ACTA ACUSTICA UNITED WITH ACUSTICA
Vol. 92 (2006) 337 – 339

A Pitch-Based Method for the Estimation of Short Reverberation Time

Mingyang Wu, DeLiang Wang
Department of Computer Science & Engineering and Center for Cognitive Science, The Ohio State University, Columbus, OH 43210-1277, USA.
MingyangWu@fairisaac.com, dwang@cse.ohio-state.edu

PACS no. 43.55.Mc

Summary

Reverberation corrupts harmonic structure in voiced speech. We observe that the pitch strength of voiced speech segments is indicative of the degree of reverberation. Consequently, we present an estimation method of reverberation time (T_{60}) based on pitch strength. The pitch strength is measured by deriving the statistics of relative time lags, defined as the distances from the detected pitch periods to the closest peaks in a correlogram. The monotonic relationship between the measured pitch strength and reverberation time learned from a corpus of reverberant speech with known reverberation times yields an estimate of T_{60} up to 0.6 seconds.

1. Introduction

Reverberation time is an important quantity characterizing room acoustics. Traditional room reverberation time measurements use synthetic sounds, such as an impulsive excitation, a white noise, or a swept sine wave (for example, see [1]). Such measurements are often cumbersome in practice; for example, they need to be conducted when a room is unoccupied by people [2]. On the other hand, experienced acousticians are able to estimate reverberation times rather precisely by listening to speech or music in a room, and in many situations it is desirable to estimate this quantity from reverberant speech directly. Such blindly estimated reverberation times can facilitate acoustic processing tasks in reverberant environments such as reverberant speech enhancement.

Cox *et al.* [2] proposed a reverberation time estimation algorithm from reverberant speech using artificial neural networks. Reverberation smears the temporal structure in speech and, therefore, flattens its energy envelopes; the flatness indicates the degree of reverberation. Taking advantage of this fact, a neural network model is trained on speech samples with known reverberation times and later used to determine the reverberation time in a room. However, speech utterances are restricted to individually pronounced digits and uncontrolled situations are not considered. In a subsequent study, Li and Cox [3] extended the neural network approach to the estimation of speech transmission index which predicts speech intelligibility in a transmission channel that may contain both reverberation and additive noise. Supervised training uses the extracted envelope spectrum of speech. The method works well with a given excitation but is not accurate with arbitrary speech. Recently, Ratnam and his coworkers [4] [5] attempt to estimate the reverberation time by modeling a reverberation tail as an exponentially damped Gaussian white noise process, and the reverberation time is estimated employing a maximum-likelihood procedure. Their model assumes source signals with abrupt offsets and long gaps between sound segments. Consequently, the authors report that the estimated reverberation times are in good agreement with the actual values when the source signals are white noise bursts and hand-claps. The gradual offsets in speech sounds, however, violate the model assumptions and, therefore, introduce significant bias to model estimates, particularly when the reverberation times are relatively short and source signals are connected speech.

Besides other manifestations of reverberation in a speech signal, reverberation corrupts harmonic structure in voiced speech; that is, the harmonicity (or periodicity) that characterizes voiced speech is weakened. We have found that the degree of corruption can be used as an indication of reverberation [6]. In this letter, we develop a pitch-based method for blind estimation of the reverberation time (T_{60}), which is the time taken for the sound level to drop by 60 dB after the excitation is turned off. The next section gives the detailed explanation of the method and Section III concludes this letter.

