Audio Engineering Society

Convention Paper 9404

Presented at the 139th Convention
2015 October 29–November 1 New York, USA

Effect of Reverberation on Overtone Correlations in Speech and Music

Sarah R. Smith[1], Mark F. Bocko[1]

[1] *University of Rochester, Rochester, NY, 14627, USA*

Correspondence should be addressed to Sarah Smith (sarahsmith@rochester.edu)

ABSTRACT

This paper explores the effect of reverberation on audio signals that possess a harmonically rich overtone spectrum such as speech and many musical instrument sounds. A proposed metric characterizes the degree of reverberation based upon the cross correlation of the instantaneous frequency tracks of the signal overtones. It is found that sounds that exhibit near perfect correlations in an anechoic acoustic environment become less correlated when passed through a reverberant channel. These results are demonstrated for a variety of music and speech tones using both natural recordings and synthetic reverberation. The proposed metric corresponds to the speech transmission index and thus may be employed as a quantitative measure of the amount of reverberation in a recording.

1. INTRODUCTION

In addition to reducing the clarity of audio and speech, the presence of reverberation can be a major compromising factor in the removal of vocal tracks from recorded audio and other source separation tasks. In applications such as video conferencing when the microphone may be at some distance from the person speaking, the transmitted signal may contain a significant reverberant sound field component relative to the direct sound. While a large amount of reverberation may lead to a significant reduction in speech intelligibility, even smaller effects can cause undesirable distortions to the audio. Similarly, music recordings made in a concert hall with a significant amount of reverberation may pose challenges to music information retrieval tasks such as acoustic source separation and automated music transcription. For these and other reasons, there is significant motivation to remove reverberation from recordings prior to transmission or further analysis.

When a measurement of the impulse response of an acoustic space is available and can be identified as minimum phase, it is possible to find an inverse filter that will exactly cancel the reverberation. However, most acoustical systems are not minimum phase and independent impulse response measure-

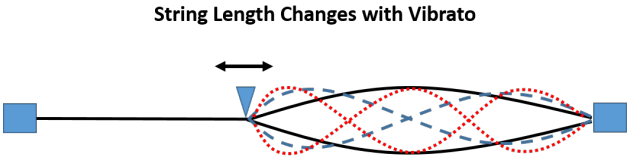

Fig. 1: All modes of a vibrating string are affected by changing the string length

ments of recording environments seldom are available [1]. If the available impulse response is non-minimum phase, it is necessary to identify an inverse filter whose output best approximates a delta function when applied to the room impulse response. While there are many possible statistics that can be used to evaluate the reconstructed impulse, a perceptual motivated metric is generally best. One such metric is the speech transmission index (STI), which was developed to measure an system's ability to preserve the variations of an amplitude modulated signal through the use of a modulation transfer function and has been shown to correspond well with perceptual metrics of speech intelligibility [2]. However, since the STI is designed to measure changes in intelligibility, it is most useful for systems that contain large amounts of reverberant distortion. In this paper, we analyze the effects of reverberation on pitched sounds including spoken vowels and musical instrument tones. The presence of reverberation is found to reduce the mutual correlations among the overtones of these sounds and can be used as a metric for evaluating signal reconstruction in the absence of a measured impulse response.

1.1. Physics of correlated overtones

Many naturally generated sounds consist of a fundamental frequency component and a series of harmonic overtones. These overtones are created when multiple resonant modes of the system generating the sound are excited simultaneously, as is true of most musical instruments and the human voice. For instance, the waveform at the bridge of a bowed string closely resembles a sawtooth wave, and includes frequency components at integer multiples of the fundamental [3]. Similarly, the glottal waveform of spoken vowels is highly periodic and contains

