An Assessment of the Improvement Potential of Time-Frequency Masking for Speech Dereverberation

Chenxi Zheng[1], *Tiago H. Falk*[2], *Wai-Yip Chan*[1]

[1]Department of Electrical Engineering, Queen's University, Kingston, Canada
[2]Institut National de la Recherche Scientifique (EMT), Montreal, Canada

Abstract

The effect of ideal time-frequency masking (ITFM) on the intelligibility of reverberated speech is tested using objective measurement, namely STI and PESQ scores. The best choice of ITFM threshold is determined for a range of reverberation times (RTs). Four existing dereverberation algorithms are also assessed. Objective test results and informal subjective listening show that IFTM provides great intelligibility improvement for all RTs and outperforms the existing dereverberation algorithms, one of which assumes perfect knowledge of the room impulse response. While ITFM provides only a best possible performance bound, our results demonstrate the potential improvement that could be obtained using time-frequency masking for speech dereverberation.

Index Terms: Dereverberation, speech intelligibility, speech transmission index, speech quality, time-frequency masking

1. Introduction

In computational auditory scene analysis research, ideal time frequency masking (ITFM) is reported to improve noisy speech intelligibility [1]. More recently, subjective tests were used to characterize the effects of ITFM (and the masking threshold) on noise-corrupted speech intelligibility and provided insights on how to better build noise suppression algorithms [2]. In a similar vein, time-frequency masking (TFM) was shown to also improve reverberated speech intelligibility [6]; the environments studied, however, only encompassed those with short reverberation time (RT) values, i.e., below 400 ms.

Existing single-channel dereverberation algorithms (e.g., [3, 4, 5]) are known to only subtly improve speech intelligibility. Motivated by the findings reported in [2, 6], we explore the benefits obtained with binary masking for reverberated speech across a wider range of RT values, encompassing both smaller (e.g., offices) and larger (e.g., theaters) enclosures. In order to garner the potential of using TFM for quality and intelligibility improvement, we use ITFM. While ITFM is not practical - it requires the knowledge of the original clean speech signal to compute the binary mask - it offers a benchmark of the best possible attainable performance.

In this paper, a series of experiments are performed in order to systematically analyze the ability of ITFM to improve both the quality and intelligibility of reverberated speech. The effect of the masking threshold and its relationship with RT are also studied. The remainder of this paper is organized as follows. Section 2 describes the reverberated speech model and ITFM processing scheme. In Section 3, after describing two databases and four benchmark dereverberation algorithms, four experiments are conducted to assess not only the potential of ITFM in intelligibility improvement, but also elements affecting ITFM performance. Conclusions are drawn in Section 4.

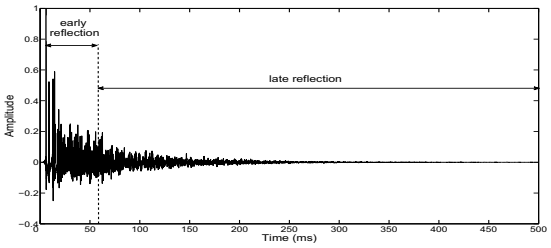

Figure 1: Waveform of a representative room impulse response

2. Speech and masking model

In this section, a brief description of the reverberated speech and ITFM models are given.

2.1. Speech model

In a reverberant room, the reverberated speech $z(n)$ results from the convolution of the clean speech signal $s(n)$ and the room impulse response (RIR) $h(n)$ as

$$z(n) = \sum_{i=0}^{Q-1} h(i)s(n-i), \quad (1)$$

where Q is the length of $h(n)$. Fig. 1 depicts a representative RIR generated by the so-called image method [7].

The RIR can be partitioned into three components: the direct signal, early reflections, and late reflections. The direct signal is the strongest impulse corresponding to the direct path from the speech source to the listener. Early reflections are the impulses that arrive within 50 ms after the direct signal. Early reflections are known to cause short-term reverberations or "coloration" effects. Early reflections can boost signal energy as well as emphasize modulation frequency content around 4 Hz [8], thus have minimal effects on intelligibility. Late reflections, in turn, which arrive at time intervals greater than 50 ms post the direct impulse, smear the speech spectrum and can severely reduce signal quality and intelligibility. Late reflections cause the so-called long-term reverberations or echoes.

