Odyssey 2012
The Speaker and Language Recognition Workshop
25-28 June 2012, Singapore

# Linear Prediction Modulation Filtering for Speaker Recognition of Reverberant Speech

*Bengt J. Borgström and Alan McCree*

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

## Abstract

This paper proposes a framework for spectral enhancement of reverberant speech based on inversion of the modulation transfer function. All-pole modeling of modulation spectra of clean and degraded speech are utilized to derive the linear prediction inverse modulation transfer function (LP-IMTF) solution as a low-order IIR filter in the modulation envelope domain. By considering spectral estimation under speech presence uncertainty, speech presence probabilities are derived for the case of reverberation. Aside from enhancement, the LP-IMTF framework allows for blind estimation of reverberation time by extracting a minimum phase approximation of the short-time spectral channel impulse response. The proposed speech enhancement method is used as a front-end processing step for speaker recognition . When applied to the microphone condition of the NIST-SRE 2010 with artificially added reverberation, the proposed spectral enhancement method yields significant improvements across a variety of performance metrics.

## 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 was 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. In [4]-[6], modulation filters are designed to invert an exponentially decaying model of the acoustic channel impulse response. In [7]-[11], the authors design data-driven modulation filters according to the minimum mean-square error (MMSE) criterion. While improvements in perceptual quality are observed in these cases, the studies require oracle information regarding the room impulse response.

In this paper, we propose a method for spectral enhancement of reverberant speech based on inversion of the modulation transfer function. We discuss the MTF, 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, and implement it as a low-order IIR filter in the modulation envelope domain. The proposed method adapts to current acoustic conditions and therefore does not require oracle knowledge of the room impulse response. To improve reverberation suppression within inactive time-frequency speech regions, we explore spectral estimation under speech presence uncertainty, and derive speech presence probabilities for the case of reverberation. Although the proposed spectral enhancement method is applicable to a variety of applications, in this study it is applied as a pre-processing step for speaker recognition of reverberant speech.

Inferring the severity of acoustic distortion in an observed speech signal is useful for many applications. Whereas estimation of signal-to-noise ratio for additive background noise has been widely studied, eg. [12], blind estimation of reverberation time remains an important topic. In [13] and [14], reverberation time is estimated by locating abrupt stops in speech and analyzing decay rates in short-time subband energy envelopes. In [15], the authors present a data-driven system using a support vector machine. In this paper, we present a method for blind estimation of reverberation time based on the LP-IMTF framework. Using linear prediction of short-time spectral energy envelopes, a minimum phase approximation of the short-time spectral channel impulse response is determined, from which the reverberation decay time can be extracted. Blind estimation of reverberation time can utilized for designing condition-adaptive speaker recognition systems, which is a topic of future work.

This paper is organized as follows. In Sec. 2, the LP-IMTF filtering framework is proposed, and spectral estimation under speech uncertainty is discussed. Sec. 3 presents blind estimation of reverberation time. Sec. 4 includes experimental results for speaker recognition of reverberant speech. Finally, conclusions and future work are provided in Sec. 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.

## 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],[9]

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

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 $1 \leq k \leq N_{ch}$ and $m \geq 0$ refer to the channel and time index, respectively. In this study, we assume that $H_k(0)$=1. 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. It should be noted that there exist a multitude of parameters which can be used to characterize room impulse responses. However, throughout this study we utilize $t_{60}$ to efficiently summarize the severity of reverberation.