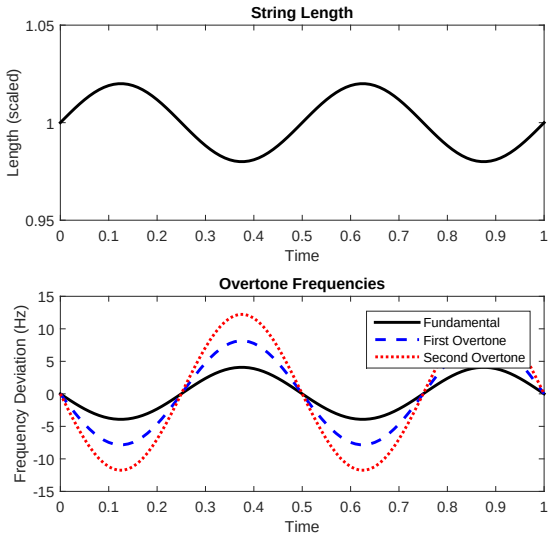

Fig. 2: Overtones should be correlated in string vibrato

strong overtones [4].

However, these systems are often modulated in semantically important ways. Amplitude and frequency fluctuations in speech convey important information about the context or emotion of the speaker and musicians will often introduce slight modulations into their tone for expressive effect. Acoustically, these fluctuations result from a modulation of system parameters such as the length of the string or tension of the vocal cords. Since these changes affect the whole acoustic system uniformly, it is expected that all of the resulting overtones should be similarly altered. In the case of a bowed string, when the length of the string is adjusted, the frequencies of all the modes should change accordingly. This effect is illustrated in Figs 1-2. As the length of the string is decreased, the frequencies of all of the modes increase proportionally. This expectation should also extend to the mechanisms of woodwind and vocal music: alterations that modulate the frequency of a fundamental pitch will alter all modes of the acoustic system.

For the purposes of this analysis, the input signal is modeled as the sum of many sinusoidal components, each of which may be independently mod-

ulated (Eq. 1). However, in order to isolate the frequency deviations of each overtone, it is helpful to separate out the constant frequency components from each overtone, as in Eq. 2. In this model, f_0 corresponds to the average fundamental frequency (in Hz) of the tone, around which the modulation occurs. For a perfectly harmonic tone, the instantaneous phase $\theta_n(t)$ of each harmonic should resemble a scaled version of the phase of the first harmonic (i.e. $\theta_n(t) = n\theta_1(t)$). Since the harmonic scale factor n is included in the first term of Eq. 2, this is equivalent to the case where $\phi_n(t) = n\phi_1(t)$.

$$x(t) = \sum_{n=1}^{n=\infty} a_n(t)\cos(\theta_n(t)) \quad (1)$$

$$x(t) = \sum_{n=1}^{n=\infty} a_n(t)\cos(2\pi n f_0 t + \phi_n(t)) \quad (2)$$

From here, the instantaneous frequency of an overtone can be defined as the time derivative of the instantaneous phase, $f_n(t) = \frac{d}{dt}\theta_n(t) = 2\pi n f_0 t + \frac{d}{dt}\phi_n(t)$. It should be noted that for a perfectly harmonic sound, both the instantaneous phases and frequencies should be perfectly correlated. However, for sounds with slightly non harmonic overtones, a constant offset in the frequency of an overtone will result in a linearly increasing phase in our model, and care should be taken to remove such offsets when considering an instantaneous phase trajectory.

2. SIGNAL ANALYSIS

Although many different methods exist to track the instantaneous frequency of a sound and its overtones, the desire to track many overtones imposes a few restrictions. Most importantly, the magnitude of the spectrum of many sounds drops off at high frequencies and it is not uncommon to encounter overtones whose amplitudes are only a few decibels (dB) above the surrounding noise floor. In order to accurately track 25 to 30 overtones of the fundamental, it is important to choose an algorithm that is robust in conditions with low signal to noise ratio (SNR). For example, the overtone amplitudes of a typical cello tone recorded in an anechoic chamber are shown in Fig. 3. In this case the upper overtones have amplitudes roughly 50 dB below the fundamental, rapidly approaching the noise floor of even a quiet room. Secondly, the frequency deviations of higher overtones often become large compared to the fundamental frequency, making it impractical to isolate individual harmonics using a bandpass filter bank. These considerations motivate the use of an algorithm based on spectrogram peak tracking, which is very robust to noise and does not require individual harmonics to be isolated prior to processing [7]. However, the addition of additional analysis steps before and after the spectrogram tracker can improve the accuracy and robustness of the initial results. The implementation described below was calibrated using synthetic tones and additive noise and was found to perform well even in as little as 0 dB SNR.

