far-field. The result indicates that most degradation comes from the reverberation.

2.1. Linear Least-squares (LS) Solution

A linear time-invariant system parameters can be estimated using least-squares (LS) technique from the following equa-

*P.Jinachitra is sponsored by Toyota ITC, Palo Alto, USA.

Signal	Close-talk	Far-field	Noisy close-talk
Error rate	9.1%	17.1%	9.8%

Table 1. Speech recognition error rate for close-talk, far-field and close-talk with the same SNR as far-field

tion [8]

$$\mathbf{Rw} = \mathbf{d} \tag{1}$$

In this paper, the inverse system is modeled as an FIR filter directly. So $\mathbf{w}$ will be the required inverse filter coefficients, $\mathbf{d}$ is the close-talk signal while $\mathbf{R}$ contains far-field data.

By directly modeling the inverse system, we avoid accumulated error and the problem of inverting a deep notch commonly associated with inverse filtering approaches. However, since the room reverberation in general is non-minimum phase, the structure needs to allow for this compensation in some form e.g. using non-causal filter. However, letting the filter have long anti-causal taps raises the problem of pre-echo which brings down the performance of the speech recognizer considerably (similar to Haas effect but reverse in time order).

In this work, the filter is allowed to have shorter anti-causal part to keep the pre-echo subdued, resulting in an asymmetric shape (practically implemented causally with delay). Similar truncation is also used to deal with pre-echo problem in a matched filtering algorithm for microphone array in [7].

2.2. Mean Squared Error (MSE) Solution

Instead of fitting the speech data themselves using least-square, statistical minimum mean square error (MMSE) solution can be derived by replacing $\mathbf{R}$ with, $\mathbf{R_{yy}}$, the auto-correlation matrix of $\mathbf{y}(t)$ and $\mathbf{d}$ by, $\mathbf{R_{yx}}$ the cross correlation of $\mathbf{y}(t)$ and $\mathbf{x}(t)$ [8]. This is the so-called Wiener filter. The computational complexity of matrix inversion is greatly reduced because of the Toeplitz structure in $\mathbf{R_{yy}}$. The recognition results have been verified to be more or less the same as those of LS in our experiments (within 1% difference). Therefore, only recognition results from MMSE solution are shown for time-domain solutions in Figure 1 and 2.

2.3. Weighted Least Squares (WLS)

In [3], it is demonstrated that long reverberation has more negative impact on recognition than the signal-to-reverberation ratio. The binary-weighted least square (BWLS) solution is then proposed which penalizes the squared error more in the tail region while leaving as "don't care", a period of time right after the direct signal arrives. It is assumed that the forward path impulse response is known in advance. The estimated filter was shown to give a better recognition performance than conventional least squares.

Error rate	T_e=18.7ms	T_e=25ms
Early reverb only	11.17%	11.17%
Late reverb only	18.06%	17.4%

Table 2. Recognition error rate (%) for early or late reverb part only for early reflection period T_e taken to be 18.7 ms and 25 ms

Room reverberation can be divided into two parts : the early and late reverberation. Early reverb refers to the part where room impulse response is still sparse while late reverb is the later denser part. Most of the spectral coloration, which surely affects speech recognition features, comes from early reverberation. The question then arises on how much each part of the reflections affect the recognition performance. To find out, we conducted an experiment where a synthetic room impulse response is generated using an image method [9]. A close-talk speech signal is then filtered by the whole impulse response, the early reflection part and then the late reflection part (including direct signal delay), resulting in three copies of reverberant signals. The dimension of the room used is 4 x 4 x 2.5 m with a reverberation time (T60), for the energy to drop 60 dB, of 0.23 seconds. Table 2 shows the recognition error rate for different durations taken to be an early reflection period. The value of $T_e = 18.7$ ms is the time when the impulse response's region of sparse delays end and dense non-zero values starts. The other value of $T_e = 25$ ms is for comparison with result in [3].

The results supports the claim in [3] that the late echoes have more negative impact on recognition than the early ones. However, it has been found in our experiments, both with synthetic and real reverberation, that letting the filter taking values freely in the "don't care" region as in BWLS leads to severe spectral distortion (mainly lowpassed) despite the reduction in trailing echoes. This leads to inferior recognition results. Instead, by weighting the previously "don't care" region by just a small weight, the estimated filter is constrained to be well-behaved in both regions. The relative weights allow more control in a trade-off between early and late reflection regions. The results applying WLS to a synthetic reverberated speech mentioned earlier, using 1024-tap filter, is given in Table 3. The discrepancy from [3] is most likely because of shorter utterances used here so that the benefit of suppressing more of late reverb over short reverb is not as pronounced.