### 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) [16]. 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 [17]. 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

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

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.
- 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 (9)-(10). 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 speaker recognition, and should therefore not affect the IMTF filter shape. During implementation, normalized autocorrelation coefficients for channel $k$ can be estimated using a neighborhood of frequency channels

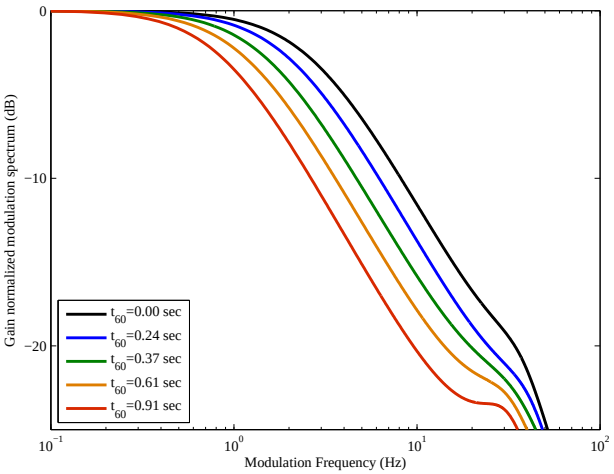


Figure 1: Gain-normalized all-pole modulation spectra of example speech in the presence of reverberation of varying degree, for the frequency channel with center frequency 1500 Hz, and for $P$=6

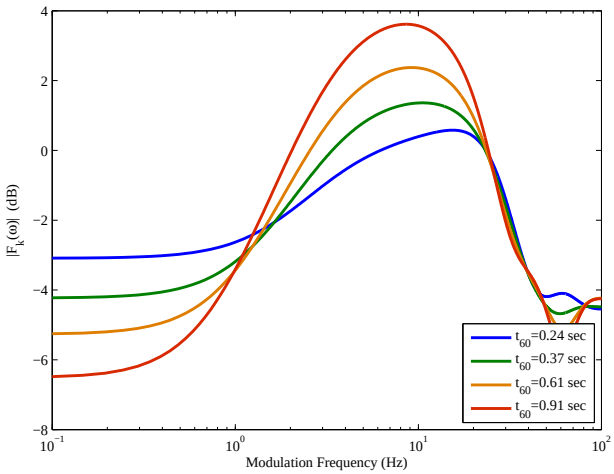


Figure 2: Proposed IMTF filters for example speech in the presence of reverberation of varying degree, for the frequency channel with center frequency 1500 Hz, and for $P$=6

$$r_{Y,k}(\tau) = \frac{\sum_{l=-Nr}^{Nr}\sum_{n=1}^{T}|Y_{k+l}(n)|\,|Y_{k+l}(n+\tau)|}{\sum_{l=-Nr}^{Nr}\sum_{n=1}^{T}|Y_{k+l}(n)|^2} \quad (7)$$

where $T$ denotes the number of frames, and $N_r$ controls the amount of inter-channel smoothing. The use of information from adjacent channels avoids discontinuous behavior of the LP-IMTF filter with respect to frequency channel. Additionally, it allows normalized autocorrelation coefficients to be estimated reliably for shorter observed utterances. Note that summation indices in (7) are restricted by lower and upper channel limits.

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

$$|M_{Y,k}(\omega)|^2 \approx \frac{\sigma_{Y,k}^2}{\left|1-\sum_{l=1}^{P} a_{Y,k}(l)\exp(-j\omega l)\right|^2} \quad (8)$$

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, for the example frequency channel with center frequency 1500 Hz, and for $P$=6. In this example, reverberation is added artificially to microphone interview speech from the 2010 NIST-SRE, using room impulse responses from [18]. It can be observed in Fig. 1 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

$$|F_k(\omega)| = \left|\frac{\sigma_{X,k}\left(1-\sum_{l=1}^{P} a_{Y,k}(l)\exp(-j\omega l)\right)}{\sigma_{Y,k}\left(1-\sum_{l=1}^{P} a_{X,k}(l)\exp(-j\omega l)\right)}\right| \quad (9)$$

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. Fig. 3 provides an illustrative example for the behavior of the LP-IMTF filter across short-time frequency channels, for a reverberation time of 0.91 seconds. The bandwidth of the LP-IMTF filter is observed to generally increase for higher frequency channels, with maximum filter depth at lower frequency channels. It is interesting to note that the frequency responses illustrated in Fig. 2 are consistent with studies on the relative importance of modulation frequencies for human speech perception [19], and automated speech applications [20], [21].

