15th European Signal Processing Conference (EUSIPCO 2007), Poznan, Poland, September 3-7, 2007, copyright by EURASIP

OBJECTIVE MEASUREMENT OF COLOURATION IN REVERBERATION

Jimi Y. C. Wen, Patrick A. Naylor

Imperial College London, UK
E-mail: {yung-chuan.wen,p.naylor}@imperial.ac.uk

ABSTRACT

An objective measure for the perceived effect of reverberation is an essential tool for research into dereverberation algorithms. There are two different effects that contribute to the total perceived reverberation: colouration and reverberation decay tail effect. In this paper we aim to investigate an objective measurement of colouration of the speech signal in the presence of room reverberation. This measurement only requires the original reference signal and output signal and does not require an estimate of the room impulse response. The colouration measurement is performed only at onsets of speech activity in each bin of a short time FFT analysis.

1. INTRODUCTION

Reverberation is caused by the multi-path propagation of acoustic signals from source to receiver. Reverberant speech can be described as sounding distant with noticeable echo and colouration. Reverberation has a negligible effect in telephony applications with traditional handsets. However, in hands-free systems, reverberation affects the quality and intelligibility of speech and is a significant problem for telecommunications, speech recognition applications, and hearing aids [1]. Dereverberation is the process of forming an estimate of the original source from one or more observations of the reverberant signal. Several dereverberation algorithms have been proposed and can be considered in three categories: (i) speech enhancement processing (ii) beamforming and (iii) blind channel estimation and inversion algorithms [1]. There is a need to have a reliable and consistent objective measure for the perceived reverberation or reverberation reduction, which can allow the perceptual significance of a particular algorithm's processing to be evaluated. Reliable quantitative measurement of the level of reverberation in a speech signal is particularly difficult and a unanimously accepted methodology has yet to emerge [1]. It has been observed that most of the current objective speech quality measures that give good prediction of overall speech quality were developed in the context of speech coding and do not give good and consistent prediction for effects of reverberation.

Various spectral difference measures are employed to quantify algorith performance when evaluating reverberant and dereverberated speech. For example, the Bark Spectral Distortion (BSD) [2] is an often used objective measure for speech quality. Assuming the only 'distortion' present is the room reverberation, the BSD measures the room reverberation as a perceptually weighted spectral difference of the original and reverberant signal. This difference consists of two parts: the colouration effect, and the reverberation decay tail effect [3, 4, 5]. The colouration effect is due mostly to the stronger early room reflections [6], and the reverberation decay tail effect is due to the later response of the acoustic room impulse response. The direct sound added to one or more strong reflections result in a spectrum of peaks and valleys. For delay times up to about 25 ms the colouration is mainly due to frequency-response variations [7]. Some speech enhancement methods for dereverberation can also introduce artifacts that cause colouration. Whether the colouration arises from room reverberation or algorithm processing, specific colouration estimates can also help in the design of enhancement methods for equalization. Schroeder et al. [8] investigated the colouration caused by one strong reflection using white noise as input signal. The authors proposed two objective criteria to predict the perception of colouration: a spectral criterion and an autocorrelation criterion. The spectral criterion states that colouration is perceptible if the level difference between max-value and min-value of the short-time power spectrum exceeds a threshold A_0. The autocorrelation criterion states that colouration is perceptible if the ratio of the max value of the short-time autocorrelation function for any nonzero delay to its value at zero delay exceeds a threshold B_0. Berkley later suggested that spectral deviation of a room transfer function could predict spectral colouration, and showed that the variance of the frequency response irregularities was well correlated to the subjective perception of coloration [9]. Both Schroeder's and Berkley's measure require room responses for spectrally flat input signals.

Room impulse responses (or how this response is modified by a dereverberation algorithm) may not be available or may be difficult to estimate for time-varying processing, nonlinear processing, or systems with the combination of both. For example, in [10], the authors used a dereverberation algorithm that does not give an equivalent impulse response. In these cases it is obviously not possible to employ measures based on the impulse response. Therefore a measure which only requires some form of the difference of the reverberant and original reference speech signals would be more practical. In this paper we aim to investigate an objective measurement of colouration that measures only the colouration of the speech signal in the presence of room reverberation, and this measurement only requires the reference and output signals. The outline of the paper is as follows. Section 1 is an introduction to the application and objectives. Section 2 is the formulation of colouration estimate using onset detection. Sections 3 and 4 describe the testing procedure and results of the experiments. We conclude with Section 5.

