Spectro-Temporal Processing for Blind Estimation of Reverberation Time and Single-Ended Quality Measurement of Reverberant Speech

Tiago H. Falk, Hua Yuan, and Wai-Yip Chan

Department of Electrical and Computer Engineering
Queen's University, Kingston, Ontario, Canada
Email: {falkt, yuanh, chan}@ee.queensu.ca

Abstract

Auditory spectro-temporal representations of reverberant speech are investigated for blind estimation of reverberation time (RT) and for single-ended measurement of speech quality. The auditory representations are obtained from an eight-filter filterbank which is used to extract the modulation spectra from temporal envelopes of the speech signal. Gaussian mixture models (GMM), one for each modulation channel and trained on clean speech signals, serve as reference models of normative speech behavior. Consistency measures, computed between reverberant test signals and each GMM, are mapped to an estimated RT and to an estimated quality score. Experiments show that the proposed measures achieve superior performance relative to current "state-of-art" algorithms.
Index Terms: Reverberation time, quality measurement, GMM, modulation spectrum, consistency.

1. Introduction

When speech is produced in an enclosed environment, the acoustic signal follows multiple paths from source to receiver, resulting in reverberations. Reverberant signals sound distant and suffer from perceptual artifacts such as coloration and echoes. With the advances in hands-free telephony, reverberation has become a burden, in particular, to applications such as automatic speech recognition (ASR) and hearing aids. Quantifying reverberation is not easy and, commonly, the so-called reverberation time (RT) is used. RT, by definition, is the interval required for the sound energy to decay by 60dB after the sound source is turned off. Larger RT results in speech signals with decreased quality and intelligibility. Traditionally, room impulse responses or room geometry and absorptive properties are used to measure RT.

More recently, signal-based RT estimators that do not take into account room characteristics have been investigated; such "blind" estimation methods are important for real-time applications. In [1], a partially blind method is proposed where room characteristics are learned via a neural network approach. In [2], the diffuse tail of the reverberation is modeled as an exponentially damped Gaussian white noise. The time constant of the decay, obtained via a maximum-likelihood procedure, is used to blindly estimate RT. In the dereverberation literature, linear prediction (LP) residuals are widely used to quantify reverberation; several algorithms have been proposed based on this paradigm (e.g., [3, 4]). The idea is that for clean voiced speech segments, the LP residuals have strong peaks corresponding to glottal pulses; these peaks are spread in time for reverberant speech. The kurtosis of the LP residual is shown to be a good measure of RT [3].

In this paper, auditory spectro-temporal (ST) representations of speech are investigated for blind estimation of RT, as well as for single-ended quality measurement of reverberant speech. It is known that reverberation alters the modulation spectra of the speech signal (see e.g., [5]). Modulation frequencies are defined as the spectral content of the temporal envelope of the speech signal. Cues obtained from ST representations have been used by the ASR community to enhance features extracted from noisy speech (e.g., [6]). To the best of our knowledge, ST representations have not yet been used to estimate RT. Although ST representations have been used to estimate the quality of speech codecs and transmission systems [7], we show that the measure proposed here estimates the quality of reverberant speech more accurately.

Gaussian mixture models (GMM), trained on clean speech, are used as models of normative speech behavior. Different GMMs are obtained for eight different modulation frequency bands. Here, two experiments are performed. The first investigates the effects of reverberation on the modulation spectra. In the second, a more practical scenario is investigated where reverberation *and* acoustic background noise are present. In both cases, consistency measures, computed from degraded speech signals relative to each GMM, are mapped to an estimated RT and to an estimated quality score. We show that with simple linear mappings, improved estimation performance can be attained relative to current state-of-art schemes.

2. ST Representations of Speech

In order to obtain an auditory spectro-temporal representation akin to the one described in [8], the following processing steps were performed. First, the speech signal

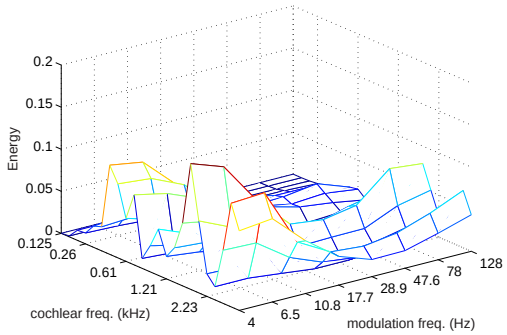

Figure 1: ST representation of a reverberant signal with $RT = 0.3$s.