Since (5) does not account for phase, there exists 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 (9), 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

$$\left|\hat{X}_k(m)\right| = \frac{\sigma_{X,k}}{\sigma_{Y,k}}\left(|Y_k(m)| - \sum_{l=1}^{P} a_{Y,k}(l)\,|Y_k(m-l)|\right) + \sum_{l=1}^{P} a_{X,k}(l)\left|\hat{X}_k(m-l)\right| \quad (10)$$

Each frequency band of the observed short-time spectra is filtered with (10) to obtain enhanced spectral components.

To guarantee non-negativity in (10), processed spectral values must be floored. Spectral flooring, however, may result in undesirable nonlinear effects, which can be reduced by applying gain smoothing along time index and/or frequency channel. Traditional speech enhancement approaches designed to

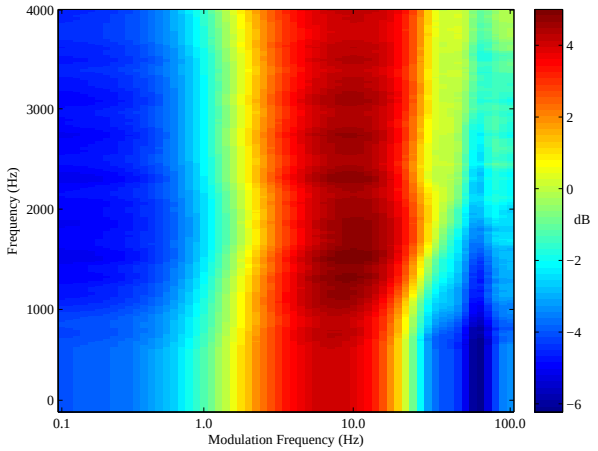


Figure 3: The proposed IMTF filter for example speech in the presence of reverberation with $t_{60}$=0.91 sec, and for $P$=6

combat additive noise often apply gain smoothing along time. However, in the case of reverberation suppression, such post-processing often reintroduces reverberant characteristics of the original degraded speech. This follows since spectral smoothing along time is mathematically similar to the short-time spectral reverberation model given by (2). In this study, we therefore apply gain smoothing along frequency channels.

It is of interest to note the connection between the LP-IMTF filter and the well-known RASTA filter, which is proposed in [22] as a front-end processing step for robust ASR or speech enhancement. The RASTA filter serves as an empirically-tuned IIR bandpass filter in the modulation spectral domain, designed to compensate for reverberation and slowly-varying additive noise components. When used for enhancement, the filter is applied to standard short-time spectra. As previously discussed, the LP-IMTF filter adapts to observed acoustic conditions by determining the ratio of all-pole models of modulation spectra. The RASTA filter can be interpreted as a specific instance of the LP-IMTF solution. In this interpretation, the frequency response of the RASTA filter can be decomposed into underlying all-pole modulation spectra of clean and observed speech. Comparison with LP-IMTF filters determined for a set of room impulse responses with varying reverberation times shows the RASTA filter to roughly correspond to a specific LP-IMTF solution with $t_{60}$=0.8.

### 2.4. Spectral Estimation Under Speech Presence Uncertainty

Spectral enhancement presented in Sec. 2.3 is designed under the assumption that active speech is present throughout time and frequency, and may therefore underattenuate reverberation within time-frequency segments of inactive speech. In this section, we apply speech presence probabilities (SPPs) during modulation filtering to improve spectral enhancement. Similar to [23] and [24], we assume an underlying two-state model wherein $\mathcal{H}_0$ and $\mathcal{H}_1$ correspond to inactive and active speech, respectively

$$\mathcal{H}_0 : |Y_k(m)| = \sum_{l=1}^{\infty} |H_k(l)|\,|X_k(m-l)| \tag{11}$$

$$\mathcal{H}_1 : |Y_k(m)| = |X_k(m)| + \sum_{l=1}^{\infty} |H_k(l)|\,|X_k(m-l)|$$