2. COLOURATION ESTIMATION

As mentioned in section 1, spectral difference measures such as the BSD measure the room reverberation as a perceptu-

ally weighted spectral difference between the original and reverberant signal. This difference consists of two parts: the colouration effect, and the reverberation decay tail effect. In [5], the authors presented a measure R_{DT}, which aims to measure the reverberation tail effect independently of the colouration. This is achieved by computing the measure over frames where the reverberation tail effect is strongest compared to the colouration, such as after the end-points. Now we propose a measurement for the colouration.

2.1 Approach

The aim is to measure the colouration added due to the room reverberation or speech enhancement algorithm but independent of the reverberation decay tail effect. We use the assumption that the colouration is dominated by the early part of the impulse response, which we denote the early reflections. Our approach is to estimate the effect of room/process colouration by examining onsets in the input speech at different frequencies using time-frequency analysis. The only points where we have 'visibility' of the effect of early reflections are at speech onsets. By working in the time-frequency domain it is possible to find more onsets to improve the accuracy and robustness of the colouration estimate comparing to finding onsets in the time domain. By 'visibility', we mean where the time-aligned differences of the input/output short-time spectra are mostly due to the early reflection and the effect of reverberation decay tail of previous speech frames is negligible, therefore measuring the colouration independent of reverberation.

We define $S_{in}(n,k)$ as the magnitude of the of short-time Fourier transform (STFT) of the reference speech at time frame n and frequency index k. The corresponding magnitude STFT of the reverberant speech, $S_{out}(n,k)$ can be defined as:

$$S_{out}(n,k) = \sum_{i=0}^{n-1} S_{in}(n-i,k) H_{env}(i,k), \qquad (1)$$

where $H_{env}(i,k)$ is a reverberation decaying factor of the i-th previous speech frame. The product $S_{in}(n-i,k)H_{env}(i,k)$ is the speech energy of i-th previous frame contributing to the output speech of the current frame. Assuming the room reverberation has an exponential decay in the time-frequency domain, the reverberation decay, $H_{env}(i,k)$ has the following form:

$$H_{env}(i,k) = e^{-\lambda_k i}, \qquad (2)$$

where λ_k is the decay rate at k-th bin. The contribution, $C(n,i,k)$ to frame n of the output speech due to the input speech energy at $S_{in}(n-i,k)$ to the next i frame of the input speech is related by the decay, λ_k of the reverberation at the frequency k with the following form:

$$C(n,i,k) = S_{in}(n-i,k) H_{env}(i,k). \qquad (3)$$

Now redefining $C(n,i,k)$, $S_{in}(n,k)$ and $H_{env}(i,k)$ in dB by taking logs we write

$$C(n,i,k) = S_{in}(n-i,k) - \lambda'_k i \quad \text{dB}, \qquad (4)$$

where $\lambda'_k = -20\log_{10}(e)\lambda_k$. The contribution energy at the i-th previous frame is dependent on two factors: the value of i and the input speech energy at the i-th previous frame.

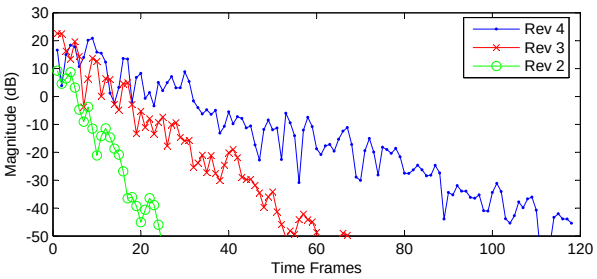

Figure 1: Example of the linearity of reverberation decays for frequency bin 50 for three different reverberation lengths.