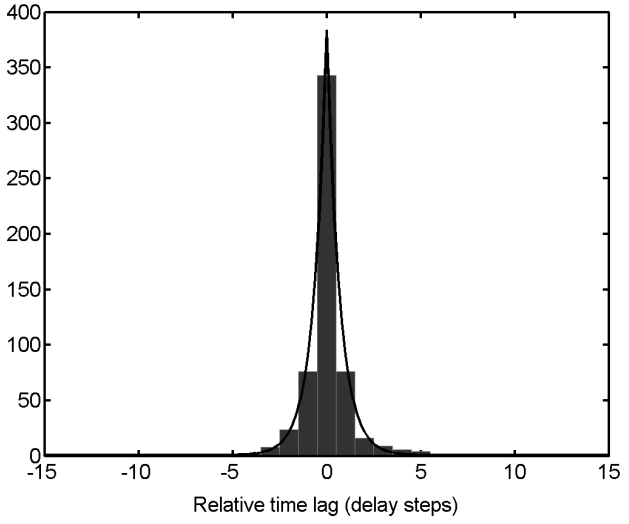
2. Proposed measure

A speech signal can be classified into three different sections: voiced, unvoiced, and silence. Obviously, pitch-based estimation of reverberation could be based only on voiced time frames. Moreover, in a noisy background, some frequency channels in a voiced frame may be severely corrupted by noise. This estimation should be thus based on the signals from "clean" frequency channels. In order to satisfy these criteria, our method, detailed below, employs a simplified version of a recent multipitch tracking algorithm [7]. That algorithm can track pitch periods reliably and can also be used to provide voiced/unvoiced labeling. In addition, it has a channel selection method for identifying weakly corrupted frequency channels on which the pitch-based measure is based.

The pitch-tracking algorithm consists of four stages. In the first stage, the input signal is sampled at 16 kHz and then filtered into 55 frequency channels by a bank of fourth-order gammatone filters [8] with center frequencies equally distributed on the equivalent rectangular bandwidth scale between 80 Hz and 800 Hz. As a result of using information from only lower frequency channels, strictly speaking, the estimated reverberation time is of the lowpass nature (lower than 1275 Hz), although no distinction is made between broadband and narrowband definitions of T_{60} in the evaluation. At the end of the first stage, normalized correlograms (autocorrelations) are computed using a window size of 16 ms in all channels. Channel and peak selection forms the second stage. Based on the shapes of normalization correlograms, only channels weakly corrupted by noise are selected and passed to later processing (see [7] for detailed explanations). Depending on speech and reverberation, channel selection typically retains most of the channels, e.g. over 85% for anechoic speech. The third stage integrates periodicity information across all channels and the final stage forms continuous pitch tracks using a hidden Markov model. This simplified version of the algorithm is restricted to single pitch tracks, as the present study only deals with single speech sources. Our experiments show that this algorithm performs reliably under modest reverberant conditions.

Our key observation is that the differences between a detected pitch period by the pitch tracker and the time lags from the closest peaks of normalized correlograms in selected channels indicate the level of degradation in the harmonic structure caused by

Received 29 December 2004,
accepted 11 October 2005.



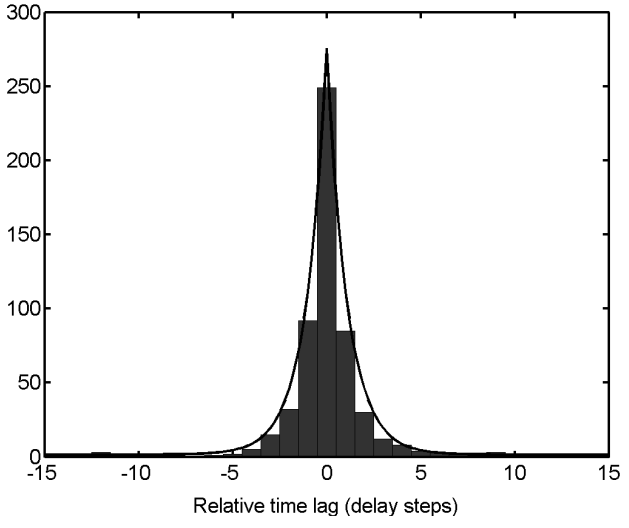


Figure 1. Histogram and estimated distribution of relative time lags in channel 40 (center frequency is 507 Hz) of (a) anechoic speech, and (b) reverberant speech with reverberation time of 0.4 s. The bar graphs represent histograms and the solid lines represent the estimated distributions.

reverberation. More specifically, a relative time lag δ is defined as the distance from the detected pitch period to the closest peak in correlogram. We collect the δ statistics from the selected channels across all voiced frames from anechoic speech utterances randomly chosen from the TIMIT database [9] for every channel separately. As a typical example, the δ histogram for channel 40 is shown in Figure 1a. As can be seen, the distribution is sharply centered at zero with a small spread. This spread, however, is not an artifact due to the inaccuracy of a pitch tracker. A signal composed of an ideal stationary harmonic structure is extremely clean. In this case, the relative time lags collected from the signal have the same value of zero, and the distribution has zero spread. Due to the nonstationary nature of speech, the distribution spread of natural speech is greater than zero (see Figure 1a).