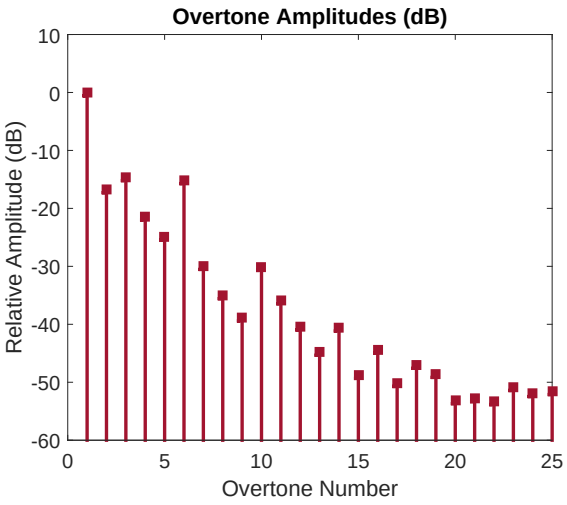

Fig. 3: Overtone spectrum of a cello tone

2.1. Frequency Tracking

In light of the tracking considerations mentioned above, a three stage algorithm is used to calculate the instantaneous frequency trajectories $f_n(t)$. In the first stage, an overall pitch trajectory is calculated using the YIN algorithm [8]. This provides a reasonable estimate of $f_1(t)$ that serves as an initial guide for later tracking refinements and identifies the amount of frequency deviation present in the sound. In the case of musical vibrato, a good estimate of the width and rate of vibrato is also obtained. The second stage uses a spectrogram based peak tracker to capture any deviation from harmonicity [7]. Using

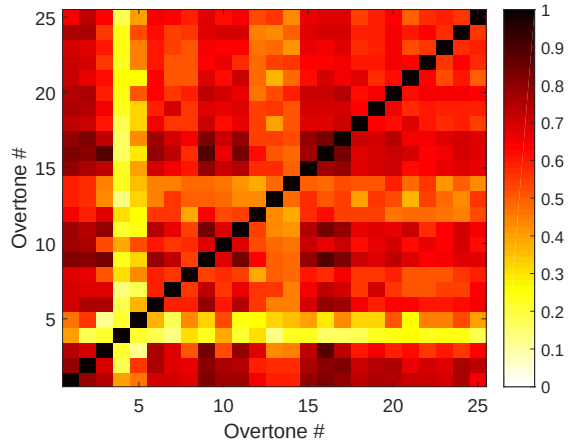

Fig. 4: Correlation matix

the spectrogram magnitudes, the algorithm identifies all of the local peaks in each frame and attempts to assign them to one of the overtone trajectories. In cases where a local peak in the spectrogram is not found for a given overtone, the YIN pitch estimate for the given frame is scaled by the overtone number and used instead, ensuring the best possible estimate for use in the final correction stage. Since this stage uses only the spectrogram magnitudes, it is possible to locate and track very low amplitude peaks with minimal interference from the surrounding noise.

In order to further refine the accuracy of the frequency estimates, the final analysis stage uses the spectrogram phase information to adjust the previous frequency estimates in each frame. For each overtone and frame, the discrete time Fourier transform (DTFT) is evaluated at the frequency estimated in stage 2 (f_{eval}) for both the current frame m and the following frame [9]. The A final instantaneous frequency estimate can then be found using the phases of the corresponding DTFT components (here labeled ϕ_m and ϕ_{m+1}) and the hop time t_{hop} according to Eq. 3.

$$f_{est} = f_{eval} + \frac{\phi(m+1) - \phi(m)}{2\pi t_{hop}} \tag{3}$$

However, since the DTFT phase is easily influenced by surrounding noise, the accuracy is improved when the stage 2 estimate is as close as possible to the correct frequency. For this reason, it is important to use the frequencies from the peak tracking analysis as opposed to performing this correction based on the phases of the nearest spectrogram bin.

