THE LINEAR PREDICTION INVERSE MODULATION TRANSFER FUNCTION (LP-IMTF) FILTER FOR SPECTRAL ENHANCEMENT, WITH APPLICATIONS TO SPEAKER RECOGNITION

Bengt J. Borgström and Alan McCree

MIT Lincoln Laboratory
Lexington, MA 02420
{jonas.borgstrom,mccree}@ll.mit.edu

ABSTRACT

We propose a method for spectral enhancement of reverberant speech based on inverting the modulation transfer function (MTF). Using all-pole models of modulation spectra allows the linear prediction inverse MTF (LP-IMTF) filter to exhibit a smooth frequency response, and allows it to be implemented as a low-order IIR filter in the modulation envelope domain. The proposed filter adapts to current acoustic conditions without relying on explicit information regarding reverberation time.

Additionally, the LP-IMTF framework allows for estimation of useful side information, such as local signal-to-reverberation ratios and band-specific reverberation times. As example applications, the LP-IMTF system is applied to enhancement and speaker recognition of reverberant speech, and significant performance improvements are achieved.

1. INTRODUCTION

When observed in an enclosed environment, speech signals will generally experience distortion due to reverberation, which is caused by multi-path propagation of sound from source to sensor. Human intelligibility has been widely shown to degrade in the presence of reverberation [1], as has the performance of automated speech systems such as automatic speech recognition (ASR) and speaker recognition [2]. It is therefore of interest to enhance spectra of reverberant speech.

The concept of the modulation transfer function (MTF) is introduced by Houtgast and Steeneken in [1] to characterize the acoustic channel encountered when observing speech within an enclosed space. Specifically, they explore the effect of reverberation on the modulation index of the intensity envelope for an input signal, and the resulting effect on speech intelligibility.

In [3], Langhans and Strube aim to suppress acoustic distortion by inverting the magnitude of the MTF in order to reshape the modulation spectrum of degraded speech. The inverse modulation transfer function (IMTF) filter has since been explored as a means by which to suppress the effects of adverse acoustic environments on speech signals, thereby improving perceptual quality of resynthesized speech [4],[5].

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through Air Force Contract FA8721-05-C-0002. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, or the U.S. Goverment.

In this paper, we propose a method for spectral enhancement of reverberant speech. We discuss the modulation transfer function, and its behavior for speech with convolutional distortion. We utilize all-pole modeling of modulation spectra of clean and degraded speech to derive the LP-IMTF filter, which adapts to current acoustic conditions, and implement it as an IIR filter in the modulation envelope domain. The proposed spectral enhancement method is applicable to a variety of applications. In this study, it is applied to enhancement and speaker recognition of reverberant speech. Aside from spectral enhancement, the proposed framework provides useful side information, namely frame-level signal-to-reverberation ratios (SRRs), and frequency band-specific reverberation times.

This paper is organized as follows. Sec. 2 discusses the modulation transfer function and its behavior for speech with convolutional distortion, and derives the proposed LP-IMTF filter. Sec. 3 discusses extraction of side information. Experimental results for enhancement and speaker recognition are included in Sec. 4, and Sec. 5 provides conclusions.

2. THE LINEAR PREDICTION INVERSE MODULATION TRANSFER FUNCTION (LP-IMTF) FILTER

2.1. The Modulation Envelope Domain

A discrete speech signal observed in a reverberant environment can be expressed as

$$y(n) = \sum_{l=0}^{\infty} h(l)\, x(n-l) \tag{1}$$

where $x(n)$ is the underlying clean speech and $h(n)$ is the causal room impulse response. Short-time spectral analysis of $y(n)$ reveals channel-specific trajectories of spectral magnitudes along time, i.e. modulation envelopes. When applying short-time spectral analysis, the relationship from (1) becomes difficult to express mathematically, and instead short-time spectra are approximated as [3],[6]

$$|Y_k(m)| = \sum_{l=0}^{\infty} |H_k(l)|\, |X_k(m-l)| \tag{2}$$

where $X_k(m)$ and $Y_k(m)$ denote the short-time Fourier transforms (STFTs) of $x(n)$ and $y(n)$, respectively. $H_k(m)$ characterizes the inter-frame effect of reverberation, and k and m refer to the channel and time index, respectively. From (2), the effect of reverberation along short-time spectral envelopes is modeled as a channel-wise convolution. To capture the "smeared" nature typically observed in spectrograms of reverberant speech, $|H_k(m)|$ is generally defined as a causal low-pass envelope. The decay rate of $|H_k(m)|$ is then related to reverberation time, which is commonly measured as t_{60}, i.e. the time required for $h(n)$ to attenuate by 60 dB.