Since we are interested in improving quality and intelligibility, we decompose the reverberated speech signal $z(n)$ into two components. The first component, namely $z_e(n)$, includes the direct path signal $s(n)$ and the early reflections. The second, $z_l(n)$, includes the late reflections. The decomposition is

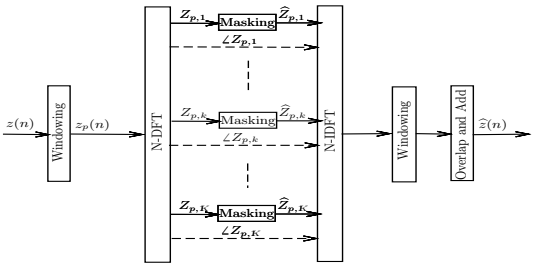

Figure 2: ITFM processing steps

given by:

$$z(n) = z_e(n) + z_l(n) = \sum_{i=0}^{Q_e-1} s(n-i)h(i) + \sum_{i=Q_e}^{Q-1} s(n-i)h(i), \quad (2)$$

where Q_e represents the length of the early reflections.

As mentioned above, reverberated speech quality and intelligibility are mostly compromised due to the late reflections. Commonly, late reflections are modeled as an exponentially damped Gaussian white noise process. This assumption is key to the use of spectral subtraction techniques, originally designed for additive noise suppression, for dereverberation (e.g., [3, 4]); in summary, late reverberations are treated as additive noise. Since ITFM has been shown to improve intelligibility for noise-corrupted speech, we explore its efficacy for dereverberating speech. ITFM is described next.

2.2. Ideal time frequency masking

With ITFM, access to both the clean speech signal $s(n)$ and the reverberated speech signal $z(n)$ is required. The processing steps applied to $z(n)$ are shown in Fig. 2. First, $z(n)$ is windowed by an analysis window. An N-point DFT is then taken and the magnitude spectrum $Z_{p,k} = |z_{p,k}|$ ($p = 1...P, k = 1...K$) is input to the masking processing module; here p indexes the windowed speech frame and $k = \frac{N}{2}$ the DFT coefficients. The modified magnitude spectrum $\widehat{Z}_{p,k} = |\widehat{z}_{p,k}|$ and unmodified phase spectrum of the reverberated speech $\angle z_{p,k}$ are then input to an N-point IDFT and further windowed by a synthesis window. Overlap-and-add is used to reconstruct the enhanced signal $\widehat{z}(n)$. In our experiments, a square-root Hann window of length 20 ms is used both as the analysis and synthesis windows; 50% frame overlap is used.

For the masking processing module, the output $\widehat{Z}_{p,k}$ is the product of $Z_{p,k}$ and a binary mask $I_{p,k}$. The mask I is obtained by comparing the spectral magnitude of the clean speech signal $S_{p,k}$ and reverberated speech signal $Z_{p,k}$. The following rules are used to obtain the binary mask:

$$I_{p,k} = \begin{cases} 1 & Z_{p,k} < \theta S_{p,k}; \\ 0 & Z_{p,k} \geq \theta S_{p,k}, \end{cases} \quad (3)$$

where θ is a masking threshold parameter which controls how severely spectral components are suppressed; $\theta = \sqrt{2}$ is commonly used for noise suppression.

3. Experiments

In this section, the intelligibility of reverberated and processed speech signals are objectively assessed using the speech transmission index (STI) and the International Telecommunications Union ITU-T PESQ speech quality measurement algorithm [15]; PESQ scores have been shown to correlate with intelligibility ratings [11]. Here, two speech based derivatives of STI are explored, namely STI1 [9] and STI2 [10]. In the experiments described below, STI and PESQ scores are averaged over the entire data sets. Time-alignment was applied to compensate for direct-path delays in the reverberated speech prior to ITFM and STI computation; PESQ is already equipped with an internal time-alignment algorithm.

3.1. Databases

Two databases are used in our experiments. The first consists of 128 clean speech files, spoken by two male and two female subjects, artificially corrupted using the SImulation of REal ACoustics (SIREAC) tool [12], with RT values ranging from 0.1-2 s. The second database consists of a corrupted version of the Wall Street Journal November 92 speech testset (330 sentences uttered by eight different speakers). The clean speech files are corrupted by a recorded six-channel room impulse response measured by a linear microphone array in four different enclosures with reverberation times of 274, 319, 422 and 533 ms [13]. Both databases are originally sampled at 16 kHz but were downsampled to 8 kHz due to restrictions in the PESQ algorithm. Both the clean and reverberated speech files were level-normalized to -26 dBov using the P.56 voltmeter [14].

3.2. Benchmark dereverberation algorithms

In order to gauge the benefits of using ITFM for dereverberation, four multi-channel benchmark algorithms are used, namely, delay-and-sum beamforming (DSB), cepstral liftering, subspace-based dereverberation, and matched inverse filtering. The latter assumes the availability of the RIR, and like ITFM is impractical. The reader is referred to [13] for more details about these multichannel dereverberation algorithms.