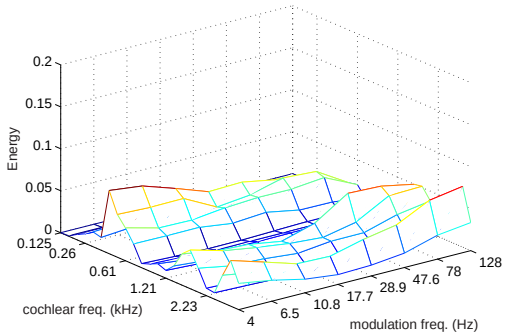

Figure 2: ST representation of a reverberant signal with $RT = 1$s.

level is normalized to -26dBov. The normalized speech signal is then filtered by a bank of critical-band filters. A critical-band gammatone filterbank, with 23 filters, is used to emulate the processing performed by the cochlea. The filter center frequencies range from 125Hz to 3.5kHz (speech signals used in our experiments have a sampling rate of 8kHz). The bandwidth of the filters are characterized by the equivalent rectangular bandwidth (ERB); the filter centered at 125Hz has a bandwidth of 38Hz, whereas the filter centered at 3.5kHz has a bandwidth of 410Hz. The Hilbert envelope is then obtained from each of the 23 signals output from the cochlear filterbank. An eight-filter modulation filterbank is applied to each temporal envelope. The center frequencies (f_c) and bandwidths (BW) of the modulation filters are described in Table 1. The output of the auditory processing module (for frame *j*) is termed $\mathbf{X}_j = \{\mathbf{x}_{1,j}, \ldots \mathbf{x}_{8,j}\}$ where each $\mathbf{x}_{i,j}, i = 1, \ldots, 8$, represents a 23-dimensional vector with one energy value for each cochlea frequency band.

Figures 1 and 2 depict the ST representation of a female speaker for an RT of 0.3s and 1s, respectively. The plots show the effects of increasing RT on different modulation frequency bands. Note that due to the increasing bandwidth of the filters, energy normalization, in principle, should be performed. For this study, however, normalization is not crucial since the computed measures are obtained *relative* to models of normative clean speech behavior. It is also important to emphasize that the energy envelopes obtained across cochlear frequencies, for the first few modulation channels (i.e., in accordance with normative speech behavior), resemble spectral envelopes obtained from LP analysis of the speech signal. For brevity, we omit plots that better illustrate this behavior.

3. GMMs and Consistency Calculation

Eight Gaussian mixture models are used to model $\mathbf{x}_i$, $i = 1, \ldots, 8$, for active speech frames. A Gaussian mixture density is a weighted sum of M component densities $p(\mathbf{x}|\boldsymbol{\lambda}) = \sum_{k=1}^{M} \alpha_k b_k(\mathbf{x})$, where $\alpha_k \geq 0, k = 1, ..., M$ are the mixture weights, with $\sum_{k=1}^{M} \alpha_k = 1$, and $b_k(\mathbf{x})$ are K-variate Gaussian densities with mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$. The parameter list, $\boldsymbol{\lambda}$=$\{\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_M\}$, defines a particular GMM, where $\boldsymbol{\lambda}_k$ = $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \alpha_k\}$. On our experiments, 32 diagonal covariance matrix Gaussian components are used.

The consistency between the reverberant test speech signal and each of the eight reference GMMs is computed as the normalized log-likelihood

$$c_m(\{\mathbf{x}_{m,j}\}_{j=1}^{N}) = \frac{1}{N}\sum_{j=1}^{N} \log(p_m(\mathbf{x}_{m,j}|\boldsymbol{\lambda})), m = 1, ..., 8$$

where $\{\mathbf{x}_{m,j}\}_{j=1}^{N}$ represents the feature vectors of the N active speech frames for modulation frequency band m, and p_m are the reference GMMs. Larger c_m indicates greater consistency with normative clean speech.

4. Experiment Setup

4.1. Generating Reverberant Speech

Reverberant speech is usually modeled as a convolution of the clean speech signal with the room impulse response, where the room impulse response can be measured or simulated (with e.g., the image model). Here, the SIREAC (SImulation of REal ACoustics) tool [9] is used to artificially generate reverberant speech with different RT. The room impulse response represents a typical office environment. The reader is referred to [9] for a more comprehensive description of the simulation tool.

We use 256 clean speech files from four male and four female speakers to train the reference GMMs. Half the sentences are in English and the other half in French. The SIREAC tool is used to artificially generate reverberant speech. In the experiment described in Section 4.2, speech is degraded with twelve different RT values: 0.1s–1s (0.1s increments), 1.5s, and 2s. In the experiment described in Section 4.3, speech is degraded with ten different RT values (0.1s–1s, 0.1s increments) and babble noise is also added at four different signal-to-noise ratios

(5dB–20dB, 5dB increments). Correlation (R) and root-mean-square error (ϵ) between reference and estimated measures are used as performance figures.