We assume complex DFT coefficients of clean speech and reverberation to be Gaussian distributed, so that spectral magnitudes follow Rayleigh distributions

$$p\left(|Y_k(m)|\,\middle|\,\mathcal{H}_0\right) = \frac{2\,|Y_k(m)|}{\sigma^2_{R,k}(m)} \exp\left(-\frac{|Y_k(m)|^2}{\sigma^2_{R,k}(m)}\right) \tag{12}$$

$$p\left(|Y_k(m)|\,\middle|\,\mathcal{H}_1\right) = \frac{2\,|Y_k(m)|}{\sigma^2_{X,k}(m) + \sigma^2_{R,k}(m)} \times \exp\left(-\frac{|Y_k(m)|^2}{\sigma^2_{X,k}(m) + \sigma^2_{R,k}(m)}\right)$$

where

$$\sigma^2_{X,k}(m) = E\left\{|X_k(m)|^2\right\} \tag{13}$$

$$\sigma^2_{R,k}(m) = E\left\{\Big|\sum_{l=1}^{\infty} |H_k(l)|\,|X_k(m-l)|\Big|^2\right\}$$

We define the *a posteriori* and *a priori* signal-to-reverberation (SRR) ratios, respectively, as

$$\zeta_k(m) = \frac{|Y_k(m)|^2}{\sigma^2_{R,k}(m)}, \quad \psi_k(m) = \frac{\sigma^2_{X,k}(m)}{\sigma^2_{R,k}(m)} \tag{14}$$

Given the previously discussed two-state model, the spectral estimate provided by (10) can be interpreted as being conditioned on $\mathcal{H}_1$. Conversely, in the case of $\mathcal{H}_0$, we assume the spectral magnitude of clean speech to be zero, leading to

$$\left|\hat{X}_k(m)\right| \Leftarrow P\left(\mathcal{H}_1\,\middle|\,|Y_k(m)|\right)\left|\hat{X}_k(m)\right| \tag{15}$$

Bayes' rule allows the posterior probability of speech presence of individual time-frequency components to be expressed as

$$P\left(\mathcal{H}_1\,\middle|\,|Y_k(m)|\right) = \frac{\Lambda\left(|Y_k(m)|\right)}{1+\Lambda\left(|Y_k(m)|\right)} \tag{16}$$

where the likelihood ratio can be derived using (12) and (14)

$$\Lambda\left(|Y_k(m)|\right) = \frac{P(\mathcal{H}_1)}{1-P(\mathcal{H}_1)} \frac{P\left(|Y_k(m)|\,\middle|\,\mathcal{H}_1\right)}{P\left(|Y_k(m)|\,\middle|\,\mathcal{H}_0\right)} \tag{17}$$

$$= \frac{P(\mathcal{H}_1)}{1-P(\mathcal{H}_1)} \frac{\exp\left(\frac{\zeta_k(n)\psi_k(n)}{1+\psi_k(n)}\right)}{1+\psi_k(n)}$$

Here, $P(\mathcal{H}_1)$ denotes the prior probability of active speech.

Note that $\zeta_k(m)$ represents an instantaneous value, and can be approximated by applying power spectral subtraction to (2), resulting in

$$\zeta_k(m) \approx \frac{|Y_k(m)|^2}{\max\left\{|Y_k(m)|^2 - \left|\hat{X}_k(m)\right|^2, 0\right\}} \tag{18}$$

In speech enhancement systems designed to target additive noise, techniques for estimating the *a priori* SNR, such as the decision-directed method from [24], often incorporate a large degree of temporal smoothing in order to reduce musical artifacts. However, as previously discussed, such smoothing often reintroduces reverberation, and we instead use

$$\psi_k(m) = \max\{\zeta_k(m) - 1, 0\} \tag{19}$$

which is similar to the spectral subtraction methods presented in [25]. The likelihood ratio from (17) then reduces to

$$\Lambda(|Y_k(m)|) = \frac{P(\mathcal{H}_1)}{1 - P(\mathcal{H}_1)} \frac{\exp(\zeta_k(m) - 1)}{\zeta_k(m)} \tag{20}$$

Using (20) and (16) with (15) yields LP-IMTF filtered spectra under speech presence uncertainty.