2.2. The Modulation Spectral Domain

In the case of mild reverberation, when the room impulse response is short in duration relative to the short-time analysis window, (2) can be reduced to $|Y_k(m)| \approx |H_k(0)||X_k(m)|$, which has been used to motivate frame-based compensation techniques such as cepstral mean and variance normalization (CMVN) [7]. However, for reverberation which is more severe, distortion in (2) is a function of past short-time spectra, and frame-based algorithms may not be effective. To compensate for such effects, we look to leverage the inter-frame relationships of speech via the modulation spectrum and obtain an enhanced short-time spectrum, $|\hat{X}_k(m)|$.

The modulation spectrum is the frequency decomposition of an energy envelope extracted from a subband signal [8]. In this study, we define the modulation spectrum as

$$M_{Y,k}(\omega) = \sum_{m=-\infty}^{\infty} |Y_k(m)| \exp(-j\omega m) \tag{3}$$

with analogous terms defined for $X_k(m)$ and $H_k(m)$. Using (2) and (3), the modulation spectrum of $Y_k(m)$ becomes

$$\begin{aligned} M_{Y,k}(\omega) &= \sum_{m=-\infty}^{\infty} \sum_{l=0}^{\infty} |H_k(l)||X_k(m-l)| \exp(-j\omega m) \\ &= \sum_{l=0}^{\infty} |H_k(l)| \exp(-j\omega l) \\ &\quad \times \sum_{m=-\infty}^{\infty} |X_k(m-l)| \exp(-j\omega(m-l)) \\ &= M_{H,k}(\omega) M_{X,k}(\omega) \end{aligned} \tag{4}$$

revealing reverberation to induce a multiplicative distortion in the modulation spectral domain.

2.3. The LP-IMTF Filter

As proposed by Langhans and Strube in [3], the modulation spectrum of a degraded signal can be reshaped by inverse filtering the MTF. We aim to design an IMTF filter, $F_k(\omega)$, whose magnitude frequency response is given by

$$|F_k(\omega)| = |M_{H,k}(\omega)|^{-1} = \left| \frac{M_{X,k}(\omega)}{M_{Y,k}(\omega)} \right| \tag{5}$$

Here, knowledge regarding $|M_{Y,k}(\omega)|$ can be extracted from the observed speech signal, whereas the underlying $|M_{X,k}(\omega)|$ is unknown and must be learned from training data. We propose to use all-pole models of these modulation spectra during implementation of the IMTF filter. The motivation for this is three-fold:

- All-pole modeling provides smooth spectral transitions within modulation spectra, avoiding rapid fluctuations generally encountered when using large DFTs. This is especially important when determining the ratio of modulation spectra, as in (5), since small values in the denominator can yield large fluctuations in the resulting IMTF filter.
- All-pole modeling allows for modulation behavior to be summarized by a small set of linear prediction coefficients. $|M_{X,k}(\omega)|$ can then be efficiently trained as a small number of parameters.

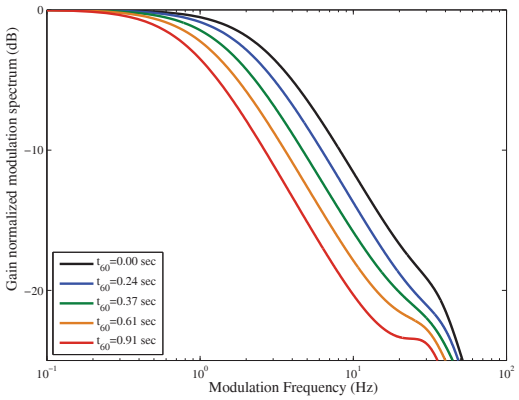

Fig. 1. Gain-normalized all-pole modulation spectra of example speech in the presence of reverberation of varying degree, for the frequency channel with center frequency 1500 Hz, and for P=6

- All-pole modeling allows for efficient implementation of the IMTF filter in the modulation envelope domain as a low-order IIR filter, as will be shown in (8)-(9). This avoids explicit transformation into the modulation spectral domain.