3.3. Assessing intelligibility improvements

In this experiment, we gauge the benefit of ITFM for intelligibility improvement by comparing STI and PESQ improvements over the four multi-channel benchmark algorithms; the second multi-channel database is used for this purpose. PESQ and STI scores of both reverberated and dereverberated speech signals are shown in Fig. 3 and 4, respectively. Since ITFM is inherently a single-channel method, the performances shown in the figures for ITFM are for reverberated speech obtained by convolving the clean signals with the RIR from one of the microphones from the array.

As can be seen from the figures, ITFM achieves the best quality and intelligibility, followed by matched inverse filtering (represented as "mat" in the plots). ITFM is also shown to outperform the remaining multi-channel dereverberation algorithms, by as much as 1 point on the 5-point mean opinion score (MOS) PESQ scale and by 0.125 on the [0,1] STI scale (at RT $=$ 533 ms). All the algorithms provide improvement in STI and PESQ scores, with the exception of cepstral liftering whose STI1 scores are below reverberated speech. Informal listening tests agree with the rank order of the dereverberation schemes in Fig. 3. Residual reverberation is audible in the processed speech of all the dereverberation schemes except ITFM. ITFM-processed speech contains audible distortions but does not sound noticeably reverberated. Matched inverse filtered speech sounds less reverberated than the other three benchmark schemes but it also contains distortions.

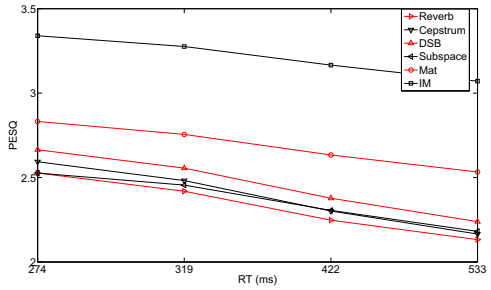

Figure 3: Gauging quality improvements using PESQ

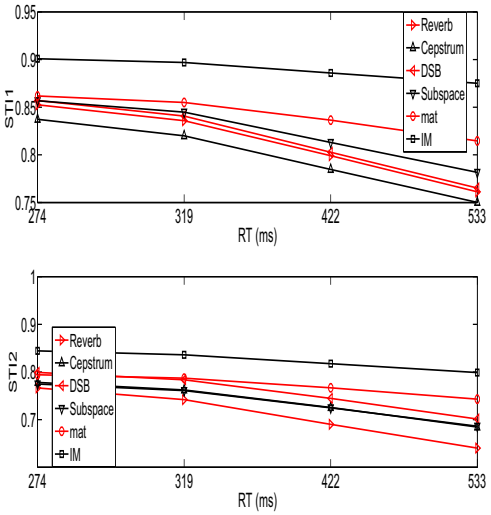

Figure 4: Gauging intelligibility improvements using STI

3.4. Assessing reverberation time effects

In this experiment, we assess the potential of ITFM for intelligibility improvement for a wide range of reverberation time values; for this purpose, the first database is used. Fig. 5 depicts STI1 and STI2 behaviour relative to increasing RT for both the reverberant and ITFM-processed signals. Since this is a single-channel dataset, the multi-channel benchmark algorithms are not used. As will be shown in Section 3.6, the optimal threshold parameter needs to be tuned for different RTs. In this experiment, the optimal threshold parameters in Fig. 9 are used.

As can be seen from Fig. 5, both STI1 and STI2 drop quickly for reverberated speech with increasing RT. The very low values (STI $\sim 0.3-0.4$) obtained for RT $= 2$ s suggest that intelligibility is severely compromised; informal listening tests corroborate such findings. For ITFM-processed speech, on the

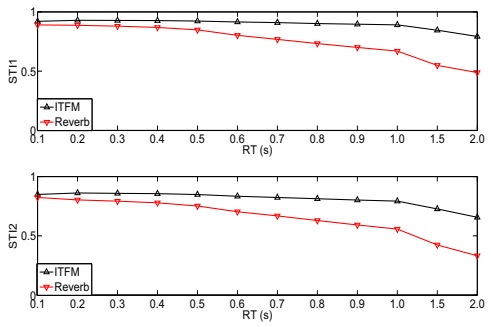

Figure 5: STI of reverberated and ITFM-processed speech for increasing RT

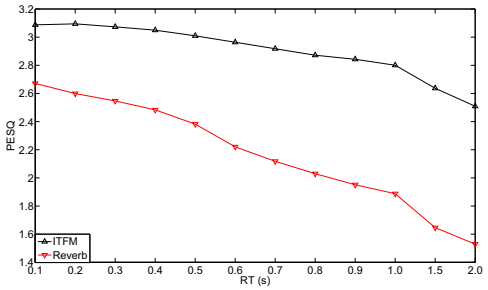

Figure 6: PESQ scores of reverberated and ITFM-processed speech for increasing RT