Room reverberation corrupts harmonic structure, and echoes from natural speech tend to spread the distribution of relative time lags. To illustrate this, we collect the statistics of relative time lags from reverberant speech generated by convolving anechoic speech with a room impulse response function generated

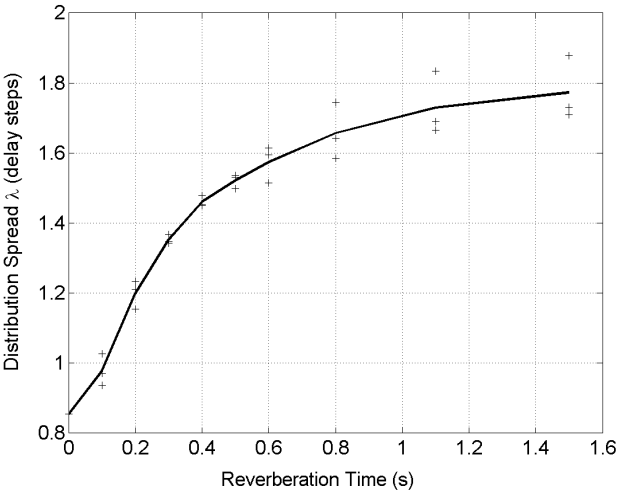


Figure 2. Average distribution spread λ of relative time lag with respect to reverberation time.

from the image model of Allen and Berkley [10] with $T_{60} = 0.4$ s. The histogram is shown in Figure 1b. The spread is wider than that of anechoic speech.

Based on the above observation, we propose to use the spread of the distribution as an indication of reverberation. In order to measure the distribution spread, we employ a mixture of a Laplacian and a uniform distribution to model the distribution in channel c (see [7] for more details):

$$p_c(\delta) = (1-q)\frac{1}{2\lambda_c}\exp\left(-|\delta|/\lambda_c\right) + qU(\delta, \eta_c), \qquad (1)$$

where $0 < q < 1$ is a partition coefficient of the mixture and λ_c is the Laplacian distribution parameter. $U(\delta, \eta_c)$ is a uniform distribution with range η_c. We set the length of the range as the reciprocal of the center frequency. Since the perceptual judgment of pitch depends primarily on the lower harmonics of a complex sound [11], our measure does not consider high-frequency channels.

We also assume a linear relationship between the frequency channel index c and the Laplacian distribution parameter λ_c,

$$\lambda_c = a_0 + a_1 c. \qquad (2)$$

The maximum likelihood method is utilized to estimate the three parameters a_0, a_1, and q in low-frequency channels (channel center frequencies lower than 1275 Hz). The estimated distributions in anechoic and reverberant speech are also shown in Figure 1a and b. As can be seen, the model distributions fit the histograms very well (see [7] for more discussion). With the estimated distributions, the distribution parameter λ_c is 0.7 in Figure 1a and 1.1 in Figure 1b.

Finally, the measure of distribution spread λ is defined as the average of λ_c's in low-frequency channels. Figure 2 shows the relationship of λ and reverberation time. Here, the reverberant signals are generated by convolving a concatenated speech signal from 2 female and 2 male utterances from four different speakers randomly selected from the TIMIT corpus with room impulse responses of various reverberation times obtained from the image model. For each reverberation time, three different impulse responses are generated in order to test the robustness with respect to different room configurations. The spread for each impulse response is indicated by a plus sign in Figure 2. The solid curve in the figure connects the average spreads at different T_{60}'s.

As shown in Figure 2, the distribution spread λ rises monotonically (almost exponentially) with increasing reverberation time. From the monotonic relationship it is straightforward to estimate the reverberation time from λ. In addition, the spread at a particular T_{60} holds relatively steady for different impulse responses when $T_{60} \leq 0.6$ s. This, along with the saturation trend beyond T_{60} of 0.6 s, suggests that our method is not accurate when T_{60} is longer than 0.6 s, and hence should not be used in this situation. The utility of our method for estimating short reverberation time is also supported by the results with more utterances presented in [12] as well as further tests not presented in this letter. Note that the T_{60} range of (0, 0.6 s) covers a wide range of room reverberations including typical living rooms [1].

Why is the proposed method not accurate for long reverberation times? Our algorithm estimates the reverberation time by examining the degree of corruption in harmonic structure. Under a range of reverberant conditions, the distribution spread λ increases as the reverberation becomes more severe. On the other hand, severely corrupted harmonic structure or non-harmonic components provide little information on the pitch, and tend to be excluded by the channel selection method in our pitch tracking algorithm. Long reverberation times cause some harmonic speech components to be severely corrupted, and as a result the distribution spread is not discriminative anymore.