2.4. Frequency Domain Approach

Here, a frequency domain counterpart of the previous least-squares in section 2.1 is presented. A convolution in time

Weight	0	0.1	1
Error rate	25.74%	17.98%	17.69%

Table 3. Recognition error rate for WLS using weights 0, 0.1 and 1 on the first 25 ms after first reflection

is transformed to a multiplication in frequency. For each DFT frequency bin, we solve for a complex least-squares solution of the filter's transform, $W(k)$, from the following equation

$$\mathbf{X}_{\{m\}}(k) = W(k)\mathbf{Y}_{\{m\}}(k) \quad (2)$$

for $k = 1, ..., N_{FFT}$, where N_{FFT} is the length of DFT and $\mathbf{X}_{\{m\}}(k)$ and $\mathbf{Y}_{\{m\}}(k)$ are vectors of complex valued input and output DFT at bin k respectively, taken from a set of STFT frames $\{m\}$ with significant energy above a threshold δ_E. After all frequency bins have been estimated, the filter $w(t)$ is obtained by $IDFT$. The immediate result will, however, be in a zero-phase form which requires "$FFTSHIFT$" operation to get $w(t)$ ready for filtering in MATLAB. Asymmetric structure can then be imposed by truncation of this non-causal filter. The pre-echo behavior also exists in this approach if long anti-causal part filter is used (including directly multiplying the estimated $W(k)$ and $Y(k)$ before IDFT). The length of FFT used must be long enough to allow for ringing of convolution in time.

There are a number of motivations behind the use of frequency domain approach. One is the computational complexity. From (2), the least squares solution only requires complex vector multiplication, as opposed to doing a matrix inverse like in time-domain LS. However, a direct comparison is not possible since it also depends on the hop-rate and the size of FFT used, among other things. A more important issue is, perhaps, frequency domain estimation allows more room for manipulation on SNR selection. It also allows for an emphasis on what matters to the application e.g. psychoacoustic for listening and feature-based spectral shaping for speech recognition. The room for better manipulation extends to band-wise processing. Since in physical room, high frequency usually decays much faster than the low frequency, we can achieve better performance and efficiency by having shorter filter for high-frequency components and longer for low-frequency ones. In this paper, however, we only compare simple energy-selective solutions as described above, leaving the rest as future work.

3. DISCUSSION

Figure 1 shows error rate of speech recognition for various length combinations of filters, applied to synthetic reverberation. A definitely non-minimum phase (NM) case, assuming no pole-zero cancelation, is constructed by adding three zeros outside a unit circle to the system. From the plot, it is clear that long filter does better in the recognition tests. However, long anti-causal part degrades the performance because of the pre-echo. A short asymmetrical length inverse filter is therefore generally preferred. Comparing the (definitely) non-minimum phase system result with the original synthetic room impulse response, the former suffers more in original far-field simulation. The leftmost points on the plot correspond to using minimum possible anti-causal length (only to compensate for direct signal delay). The severe degradation subject to the non-minimum phase case for this filter shown indicates that non-minimum phase compensation is needed. Though the result may be circumstantial, room response in general has non-minimum phase occurred only in the late reverberation region. Therefore only moderate length of anti-causal part may be enough for general use.

The general trend follows in real room recording results shown in Figure 2. The exception is a significant difference in recognition rate using long 2048-tap filter. This suggests that the discrepancy might come from non-stationarity since results for shorter filters are hardly different. As also shown in Figure 2, the solution derived via frequency domain approach gives comparable performances in all cases for length-2048 filter. In Figure 3, the mean squared error (MSE) between an ideal impulse and the equalized response for the synthetic case shows that lower MSE values do not entail better recognition. While equalization is good, pre-echo has a far more negative impact. A comparison is

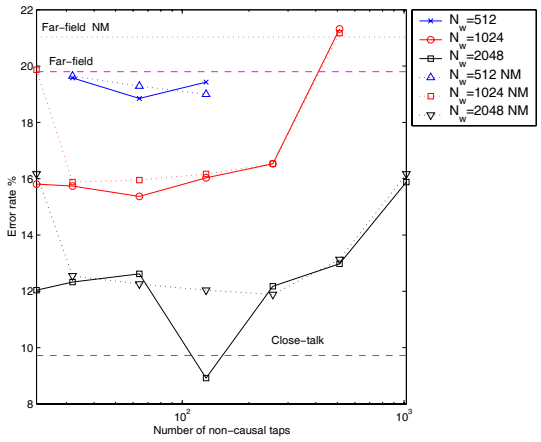

Fig. 1. Recognition error rate (%) from synthetic room reverberation for various filter length (N_w) combinations for (a) synthetic impulse response (b) non-minimum phase (NM) zeros added

given in Table 4 for length-1024 filter with anti-causal tap of 128, time-domain BWLS and an inverse filtering approach called complex-smoothed inverse filtering taken from [1], all calculated from the original synthetic impulse response. This last algorithm avoids deep-notch inverting problem by