To test the capability of the proposed measure in performing single-ended measurement of reverberant speech quality, we compare the estimated quality scores with those produced by the state-of-art ITU-T P.862 (PESQ) quality measurement algorithm [10]. Ideally, subjective scores obtained from formal listening tests, such as the mean opinion score (MOS) test [11], should be used in this comparison. However, performing these formal tests is costly and time-consuming. Nonetheless, informal subjective tests carried out at our labs suggest that PESQ provides accurate quality estimates for reverberant speech, ranking similarly with the scores obtained subjectively. Furthermore, PESQ scores are shown to correlate well with the true RT ($R = -0.95$ on our data). Note that with PESQ, the original clean speech signal is required, thus limiting its usability in online applications.

4.2. Experiment 1: Reverberation Only

In this experiment, we investigate the effects of reverberation on different modulation frequency bands. The computed consistency measures are also tested as effective features for blind RT estimation and for single-ended quality measurement of reverberant speech. Table 1 describes the correlation obtained between the computed consistency measures and the true RT for each of the eight modulation channels. As can be seen, the lowest correlation values are obtained from modulation frequencies around 14Hz and 128Hz. Results also suggest that the important frequency bands for quantifying RT lie in the 4-8Hz and in the 30-80Hz range. This information can be used to improve performance of algorithms operating in reverberant conditions.

Moreover, positive correlations are attained for the first three modulation channels, which may at first seem counter-intuitive (larger RT resulting in higher consistency with clean speech). However, for smaller RT, reflections create irregular-period pitch pulses, thus distorting the excitation spectrum and hence the LP envelope. The number of such reflections increases with increasing RT, causing the excitation to look more Gaussian-noise like, thus having less impact on the LP envelope. As an example, the spectral distortion (SD) [12], averaged over 400ms of active speech, is computed between clean and reverberant speech; an $SD = 3.51$dB is attained for $RT = 0.1$s and $SD = 2.81$dB for $RT = 1$s. The plots depicted in Fig. 3 represent an extreme case where the SD for the signal with larger RT is 2dB lower than that of the signal with smaller RT. Recall from Section 2 that the energy envelopes used to train the GMMs resemble LP spectral envelopes for lower modulation frequency bands, thus positive correlations are indeed expected. Note that this behavior is consistent with the findings reported in [4].

Table 1: Correlation between consistency and true RT for each modulation frequency band. Filter center frequency (f_c) and bandwidth (BW), in Hz, are also shown.

	Modulation Frequency Band Index							
	1	2	3	4	5	6	7	8
f_c	4	6.5	10.8	17.7	28.9	47.6	78	128
BW	2	3.5	5.9	9.8	15.9	26.4	43.2	70.8
R	0.86	0.81	0.41	-0.30	-0.67	-0.71	-0.63	-0.28

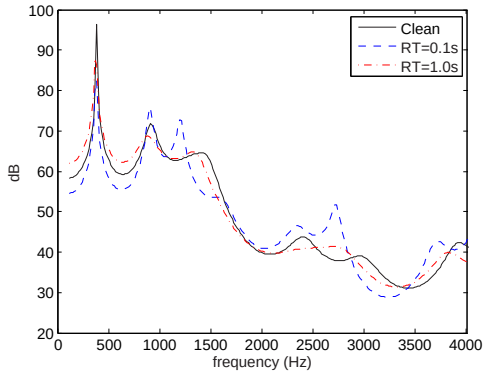

Figure 3: Comparison of 16^{th} order LP spectral envelopes for clean (solid) and reverberant speech with $RT = 0.1$s (dashed, $SD = 4.5$dB) and $RT = 1$s (dot-dashed, $SD = 2.5$dB).

In order to estimate RT and speech quality, mapping functions are devised between the eight consistencies and the measure to be estimated. The consistencies obtained from the English reverberant speech signals (total of 1536 signals) are used to train the mappings and the French reverberant signals are regarded as "unseen" and are used for testing. With a simple linear regression mapping function, an $R = 0.92$ and an $\epsilon = 0.27$ is attained between the true and estimated RT. Moreover, an $R = 0.93$ and an $\epsilon = 0.19$ is attained between the PESQ-MOS and the estimated MOS. Lower ϵ can be attained if more complex mapping functions are used; e.g., multivariate adaptive regression spline functions result in reductions in ϵ of 18% and 20% for RT and MOS estimation, respectively.