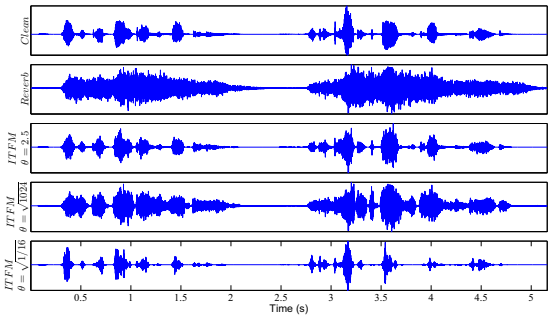

Figure 7: Top-to-bottom: waveform of clean, reverberated (RT $= 2$ s), and ITFM-processed speech for different θs

other hand, STI values decay slowly and values around $0.6-0.8$ are observed (i.e., acceptable intelligibility) at high RT values. In fact, STI values for ITFM-processed speech at RT $= 2$ s are similar to those observed for reverberated speech at RT $= 0.6$ s. Additionally, Fig. 6 plots PESQ scores attained for reverberated and ITFM-processed speech for increasing RT. Similarly, quality drops with increasing RT are slower for ITFM-processed speech relative to reverberated speech. Quality scores obtained for RT $= 2$ s for ITFM-processed speech correspond to those obtained with reverberated speech at around RT $= 0.4$ s. Informal subjective tests corroborate the quality gains obtained with ITFM processing.

In order to visually assess the gains obtained with ITFM, Fig. 7 illustrates, from top-to-bottom, the clean (uttered by a female), reverberated (RT $= 2$ s), and ITFM-processed (at different threshold values) speech waveforms. As can be seen from the ITFM-processed waveform with a threshold of $\theta = 2.5$, the majority of the clean speech envelope is restored, suggesting improved intelligibility.

3.5. Assessing masking threshold effects

In this experiment, we will assess the effect of the ITFM threshold parameter θ on intelligibility. Since STI measurements are sensitive to severe non-linear distortions observed when the threshold is small, only PESQ is used in this experiment to gauge intelligibility/quality improvements. Fig. 8 depicts the average PESQ score as a function of θ and RT. For the RT $= 2$ s curve, the thresholds between 4 and $\sqrt{2}$ attain relatively good performance. When the threshold becomes extremely large, almost all spectral components are kept and quality approaches that of the unprocessed reverberated speech signal. On the other hand, PESQ score drops quickly when the threshold becomes extremely small, i.e., when only a few spectral components are kept. This behavior can be observed from the ITFM-processed

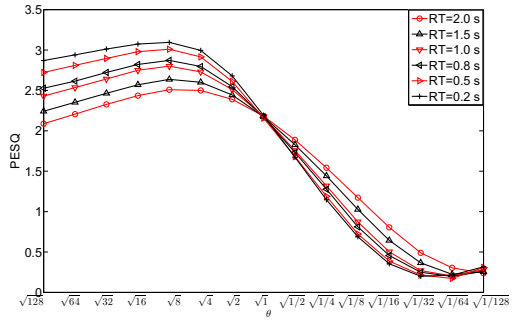

Figure 8: PESQ score as a function of RT and θ for ITFM-processed speech

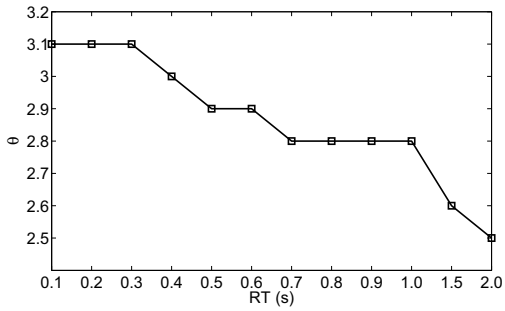

Figure 9: Best θ as a function of RT

speech waveforms depicted by Fig. 7.

3.6. Assessing the relationship between RT and θ

The optimal masking threshold depends on RT as it drives how severely spectral components are suppressed. In [16], it is suggested that the quality of reverberated speech is determined by two independent variables: RT and the RIR spectral variance. In this experiment, we study the effect of RT on the selection of threshold parameter θ. Again, PESQ is used as the quality/intelligibility criterion.