Contribution energy decreases as i increases. High speech energy at the i-th previous frame also causes a higher contribution energy at the current frame. The value of λ_k will vary for different implementation of the STFT, eg. sample rate, FFT size, window size and overlap for a given reverberation decay.

2.2 Onset Detection

The objective of the onset detection is to find speech frames that do not have a high contribution energy due to a previous frame. Instead of doing a look back operation we will implement the algorithm in the reverse time with a look ahead operation replacing n with n' for reverse time frame index. The algorithm needs to compare the current speech energy $S_{in}(n',k)$ to speech energy of the preceding frames to see if there are any frames that could potentially contribute to the current frames significantly. For real-time processing, a look back version can be implemented with straightforward modification. We define an onset to be detected when the following criteria are satisfied:

1. $S_{in}(n',k) > Max(S_{in}(n',k)) - \gamma_1 \quad \forall\, n'$
2. $S_{in}(n',k) - S_i(n'+i',k) > \gamma_2 - \gamma_3(i'-1)$

where $S_{in}(n',k)$ is the energy of the speech at reversed time frame n' and frequency bin k. i' is defined as:

$$i' = 1,2,\ldots N - n', \qquad (5)$$

where N is the total number of time frames. The values of γ_1, γ_2 and γ_3 are the design parameters of the onset detection algorithm. *Criteria 1* is to eliminate frames with low speech energy. . We have empirically chosen γ_1 to be 18 dB. *Criteria 2* is derived from (3), where γ_2 is set as the limit of the effect that any single contribution energy due to the reverberation decay tail effect of previous frames or preceding (reverse) time frames and is also determined empirically as 18 dB. The value of γ_3 is the gradient of the decay. From Fig. 1, it can be seen that the reverberation impulse response with an exponential decay envelope is approximately linear with a logarithmic scale. With a reverberation time of 1 s, the gradient of the linear decay is approximately 0.05. We will set γ_3 to the worst case of $\gamma_3 = 0.05$. Fig. 2 shows an example of the onset detection algorithm's operation.

2.3 Colouration Estimate Procedure

The colouration estimate formed from onset frames only, $F(k)$, is the average of the gain (difference of logs between

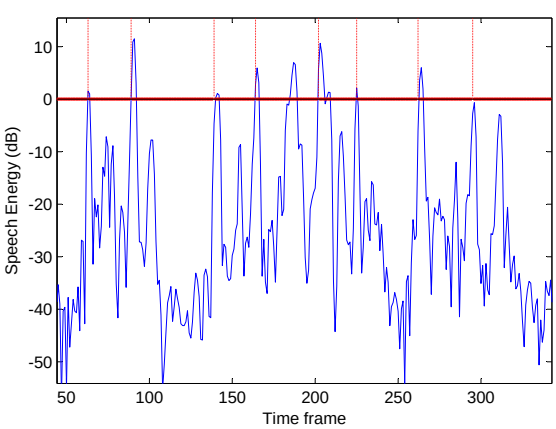

Figure 2: Example of the find onset algorithm for a particular frequency bin. Red dash spikes indicates an acceptable onset detected.

reference and output magnitude STFT) at the points of the onset for a given frequency bin, and is defined as:

$$F(k) = \sum_{n=1}^{N} D(n,k)(S_{out}(n,k) - S_{in}(n,k))/M_k, \qquad (6)$$

where $S_{out}(n,k)$ is the time-aligned output magnitude STFT, and $D(n,k) = 1$ for the onset frames, and $D(n,k) = 0$ otherwise. M_k is the number of onsets detected for the particular frequency bin k. We will compare our proposed method using onset frames only with a baseline method, which averages the gain of every frame for each frequency bin. We reduce the number of frequency bins for a 1024 point FFT from 513 ($F(k)$) to a smoothed colouration with a size of 32 ($\mathbf{f}$) by performing uniform averaging. Non-uniform averaging, such as log spacing, bark scale, etc. can also be used to obtain more perceptual advantages.

3. DESCRIPTION OF EXPERIMENTS

In this section we will describe the coloured reverberation impulse response generated for the testing of our proposed onset colouration estimation. We will also describe how the performance of our colouration estimates are evaluated.