For realistic applications, speaker recognition may be performed on noisy reverberant speech, for which (2) can be generalized as

$$|Y_k(m)| = \sum_{l=0}^{\infty} |H_k(l)|\,|X_k(m-l)| + |D_k(m)| \tag{21}$$

where $|D_k(m)|$ denotes additive noise. In such cases, speech enhancement designed for additive noise (such as [24]) can initially be applied, followed by the proposed method of spectral enhancement. Thus, it is not necessary to implement simultaneous noise and reverberation suppression.

## 3. Blind Estimation of Reverberation Time

For many applications it can be useful to estimate reverberation time from an observed speech signal. For example, knowledge regarding the level of reverberation can be leveraged to implement condition-adaptive speaker recognition systems. In this section, we propose a method for blind estimation of reverberation time based on the LP-IMTF filtering framework.

From (5) and (8), an approximation can be derived for the modulation spectrum of reverberation in the short-time spectral domain

$$\left|\hat{M}_{H,k}(\omega)\right| = \left|\frac{\sigma_{Y,k}\left(1 - \sum_{l=1}^{P} a_{X,k}(l)\exp(-j\omega l)\right)}{\sigma_{X,k}\left(1 - \sum_{l=1}^{P} a_{Y,k}(l)\exp(-j\omega l)\right)}\right| \tag{22}$$

Applying the inverse DTFT to the expression within the absolute value operator in (22) yields the difference equation for the minimum-phase estimate of the observed acoustic channel impulse response

$$\left|\hat{H}_k(m)\right| = \frac{\sigma_{Y,k}}{\sigma_{X,k}}\left(\delta(m) - \sum_{l=1}^{P} a_{X,k}(l)\,\delta(m-l)\right) + \sum_{l=1}^{P} a_{Y,k}(l)\left|\hat{H}_k(m-l)\right| \tag{23}$$

where $\delta(m)$ denotes the Kronecker delta function.

Using (23), the 60 dB reverberation time for channel $k$ can be approximated as

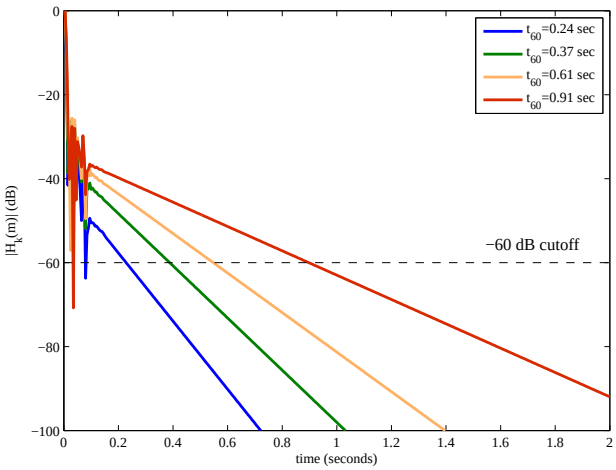


Figure 4: Acoustic channel impulse responses for example speech with various degrees of reverberation. Note that the estimated $t_{60}$, as determined by (24), corresponds to the right-most intersection of the impulse response with the $-60$ dB cutoff.