To compare performance with other blind measures, we use the kurtosis of the LP residual as a measure of RT, and the state-of-art ITU-T P.563 single-ended algorithm [13] as a measure of MOS. For fair comparisons, linear mappings trained on the English data are also devised; performance figures reported below are for the French dataset. In [3], the kurtosis of the LP residual is shown to correlate well with RT. We use the LP residuals of voiced frames computed from a 12^{th} order LP model. We note that good performance ($R = 0.83$) can be attained

only if signals with low RT are considered ($RT \leq 0.5$s). In our experiments, voicing decisions became unreliable with higher RTs. When considering all RT ranging from 0.1–2s, the kurtosis is computed from all active frames (voiced *and* unvoiced, as in [3]) and an $R = 0.52$ and $\epsilon = 0.68$ is attained. For P.563, an $R = 0.73$ and an $\epsilon = 0.52$ is attained relative to PESQ-MOS.

4.3. Experiment 2: Reverberation *and* Noise

In this second experiment, we consider a more practical application where background noise *and* reverberation are present. As an estimator of RT, the proposed measure achieves $R = 0.87$ and $\epsilon = 0.27$. Since LP residuals are sensitive to background noise we observe that a substantial decrease in performance is attained with the LP residual kurtosis measure ($R < 0.2$). Note that, in this experiment, the consistencies can also be used to blindly estimate SNR. With a linear mapping, $R = 0.94$ and $\epsilon = 1.59$ is achieved between true SNR and estimated SNR (in dB). This compares favorably with e.g., the SNR estimated by P.563, where an $R = 0.91$ and $\epsilon = 2.78$ is attained.

As an estimator of PESQ-MOS, the proposed measure attains $R = 0.82$ and $\epsilon = 0.16$. Due to the additional noise source, the range of possible PESQ scores decreases, thus a smaller ϵ is attained relative to Experiment 1. For comparisons, we use the ANIQUE quality measurement paradigm described in [7]; on our data, it resulted in better performance relative to P.563. With ANIQUE, quality scores are obtained based on the power ratio between low and high frequency modulation channels. The results reported here are based on an "in-house" implementation of the ANIQUE algorithm. The power ratio is computed between the four lower and four higher frequency modulation channels. Ratio and per-frame aggregations are performed as described in [7]. A linear mapping is devised between ANIQUE-MOS and PESQ-MOS and an $R = 0.73$ and $\epsilon = 0.22$ is attained. Clearly, the proposed measures outperform state-of-art algorithms in both RT and speech quality estimation tasks for both experiments described above.

5. Conclusions

We have investigated the use of auditory spectro-temporal processing of speech for blind estimation of reverberation time and for single-ended quality measurement of reverberant speech. Simulation results show improvement in estimation accuracy relative to current state-of-art measurement algorithms. The proposed measures are also shown to be more robust when reverberation *and* background noise are present. Furthermore, the insights obtained here can be used to improve the performance of algorithms that operate in reverberant environments, e.g., speech recognizers and dereverberation algorithms.

6. Acknowledgements

The authors would like to thank Dr. H. Günter Hirsch for providing the SIREAC simulation tool.

7. References

[1] T. Cox, F. Li, and P. Darlington, "Extracting room reverberation time from speech using artificial neural networks," *J. Audio Eng. Soc.*, vol. 49, no. 4, pp. 219–229, April 2001.

[2] R. Ratnam, D. Jones, B. Wheeler, W. O'Brien, C. Lansing, and A. Feng, "Blind estimation of reverberation time," *J. Acoustical Soc. of America*, vol. 114, no. 5, pp. 2877–2892, Nov. 2003.

[3] B. Gillespie, H. Malvar, and D. Florencio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *Proc. Intl. Conf. Acoustics, Speech, Signal Proc.*, May 2001, pp. 3701–3704.

[4] N. Gaubitch, D. Ward, and P. Naylor, "Statistical analysis of autoregressive modeling of reverberant speech," *J. Acoustical Soc. of America*, vol. 120, no. 6, pp. 4031–4039, Dec. 2006.

[5] B. Kingsbury, "Perceptually inspired signal-processing strategies for robust speech recognition in reverberant environments," Ph.D. dissertation, University of California, Berkeley, 1998.

[6] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. on Speech and Audio Proc.*, vol. 2, no. 4, Oct. 1994.

[7] D.-S. Kim, "ANIQUE: An auditory model for single-ended speech quality estimation," *IEEE Trans. on Speech and Audio Proc.*, vol. 13, no. 5, pp. 821–831, Sept. 2005.

[8] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. I – Modulation detection and masking with narrowband carriers," *J. Acoustical Soc. of America*, vol. 102, pp. 2892–2905, 1997.

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

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

[11] ITU-T P.830, "Subjective performance assessment of telephone-band and wideband digital codecs," Intl. Telecom. Union, 1996.

[12] W. B. Kleijn and K. K. Paliwal, Eds., *Speech Coding and Synthesis*. Amsterdam: Elsevier, 1995.

[13] ITU-T P.563, "Single-ended method for objective speech quality assessment in narrow-band telephony applications," Intl. Telecom. Union, 2004.