3.1 Coloured Reverberation Impulse Response Set

We use stochastic room models, $h(t)$ for the basis of our reverberation impulse response,

$$h(t) = b(t)e^{-\lambda_t t} \qquad (7)$$

where $b(t)$ is white zero-mean Gaussian distributed noise, and the time domain decay rate, λ_t is linked to the reverberation time T_{60} through [11]:

$$\lambda_t = \frac{3ln10}{T_{60}}. \qquad (8)$$

Increasing λ_t will increase the rate of decay and decrease the reverberation time T_{60}.

Room reverberation typically results in some form of band-pass colouration. We propose to add four different colourations of spectral shape shown in Fig. 3(a). We can also see that in the time domain of the colouration filters

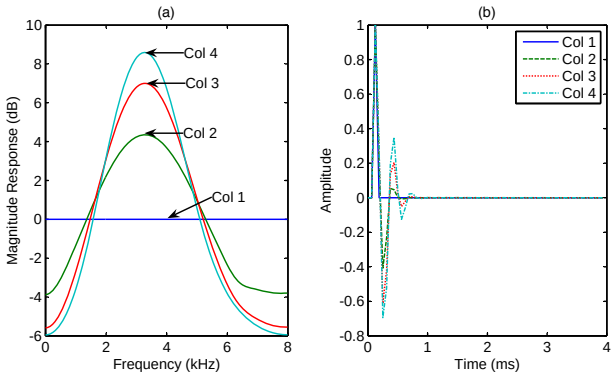

Figure 3: Colouration filters of different level : (a) frequency domain and (b) time domain.

of Fig. 3(b), the filters are short with a duration of a few milliseconds. Short filters ensure that there is no significant change to the reverberation decay tail effect when they are applied to the reverberation impulse response sets. In [6, 7], the authors stated that colouration is mainly due to the stronger early reflection from $25-50$ ms, therefore we will apply the colouration filter to just the first 32 ms of the reverberation impulse response leaving the late reverberation to be approximately white. To evaluate the consistency and robustness of the colouration estimate, we include three different speakers of the same utterance (two males, one female) and three different utterances of a same male speaker along with four colouration levels and four reverberation lengths (10 ms, 0.4 s, 0.7 s and 1 s). Speech files and impulse responses are sampled at 16 kHz. The STFT is implemented using 512 point hamming windowed frames, zero padded to a 1024 point FFT frame. Frame overlapping of 128 samples is used.

3.2 Testing Procedure

The normalized projection misalignment (NPM) [12] is used as the performance measure for channel identification in the time-domain. It measures the closeness of the reference signal to the estimated signal without regard to the scaling. We will adopt and modify the NPM measure for comparing the smoothed true reference colouration $\mathbf{f}$ (Fig. 3) to the smoothed estimated colouration $\hat{\mathbf{f}}$, given by

$$\mathrm{NPM}_f\left\langle \mathbf{f}, \hat{\mathbf{f}} \right\rangle = 20\log_{10}\left(\left\| \mathbf{f} - \frac{\mathbf{f}^T\hat{\mathbf{f}}}{\hat{\mathbf{f}}^T\hat{\mathbf{f}}}\hat{\mathbf{f}} \right\| \Big/ \|\mathbf{f}\| \right) \text{ dB}. \qquad (9)$$

Consistency and robustness of the colouration estimation are key factors for either evaluating the colouration for different rooms/algorithms or designing equalizers/enhancements. So while the absolute scaling of the colouration estimate is not critical, the relative scaling is important. We therefore show additional results for NPM_f computed over multiple colouration estimates for a given variable parameters. We call these additional NPM_f, *combined* NPM_f across a given variables parameters.