The all-pole modulation spectrum of degraded speech is determined by analyzing the normalized modulation envelope autocorrelation function $r_{Y,k}(\tau)$, defined as

$$r_{Y,k}(\tau) = \frac{E\{|Y_k(m)||Y_k(m+\tau)|\}}{E\{|Y_k(m)|^2\}} \tag{6}$$

which is estimated from the short-time spectra of the observed speech signal. Normalized autocorrelation coefficients are used in (6) since long-term average channel gains can contain speaker-specific information important for tasks such as speaker recognition, and should therefore not affect the IMTF filter shape. From $r_{Y,k}(\tau)$, the linear prediction coefficients $a_{Y,k}(l)$ and gain $\sigma_{Y,k}$ are extracted, yielding the all-pole model

$$|M_{Y,k}(\omega)|^2 \approx \frac{\sigma_{Y,k}^2}{\left|1 - \sum_{l=1}^{P} a_{Y,k}(l) \exp(-j\omega l)\right|^2} \tag{7}$$

where P is the prediction order. Analogous terms ($r_{X,k}(\tau)$, $a_{X,k}(l)$, and $\sigma_{X,k}$) are defined for the clean modulation spectrum, and determined similarly, although $r_{X,k}(\tau)$ is learned from training data.

As discussed in Sec. 2.2, the presence of reverberation can be expected to affect the shape of $|M_{Y,k}(\omega)|$. Fig. 1 provides gain-normalized all-pole modulation spectra of example speech in the presence of reverberation of varying degree. In this example, reverberation is added artificially to microphone interview speech from the 2010 NIST-SRE, using a room impulse response generator based on [9]. It can be observed in Fig. 1 that as the acoustic severity increases, modulation spectra become increasingly low-pass.

Applying all-pole modulation spectra to (5) results in the proposed LP-IMTF filter

$$|F_k(\omega)| = \left| \frac{\sigma_{X,k}\left(1 - \sum_{l=1}^{P} a_{Y,k}(l) \exp(-j\omega l)\right)}{\sigma_{Y,k}\left(1 - \sum_{l=1}^{P} a_{X,k}(l) \exp(-j\omega l)\right)} \right| \tag{8}$$

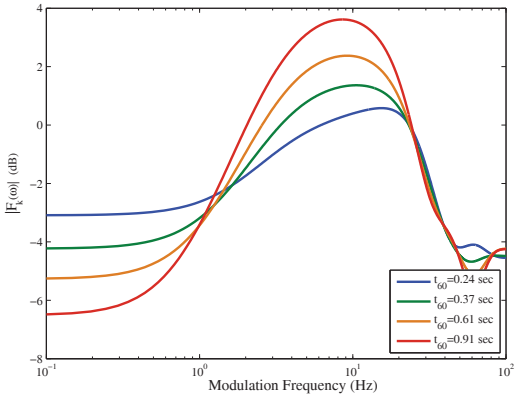

Fig. 2. Proposed IMTF filters for example speech in the presence of reverberation of varying degree, for the frequency channel with center frequnecy 1500 Hz, and for P=6

Fig. 2 illustrates the magnitude frequency response of the LP-IMTF filter obtained for example speech in reverberation of varying degree. It can be observed that the LP-IMTF solution is a bandpass filter in the modulation spectrum. Further, as the acoustic severity increases, the LP-IMTF filter exhibits increasing filter depth.

Since (5) does not account for phase, there exist multiple solutions $F_k(\omega)$ which adhere to this constraint. One such solution can be efficiently implemented by applying the inverse DTFT to the expression within the magnitude operator of (8), yielding a low-order IIR filter in the modulation envelope domain. Further, this solution is guaranteed to be minimum phase, and can therefore be expected to match the causal nature of reverberation in the short-time spectral domain. The solution is given by

$$\left|\hat{X}_k(m)\right| = \frac{\sigma_{X,k}}{\sigma_{Y,k}}\left(|Y_k(m)| - \sum_{l=1}^{P} a_{Y,k}(l)\,|Y_k(m-l)|\right) + \sum_{l=1}^{P} a_{X,k}(l)\left|\hat{X}_k(m-l)\right| \quad (9)$$

Each frequency band of the observed short-time spectra is filtered with (9) to obtain enhanced spectral components. To guarantee non-negativity, processed spectral values must be floored. Additionally, gain smoothing is applied along time and/or frequency axes to reduce nonlinear effects due to flooring.

3. EXTRACTION OF SIDE INFORMATION

Aside from providing enhanced spectra, the proposed LP-IMTF framework allows for extraction of useful side information.