$$t_{60,k} = \max_{m} \frac{m}{f_w}, \text{ such that: } 20\log_{10}\left|\hat{H}_k(m)\right| \geq -60 \tag{24}$$

where $f_w$ is the windowing rate used during short-time spectral analysis. Although frequency channel-dependent reverberation times may be of interest for certain applications, others may only require a single metric for the overall acoustic severity. To determine an overall reverberation time, we use modulation envelope autocorrelation coefficients which have been averaged with respect to frequency channel, i.e. using (7) with $N_r$=$N_{ch}$. Fig. 4 illustrates acoustic channel impulse responses for example speech with various degrees of reverberation. It can be observed that the general decay rate of the impulse responses decreases as the acoustic severity increases. Further, when (24) is applied, which corresponds to the right-most intersection of the impulse response with the $-60$ dB cutoff, the estimated $t_{60}$'s are close to their corresponding true values.

## 4. Experimental Results

To assess the effectiveness of the LP-IMTF filtering framework, enhancement was applied as front-end processing to the MIT Lincoln Laboratory Joint Factor Analysis (JFA) speaker recognition system (see [26] for details). Experiments were conducted on a subset of interview microphone data from condition 2 of the 2010 NIST-SRE corpus. SAD was performed using a channel-wise SNR measure to include active interviewee speech frames but squelch leakage from the interviewer. The subset of data used included both male and female speakers, with 6.2 K targets and 1.7 M non-targets. Simulated room impulse responses were obtained from [18] for a range of $t_{60}$'s, and reverberation was artificially added to test cuts. Note that reverberation was not added during enrollment.

Table 1 provides objective quality measures for enhanced speech using the proposed LP-IMTF framework. Enhancement was applied to three minutes of speech from the TIMIT database, artificially reverberated with room impulse responses

Table 2: Speaker recognition results on reverberant speech using the proposed LP-IMTF filtering framework, pooled across uniformly distributed reverberation times of 0.24, 0.37, 0.61, and 0.99 sec. Reverberation was artificially added using room impulse responses from [18]. Results are provided for $EER$, $DCF$, and $C_{llr}$, along with relative improvements of each.

| Algorithm | $EER$ (%) | Rel. Imp. (%) | $DCF$ ($\times 10^3$) | Rel. Imp. (%) | $C_{llr}$ | Rel. Imp. (%) |
|---|---|---|---|---|---|---|
| Clean | 5.89 | – | 3.73 | – | 0.240 | – |
| Baseline | 12.79 | – | 6.08 | – | 0.733 | – |
| LP-IMTF | 11.16 | 23.6 | 5.52 | 23.8 | 0.461 | 55.2 |
| LP-IMTF + SPP | 10.93 | 27.0 | 5.49 | 25.1 | 0.453 | 56.8 |

Table 1: Speech enhancement results on reverberant speech using the proposed LP-IMTF filtering framework. Reverberation was artificially added using room impulse responses from [18].

| | $t_{60}$ (seconds) | | | |
|---|---|---|---|---|
| Algorithm | 0.24 | 0.37 | 0.61 | 0.99 |
| SD ($\times 10^2$) | | | | |
| Baseline | 0.71 | 1.46 | 3.12 | 9.66 |
| LP-IMTF | 0.48 | 0.81 | 1.55 | 4.00 |
| LP-IMTF + SPP | 0.44 | 0.68 | 1.19 | 2.78 |
| LSD | | | | |
| Baseline | 0.99 | 1.85 | 2.93 | 5.24 |
| LP-IMTF | 0.68 | 1.04 | 1.59 | 3.04 |
| LP-IMTF + SPP | 0.64 | 0.93 | 1.34 | 2.47 |
| PESQ | | | | |
| Baseline | 2.61 | 2.25 | 2.01 | 1.76 |
| LP-IMTF | 2.73 | 2.40 | 2.18 | 1.88 |
| LP-IMTF + SPP | 2.74 | 2.43 | 2.20 | 1.92 |