4. RESULTS

In this section we show different comparisons to illustrate the improvements of colouration estimation using the proposed

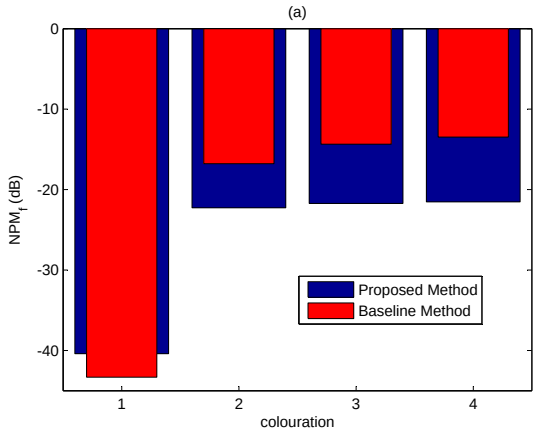

Figure 4: Performance of averaged (6 speech sets) NPM_f measures of: 4 different reverberation times for varying colouration.

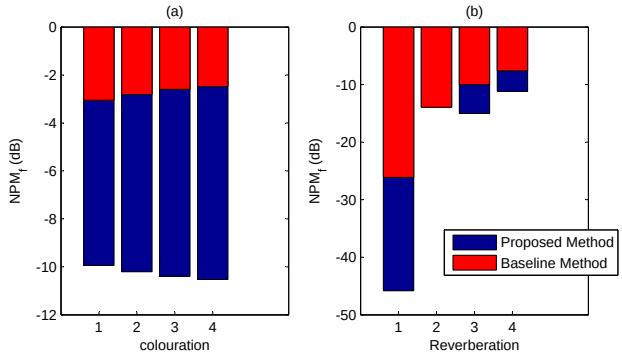

Figure 5: Performance of averaged (6 speech sets) *combined* NPM_f measures across: (a) different reverberation times for varying colouration (b) different colouration for varying reverberation times.

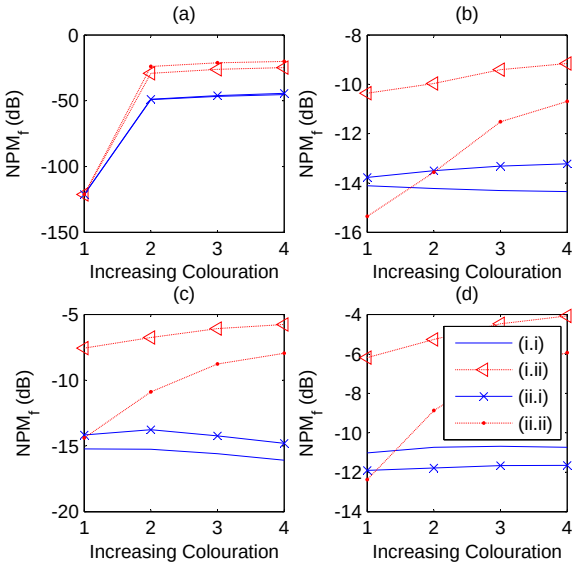

Figure 6: *Combined* NPM_f measure across (i.i) different speakers-Proposed method, (i.ii) different speakers-Baseline method, (ii.i) different utterance-Proposed method, (ii.ii) different utterance-Baseline method for different reverberation times (a)(b)(c)(d) and different colouration.

method compared to the baseline method.

4.1 General Performance

The average of all NPM_f measures is -26.5 dB for the propsed method and -22.0 dB for the baseline method. Fig. 4 shows the average of the NPM_f for different speakers, utterances and reverberations for different colouration levels. It can be seen that increasing colouration increases the improvement of the proposed method over the baseline. However since the averaging in Fig. 4 is across all variations of reverberation, utterance and speaker, we will use the *combined* NPM_f to evaluate the performance in more detail.

4.2 Reverberation and Colouration Consistency

Figure 5(a) shows the *combined* NPM_f across different reverberation times for varying colouration averaged over different speakers/utterances. This comparison is used to test the consistency of the colouration estimate against different reverberation lengths. Regardless of the reverberation length, the colouration should be consistent from the colouration estimate. It can be seen that the proposed method has about $6-8$ dB improvement compared to the baseline method. It can also been seen that for stronger colouration, ie. stronger early reflection, the performance of the proposed method improves.

Figure 5(b) shows the *combined* NPM_f across different colouration for varying reverberation time averaged over different speakers/utterances. This comparison is used to test the consistency of the colouration estimate against different colouration levels. There are also general improvements for the proposed method over the baseline method, however both methods suffer with increasing reverberation.