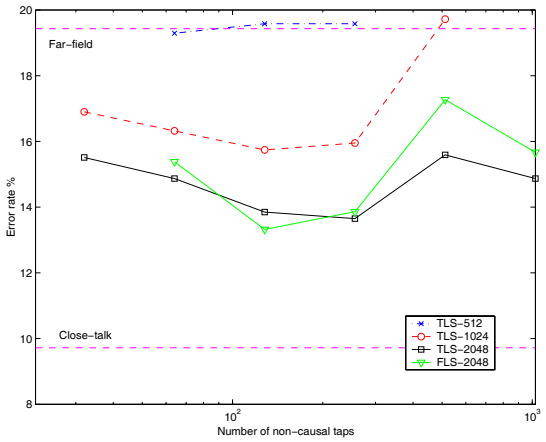

Fig. 2. Recognition error rate (%) from real recording for filter length (N_w) of 512, 1024 and 2048 in time (TLS-513,1024,2048) and frequency domain (FLS-2048).

Method	a-MMSE	BWLS	CplxSm
Error rate	8.92%	21.17%	22.77%

Table 4. Error rate comparison between time-domain MMSE asymmetric inverse filter (a-MMSE), BWLS and the complex-smoothed inverse filtering (CplxSm) on synthetic room impulse response

smoothing in frequency domain. It aims for pleasant listening, avoiding artifacts which commonly come with compensation of non-minimum phase zeros. As the results show, this does not necessarily lead to good speech recognition performance. Though, this may partly be because there is not much problem with non-minimum phase in this test. Also,the impulse response used here may not be as long as the effective cases of large concert hall and auditorium demonstrated in the paper.

4. CONCLUSION

We showed in this paper the comparative effects of reverberation on speech recognition. A few different approaches to arriving at an asymmetric-length non-causal linear FIR filter and others have been presented. The recognition improvement over far-field signal has been shown in both synthetic and real room recordings. The experimental results encourage more consideration on speech recognizer oriented design of dereverberating algorithms.

5. REFERENCES

[1] P.Hatziantoniou and J.Mourjopoulos, "Results for room acoustics equalisation based on smoothed responses," in *114th AES Convention*, March 2003.

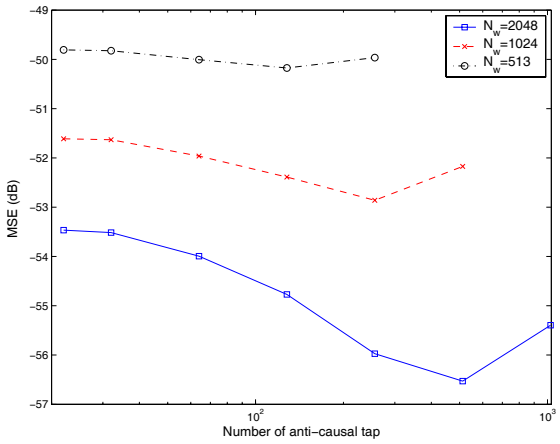

Fig. 3. MSE of equalized response with an ideal impulse for filter length (N_w) of 512, 1024 and 2048 against different number of anti-causal taps, derived from time domain MMSE.

[2] E.Armelloni E.Ugolotti A.Bellini, G.Cibelli and A.Farina, "Car cockpit equalization by warping filters," *IEEE Transactions on Consumer Electronics*, vol. 47, no. 1, pp. 108–116, February 2001.

[3] B.W.Gillespie and L.E.Atlas, "Acoustic diversity for improved speech recognition in reverberant environments," in *ICASSP*, 2002.

[4] J.Droppo L.Deng and A.Acero, "Enhancement of log mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise," *IEEE Trans. on Speech and Audio Processing*, vol. 12, no. 2, pp. 133–143, March 2004.

[5] Y.Pan and A.Waibal, "The effects of room acoustics on mfcc speech parameter," in *ICSLP*, 2000.

[6] M.S.Bradstein, "An event-based method for microphone array speech enhancement," in *ICASSP*, 1999, pp. 953–956.

[7] J.Flanagan D.Rabinkin, R.Renomeron and D.F.Macomber, "Optimal truncation time for matched filter array processing," in *ICASSP*, 1998, vol. 6, pp. 3629–3632.

[8] S.M.Kay, *Fundamentals of statistical signal processing*, Prentice Hall, 1993.

[9] D.A.Berkely J.B.Allen, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, pp. 943–950, April 1979.