As can be seen in Fig. 8, for smaller RT, relatively larger threshold values attain best intelligibility. This is because the smaller is RT, the higher the speech to reverberation ratio. Speech time-frequency components are less corrupted by reverberation and should be more likely to be kept by using a lager θ in ITFM. As θ decrease below one, more components are suppressed; the larger RT speech benefits more so that the ITFM processed speech has better PESQ score for lager RTs. Nevertheless, the optimal threshold value to use is greater than one for all RTs, and is plotted in Fig. 9. For RT $\geq T_0 = 2$ s, $\theta = 2.5$ is recommended, but larger θ is recommended for RT $< T_0$. The slopes of the curves in Fig. 8 suggest that it is better to err on the side of using a larger than optimal θ (i.e. lesser suppression) than smaller. As the optimum threshold depends on RT, the blind RT estimator in [17] can be used to adjust θ. Threshold parameter θ adaptive to blind RT estimation and the dependence of θ on RIR spectral variance will be studied in the future.

4. Conclusion

In this paper, ideal time-frequency masking (ITFM) is used to gauge the potential benefits of using binary masks for reverberated speech intelligibility improvement. Two intelligibility-related measures, namely the speech transmission index and ITU-T PESQ scores, are used to assess the effects of reverberation time, masking threshold parameter, and their interrelationship on ITFM performance. The objective measurements, combined with informal listening tests, show that significant quality and intelligibility improvements are obtained with ITFM processing. Experiments with four multi-channel dereverberation algorithms showed that ITFM can furnish substantial gains in both quality and intelligibility, thus suggesting that time-frequency binary masking is a promising method for speech dereverberation.

5. References

[1] D. Brungart, P. Chang, B. Simpson, and D. L. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time frequency segregation," *J. Acoust. Soc. Amer.*, vol. 120, pp. 4007-4018, 2006.

[2] N. Li, and P. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *J. Acoust. Soc. Amer.*, vol. 123, no. 3, pp. 1673-1682, 2008.

[3] J. M. Boucher K. Lebart. "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica*, vol. 87, pp. 359-366, 2001.

[4] H. W. Lollmann and P. Vary. "A blind speech enhancement algorithm for the suppression of late reverberation and noise," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 3989-3992, 2009.

[5] C. Avendano and H. Hermansky, "Study on the dereverberation of speech based on temporal envelope filtering," in *Proc. Fourth Int. Conf. Spoken Language*, vol. 2, pp. 889-892, 1996.

[6] N. Roman, and D. Wang "Pitch-based monaural segregation of reverberant speech," *J. Acoust. Soc. Amer.*, vol. 120, pp. 458-469, 1996.

[7] B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J Acoust. Soc. Amer.*, vol. 65, pp. 943-950, Apr 1978.

[8] T. H. Falk, C. Zheng and W.-Y. Chan, "A Non-Intrusive Quality and Intelligibility Measure of Reverberant and Dereverberated Speech," *IEEE Trans. on Audio, Speech, Lang. Process.*, pp. 1766-1774, 2010.

[9] K. Payton and L. Braida "A method to determine the speech transmission index from speech waveforms," *J. Acoust. Soc. Amer.*, vol. 106, pp. 3637-3648, 1999

[10] R. Drullman, J. Festen and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," *J. Acoust. Soc. Amer.*, vol. 95, no. 5, pp. 2670-2680, May 1994.

[11] J. Ma, Y. Hu, and P. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *J. Acoust. Soc. Amer.*, Vol. 125, no. 5, pp. 3387-3405, 2009

[12] H. Hirsch and H. Finster, "The simulation of realistic acoustic input scenarios for speech recognition systems," in *Proc. Interspeech*, pp. 2697-2700, 2005.

[13] K. Eneman and M. Moonen, "Multimicrophone speech dereverberation: Experimental validation," *EURASIP J. Audio, Speech, Music Process.*, 2007, 19 pages

[14] ITU-T P.56, "Ojective Measurement of Active Speech Level," Int. Telecom. Union, 1993.

[15] ITU-T P.862, "Perceptual evaluation of speech quality: An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Int. Telecom. Union, 2001.

[16] J. B. Allen, "Effects of small room reverberation on subjective preference," *J. Acoustic. Soc. Amer.*, vol. 71, Apr. 1982.

[17] T. H. Falk and W.-Y. Chan, "Temporal Dynamics for Blind Measurement of Room Acoustical Parameters," *IEEE Trans. Instrum. Meas.*, Vol. 59, No. 4, pp. 978-989, April 2010.