4.3 Speaker and Utterance Consistency

Figure 6 shows the *combined* NPM_f across different speakers/utterances for both methods and for different reverberations and colourations. It can be seen that the proposed method is more consistent across different speakers and utterances for a given reverberation time and colouration, whereas the performance of the baseline method is less good. This is because computing colouration for all frames will measure all the contributions of all speech energy due to the reverberation decay tail effect. This is not the case for computing at the onset frames only, as the contribution due to previous reverberation decay tail is minimized. The consistency of the baseline method suffers more for different speakers than different utterances.

Figure 7 shows the colouration estimate for reverberation length 4 (T_{60} =1 s), colouration 4 and different speakers (a)(b)(c). It can be seen that for the propsed method, the colouration estimate is very consistent, while the consistency of the baseline method suffers from all contributions due to the reverberation decay tail effect of previous speech energy.

5. CONCLUSION

Room reverberation has two distinct effects: the colouration effect and reverberation decay tail effect. In this paper we

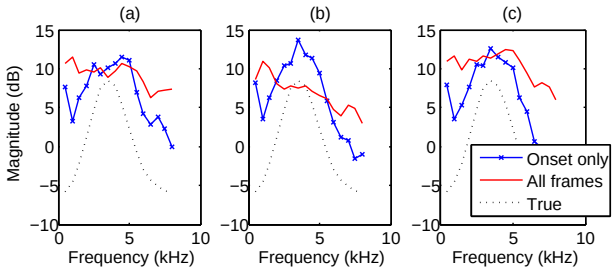

Figure 7: Comparison of colouration estimate against the true colouration for different speakers using onset only method and all frames method.

proposed a colouration measurement, which operates on onset frames only to improve the accuracy and consistent for different varying parameters. Our proposed colouration measurement only requires the reference and output signal. Experiment results show that the proposed method is more accurate, consistent and is more independent to reverberation decay effect than the defined baseline.

REFERENCES

[1] P. A. Naylor and N. D. Gaubitch, "Speech dereverberation," in *Proc. Int. Workshop Acoust. Echo Noise Control*, 2005.

[2] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE Journal on Selected Areas in Communications,*, vol. 10, no. 5, pp. 819 – 829, 1992.

[3] H. Kuttruff, *Room Acoustics*, Taylor & Francis, 4 edition, Oct. 2000.

[4] J. B. Allen, D. A. Berkley, and J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *J. Acoust. Soc. Amer.*, vol. 62, pp. 912, 1977.

[5] J. Y. C. Wen and P. A. Naylor, "An evaluation measure for reverberant speech using tail decay modeling," in *Proc. European Signal Process. Conference*, 2006.

[6] M. R. Schroeder and B. F. Logan, "Colorless artificial reverberation," *IRE Transactions on Audio*, vol. 9, Issue: 6, pp. 209 – 214, Nov 1961.

[7] Per Rubak, "Colorationin room impulse responses," in *Proc. of Joint Baltic-Nordic Acoustic Meeting*, June 2004.

[8] M. R. Schroeder, B. S. Atal, and K. H. Kuttruff, "Perception of coloration in filtered gaussian noiseshort-time spectral analysis by the ear," *J. Acoust. Soc. Amer.*, vol. 34, pp. 738, 1962.

[9] D. A. Berkley, *Acoustical factors affecting hearing aid performance*, chapter Normal listeners in typical Rooms - Reverberation Perception, Simulation, and Reduction, pp. 3–24, Baltimore : University Park Press, 1980.

[10] N.D. Gaubitch, P. A. Naylor, and D.B. Ward, "Multimicrophone speech dereverberation using spatio-temporal averaging," in *Proc. European Signal Process. Conference*, 2004, pp. 809–812.

[11] K. Lebart, J.M. Boucher, and P.N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica*, vol. 87, pp. 359–366, 2001.

[12] D. R. Morgan, J. Benesty, and M. M. Sondhi, "On the evaluation of estimated impulse responses," vol. 5, no. 7, pp. 174–176, July 1998.