3.1. Local Speech-to-Reverberation Ratio

After enhancement, some short-time spectra can still be deemed unreliable due to low signal-to-reverberation ratio (SRR). For tasks such as speaker recognition, such frames can be expected to offer little discriminative power, instead potentially introducing confusability. Therefore, unreliable frames can be dropped prior to recognition. The LP-IMTF framework allows the frame-based a posteriori SRR to be estimated using spectral subtraction

$$SRR(m) \approx \frac{\sum_k |Y_k(m)|^2}{\sum_k \max\left(|Y_k(m)|^2 - |\hat{X}_k(m)|^2, 0\right)} \quad (10)$$

A hard threshold is set, and frame-dropping (FD) can be applied accordingly.

3.2. Blind Estimation of Reverberation Time

For many applications, it may be of interest to infer the acoustic severity of an observed reverberant speech signal. We propose a method for blind estimation of reverberation time based on the LP-IMTF filtering framework. In Fig. 2 it can be observed that the LP-IMTF filter shape is related to reverberation time. Specifically, the frequency response and spectral slope of the filter seem well-correlated with t_{60} at certain modulation frequencies, eg. $\omega \approx 1$ Hz. The spectral slope of the LP-IMTF filter is given by

$$\frac{\partial |F_k(\omega)|^2}{\partial \omega} = \frac{2}{C_{X,k}^2(\omega)} \times \left[C_{X,k}(\omega)\left(B_{Y,k}(\omega) D_{Y,k}(\omega) - A_{Y,k}(\omega) E_{Y,k}(\omega)\right) - C_{Y,k}(\omega)\left(B_{X,k}(\omega) D_{X,k}(\omega) - A_{X,k}(\omega) E_{X,k}(\omega)\right)\right] \quad (11)$$

where

$$A_{Y,k}(\omega) = 1 - \sum_{l=1}^{P} a_{Y,k}(l)\cos(\omega l) \quad (12)$$

$$B_{Y,k}(\omega) = \sum_{l=1}^{P} a_{Y,k}(l)\sin(\omega l)$$

$$C_{Y,k}(\omega) = A_{X,k}^2(\omega) + B_{X,k}^2(\omega)$$

$$D_{Y,k}(\omega) = \sum_{l=1}^{P} a_{Y,k}(l)\, l \cos(\omega l)$$

$$E_{Y,k}(\omega) = \sum_{l=1}^{P} a_{Y,k}(l)\, l \sin(\omega l)$$

with analogous terms defined for $|X_k(m)|$. We propose to estimate reverberation time as a linear combination of samples of the LP-IMTF frequency response and spectral slope

$$\hat{t}_{60} = \alpha + \sum_{i=1}^{|\Omega|} \beta_i |F_k(\omega)|^2\Big|_{\omega=\omega_i} + \sum_{i=1}^{|\Omega|} \lambda_i \frac{\partial |F_k(\omega)|^2}{\partial \omega}\Big|_{\omega=\omega_i} \quad (13)$$

for some set of frequencies Ω. Here, α, β_i's, and λ_i's can be trained on reverberant development data using linear regression. In this study, using Ω=$\{1, 2, 5, 10, 20\}$ Hz showed promising results. Fig. 3 illustrates results for estimation of reverberation time obtained on reverberant speech using the discussed room impulse responses.

4. EXPERIMENTAL RESULTS

To assess the effectiveness of the LP-IMTF filtering framework, it was applied to the task of speech enhancement of reverberant speech. Enhanced spectra were combined with noisy phase, and speech was synthesized via the overlap-and-add method. Table 1 provides results for enhancement of a three minute segment of reverberant speech created by concatenating TIMIT utterances and

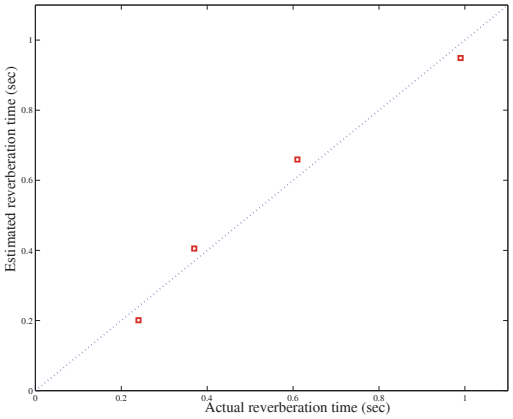

Fig. 3. Blind estimation of reverberation time: red points denote estimates, and blue denotes the diagonal.