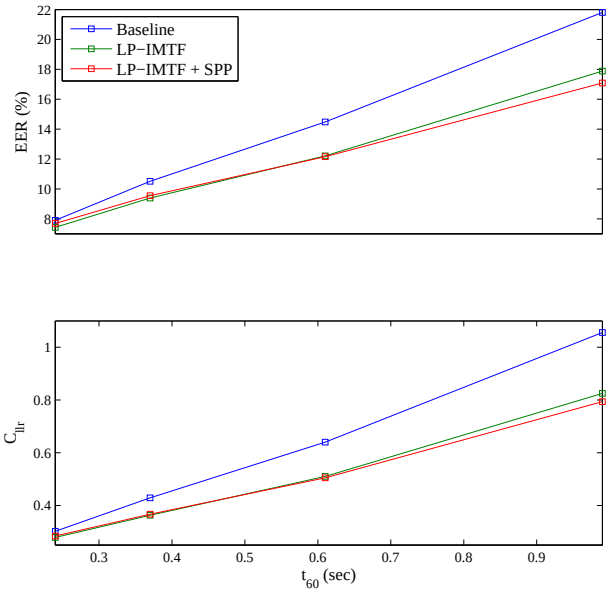


Figure 5: Speaker recognition results on reverberant speech using the proposed LP-IMTF framework. Reverberation was artificially added using room impulse responses from [18].

from [18]. Results are reported for mean spectral distortion (SD), mean log-spectral distortion (LSD), and PESQ [27]. Here, the baseline system refers to unprocessed reverberant speech signals. It can be observed that the LP-IMTF filtering framework yields significantly increased speech quality with respect to the reported metrics. Use of speech presence probabilities (SPP) provides further improvements.

Fig. 5 provides speaker recognition results on reverberant speech using the proposed LP-IMTF filtering framework. Results are provided in terms of equal error rate (EER) and the log-likelihood ratio cost ($C_{llr}$) [28]. Fig. 5 shows the LP-IMTF filter to yield significantly improved speaker recognition results across performance metrics. Spectral estimation under speech presence uncertainty provides further improvements in the more severe conditions.

Table 2 provides speaker recognition results using the proposed reverberation suppression method as a front-end processing step, pooled across reverberant conditions by assuming uniform distribution of the four RIRs used in Table 1. Results are provided in terms of EER, the 2008 NIST-SRE decision cost function (DCF), and $C_{llr}$. It should be noted that the 2010 NIST-SRE DCF was not reported since it involves a low target trial prior probability which may not be informative for severe acoustic environments. It can be observed in Table 2 that front-end reverberation suppression provides significant performance gains, resulting in 27% and 57% relative improvements for EER and $C_{llr}$, respectively.

## 5. Conclusions

This paper has proposed a method for spectral enhancement of reverberant speech based on inversion of the modulation transfer function. The LP-IMTF filter utilizes all-pole models of modulation spectra of clean and degraded speech to derive the LP-IMTF solution as a low-order IIR modulation filter. By considering spectral estimation under speech presence uncertainty, speech presence probabilities are derived for the case of reverberation. This paper proposed a method for blind estimation of reverberation time based on the LP-IMTF framework, by extracting a minimum phase approximation of the acoustic channel impulse response. When applied to speaker recognition of reverberant speech, the proposed system yields significant improvements across a variety of performance metrics.

Future work includes leveraging blind estimation of reverberation time to design condition-adaptive speaker recognition systems. Further, proposed methods can be applied other state-of-the-art speaker recognition systems, such as that from [29].

## 6. References

[1] T. Houtgast and H. J. M. Steeneken, *The modulation transfer function in room acoustics as a predictor of speech in-*

*telligibility*, Acustica, 28, pp. 66-73, 1973.

[2] P. J. Castellano, S. Sradharan, and D. Cole, *Speaker recognition in reverberant enclosures*, Proc. of ICASSP, vol. 1, 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] M. Unoki, M. Furukawa, K. Sakata, and M. Akagu, *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.

[5] M. Unoki, M. Toi, and M. Akagi, *Refinement of an MTF-based speech dereverberation method using an optimal inverse-MTF filter*, Proc. of SPECOM, pp. 323-326, 2006.