Pitch strength depends on speakers as well as utterances. For example, female speech tends to be more harmonic than male speech, and an utterance with a steady pitch contour (i.e. monotone speech) tends to show higher pitch strength than an utterance with a highly-varying pitch contour. It is expected that the distribution spread has some dependency on the utterances used. This is indeed confirmed by our experiments. On the other hand, for given utterances we always observe a monotonic relationship between the spread and the reverberation time. The dependency can simply be overcome by employing a fixed set of utterances, which unfortunately will compromise the "blindness" of the method. Another way is to use a large number of representative utterances, which will smooth out variations introduced by individual utterances. More utterances entail more computation, which is generally undesirable. Perhaps a more promising approach is to analyze speaker and utterance characteristics, which can then be exploited to predict a precise relationship of the distribution spread to T_{60}. This is an interesting topic for future research.

Besides reverberation time, another important characteristic of reverberation is signal-to-reverberant energy ratio (SRR). It depends on the distance from the source to the microphone and is correlated with reverberant speech quality [13]. In low-frequency channels, however, the relative energy of direct-path signal to early reflections mainly determines relative phase changes and magnitudes of harmonic components and should have little effect on pitch strength. As a result, while low SRR causes coloration distortion in reverberant speech, it is not expected to cause much deviation on our estimation method. This is confirmed in our informal experiments.

To our knowledge, Ratnam *et al.* [4, 5] is the only other reverberation time estimate that utilizes arbitrary speech sources. Their model tends to overestimate the reverberation time due to the gradual offsets of speech sounds. Our method, however, does not have this systemic bias and is effective for a range of reverberation times.

3. Conclusion

We have observed a monotonic relation between room reverberation and a well-established quantity in psychoacoustics – pitch. This relation in turn gives an estimate of the reverberation time utilizing only reverberant speech signal. This estimation method should be useful for many acoustic processing tasks in reverberant environments when prior knowledge of room reverberation is not available.

Acknowledgement

We thank two anonymous reviewers for their helpful comments. This research was supported in part by an AFOSR grant (FA9550-04-1-0117) and an NSF grant (IIS-0081058).

References

[1] H. Kuttruff: Room acoustics. 4th ed. Spon Press, New York, 2000.

[2] T. J. Cox, F. Li, P. Darlington: Extracting room reverberation time from speech using artificial neural networks. J. Audio Eng. Soc. **49** (2001) 219–230.

[3] F. F. Li, T. J. Cox: Speech transmission index from running speech: A neural network approach. J. Acoust. Soc. Am. **113** (2003) 1999–2008.

[4] R. Ratnam et al.: Blind estimation of reverberation time. J. Acoust. Soc. Am. **114** (2003) 2877–2892.

[5] R. Ratnam, D. L. Jones, W. D. O'Brien: Fast algorithm for blind estimation of reverberation time. IEEE Sig. Proc. Lett. **11** (2004) 537–540.

[6] M. Wu, D. L. Wang: A one-microphone algorithm for reverberant speech enhancement. Proceedings of IEEE ICASSP, 2003, Vol. 1, 844–847.

[7] M. Wu, D. L. Wang, G. J. Brown: A multipitch tracking algorithm for noisy speech. IEEE Trans. Speech Audio Process. **11** (2003) 229–241.

[8] R. D. Patterson, J. Holdsworth, I. Nimmo-Smith, P. Rice: SVOS final report, part B: Implementing a gammatone filterbank. Rep. 2341, MRC Applied Psychology Unit, 1988.

[9] J. Garofolo et al.: DARPA TIMIT acoustic-phonetic continuous speech corpus. Technical Report NISTIR 4930, National Institute of Standards and Technology, 1993.

[10] J. B. Allen, D. A. Berkley: Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. **65** (1979) 943–950.

[11] R. J. Ritsma: Frequencies dominant in the perception of the pitch of complex sounds. J. Acoust. Soc. Am. **42** (1967) 191–198.

[12] M. Wu: Pitch tracking and speech enhancement in noisy and reverberant environments. Ph.D. Dissertation, Ohio State University Department of Computer and Information Science, 2003 (available at http://www.cse.ohio-state.edu/pnl/theses.html).

[13] D. A. Berkley, J. B. Allen: Normal listening in typical rooms: The physical and psychophysical correlates of reverberation. – In: Acoustical factors affecting hearing aid performance. G. A. Studebaker, I. Hochberg (eds.). Allyn and Bacon, Needham Heights, MA, 1993, 3–14.