Table 1. Speech enhancement results on reverberant speech using the proposed LP-IMTF filtering framework

Algorithm	t_{60} (seconds) 0.24	0.37	0.61	0.99
log-spectral distortion				
Baseline	1.56	2.74	4.15	6.92
LP-IMTF	1.18	1.67	2.33	3.91
PESQ				
Baseline	2.586	2.225	1.986	1.732
LP-IMTF	2.686	2.388	2.162	1.879

applying reverberation. Log-spectral distortion and perceptual evaluation of speech quality (PESQ) [10] are used as metrics. Table 1 shows the LP-IMTF filter to provide significant improvements in speech quality.

Additionally, the LP-IMTF filter was applied as front-end processing to the MIT Lincoln Laboratory Joint Factor Analysis (JFA) speaker recognition system (see [11] for details). Experiments were conducted on the short interview microphone data from the 2010 NIST-SRE corpus, which includes both male and female speakers, with 6.2 K targets and 1.7 M non-targets. Calibration was performed using development data from the 2008 NIST-SRE corpus. Reverberation was artificially added to test cuts for a range of t_{60}'s. (Note that reverberation was added neither during enrollment, nor to development data.) Table 2 provides results for speaker recognition of reverberant speech using the LP-IMTF filter with P=6. Results are reported in terms of equal error rate (EER) and the log-likelihood ratio cost (C_{llr}) from [12]. It can be observed that LP-IMTF filtering improves robustness of speaker recognition to reverberation across the test conditions used, and frame-dropping further improves performance in the more severe conditions. Future work includes using estimated reverberation times to train acoustically matched speaker recognition systems.

5. CONCLUSIONS

This paper has explored spectral enhancement of reverberant speech based on inversion of the modulation transfer function. Using all-pole models of modulation spectra, an efficient LP-IMTF filter is derived which adapts to acoustic conditions. The LP-IMTF framework allows for extraction of side information: local SRR and reverberation time. When applied to enhancement and speaker recognition of reverberant speech, the LP-IMTF filter achieves significant improvements in performance.

Table 2. Speaker recognition results on reverberant speech using the proposed LP-IMTF filtering framework and frame-dropping (FD)

Algorithm	t_{60} (seconds) clean	0.24	0.37	0.61	0.99
$EER\,(\%)$					
Baseline	5.89	8.47	11.38	16.31	25.58
LP-IMTF	5.39	7.55	9.24	12.63	20.26
LP-IMTF+FD	5.41	7.52	9.20	12.19	19.17
C_{llr}					
Baseline	0.240	0.350	0.704	1.350	2.448
LP-IMTF	0.273	0.283	0.385	0.698	1.679
LP-IMTF+FD	0.284	0.288	0.365	0.623	1.519

6. REFERENCES

[1] T. Houtgast and H. J. M. Steeneken, *The modulation transfer function in room acoustics as a predictor of speech intelligibility*, Acustica, 28, pp. 66-73, 1973.

[2] P. J. Castellano, S. Sradharan, and D. Cole, *Speaker recognition in reverberant enclosures*, Proc .of ICASSP, pp. 117-200, 1996.

[3] T. Langhans and H. W. Strube, *Speech Enhancement by Nonlinear Multiband Envelope Filtering*, Proc. of ICASSP, vol. 7, pp. 156-159, 1982.

[4] T. Kitamura et al., *Designing modulation filters for improving speech intelligibility in reverberant environments*, Proc. of ICSLP, vol. 3, pp. 586-589, 2000.

[5] M. Unoki et al., *A method based on the MTF concept for dereverberating the power envelope from the reverberant signal*, Proc. of ICASSP, vol. 1, pp. 888-891, 2003.

[6] C. Avendano and H. Hermansky, *On the properties of temporal processing for speech in adverse environments*, Proc. of WASPAA, 1997.

[7] O. Viikki and K. Laurila, *Cepstral domain segmental feature vector normalization for noise robust speech recognition*, Speech Communication, vol. 25, pp. 133-147, 1998.

[8] S. Greenberg and B. E. D. Kingsbury, *The modulation spectrogram: in pursuit of an invariant representation of speech*, Proc. of ICASSP, pp. 1647-1650, 1997.

[9] J. B. Allen and D. A. Berkley, *Image method for efficiently simulating small-room acoustics*, JASA, 65(4), pp. 943-950, 1979.

[10] ITU-T Recommendation P.862: Perceptual evaluation of speech quality (PESQ), ITU, 2001.

[11] D. Sturim et al., *The MIT LL 2010 Speaker Recognition Evaluation System: Scalable Language-Independent Speaker Recognition*, Proc. of ICASSP, pp. 5272-5275, 2011.

[12] N. Brummer and J. du Preez, *Application independent evaluation of speaker detection*, Computer Speech and Language, vol. 20, pp. 230-275, 2006.