[6] M. Unoki, Y. Yamasaki, and M. Akagi, *MTF-Based Power Envelope Restoration in Noisy Reverberant Environments*, Proc. of EUSIPCO, pp. 228-232, 2009.

[7] H. Hermansky, E. A. Wan, and C. Avendano, *Speech Enhancement based on Temporal Processing*, Proc. of ICASSP, vol. 1, pp. 405-408, 1995.

[8] C. Avendano and H. Hermansky, *Study on the Dereverberation of Speech Based on Temporal Envelope Filtering*, Proc. of ICSLP, pp. 889-892, 1996.

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

[10] M. L. Shire and B. Y. Chen, *Data-driven filters in reverberation*, Proc. of ICASSP, vol. 1, pp. 1627-1630, 2000.

[11] T. Kitamura, K. Kinoshita, T. Arai, A. Kusumoto, and Y. Murahara, *Designing modulation filters for improving speech intelligibility in reverberant environments*, Proc. of ICSLP, vol. 3, pp. 586-589, 2000.

[12] R. Martin, *Noise power spectral estimation based on optimal smoothing and minimum statistics*, IEEE Trans. on Speech and Audio Processing, Vol. 9, No. 5, pp. 504-512, 2001.

[13] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. O'Brien, C. R. Lansig, and A. S. Feng, *Blind estimation of reverberation time*, JASA, 114(5), pp. 2877-2892, 2003.

[14] J. Y. C. Wen, E. A. P. Habets, and P. A. Naylor, *Blind estimation of reverberation time based on the distribution of signal decay rates*, Proc. of ICASSP, Vol. 1, pp. 329-332, 2008

[15] L. O. Nunes, L. W. P. Biscainho, L. Bowon, A. Said, T. Kalker, and R. W. Schafer, *Degradation Type Classifier for Full Band Speech Contaminated With Echo, Broadband Noise, and Reverberation*, IEEE Trans. on Audio, Speech, and Language Processing, Vol. 19, Issue 8, pp. 2516-2526, 2011.

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

[17] 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.

[18] S. Schimmel, M. Muller, and N. Dillier, *A fast and accurate "shoebox"' room acoustics simulator*, Proc. of ICASSP, pp. 241-244, 2009.

[19] R. van der Horst, A. R. Leeuw, and W. A. Dreschler, *Importance of temporal-envelope cues in consonant recognition*, JASA, 105(3), pp. 1801-1809, 1999.

[20] S. van Vuuren and H. Hermansky, *On the importance of components of the modulation spectrum for speaker verification*, Proc. of ICSLP, pp. 3205-3208, 1998.

[21] N. Kanedra, T. Arai, H. Hermansky, and M. Pavel, *On the relative importance of various components of the modulation spectrum for automatic speech recognition*, Speech Communication, No. 28, pp. 43-55, 1999.

[22] H. Hermansky and N. Morgan, *RASTA processing of speech*, IEEE Trans. on Speech and Audio Processing, Vol. 2, No. 4, pp. 578-589, 1984.

[23] R. McAulay and M. Malpass, *Speech enhancement using a soft-decision noise suppression filter*, IEEE Trans. Acoustics, Speech, and Signal Processing, vol. 28, no. 2, pp. 137-145, 1980.

[24] Y. Ephraim and D. Malah, *Speech enhancement using a minimum mean square error short-time spectral amplitude estimator*, IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. 32, pp. 1109-1121, 1984.

[25] M. Berouti, R. Schwartz, and J. Makhoul, *Enhancement of Speech Corrupted by Acoustic Noise*, Proc. of ICASSP, pp. 208-211, 1979.

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

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

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

[29] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, *Front-End Factor Analysis for Speaker Verification*, IEEE Trans. on Audio, Speech, and Language Processing, Vol. 19, Issue 4, pp. 788-798, 2011.