2.2. Correlation Calculation

Once the instantaneous frequency trajectories $f_n(t)$ are calculated, it is possible to calculate a correlation matrix consisting of the pairwise correlation coefficients for each pair of overtones. For each pair of overtones, the correlation coefficient is calculated from the covariance as follows:

$$\rho_{i,j} = \frac{cov(f_i(t), f_j(t))}{\sigma_{f_i}\sigma_{f_j}} \tag{4}$$

The resulting values $\rho_{i,j}$ will always be between -1 and 1, with a value of 1 corresponding to overtones whose trajectories are perfectly correlated, and a value of zero corresponding to two independent processes. This set of correlations can be represented as a square matrix and visualized as in Fig. 4. Since the diagonal elements of the matrix represent the correlation of an overtone with itself, these elements are identically unity. For each tone, an overall correlation value is obtained by averaging the off diagonal elements of the correlation matrix. Although this work analyzes the correlation of the instantaneous frequencies, it is also possible to perform a similar analysis of the instantaneous amplitudes or phases, as has been done previously [6].

3. INVESTIGATION OF VIBRATO

In order to determine if reverberation has an effect on these correlations, a selection of cello and bassoon tones was recorded in both an anechoic chamber and a recording studio. In the studio, recordings were made using two different microphones placed at different distances from the instrument. The first microphone was placed close to the instrument and captured mostly the direct signal while a second, more distant, microphone recorded more of the room reverberation. This setup allows for the comparison of three different conditions for each tone (anechoic, close mic, and far mic). As expected, the anechoic recordings give evidence of highly correlated overtones. This suggests that the instruments themselves exhibit the harmonicity that is expected from

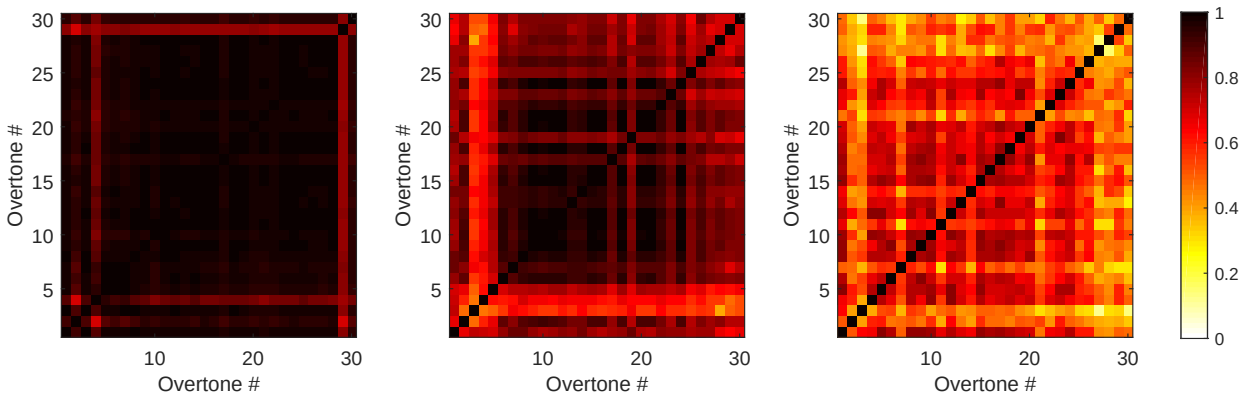

Fig. 5: Correlation is reduced in the presence of reverberation

the earlier models. However, when the same notes are recorded in a reverberant environment, these correlations are diminished. An example of this relationship is shown in Fig 5 for a cello playing vibrato on the note F4 (350 Hz). While this reduction in correlation is present even in the close microphone recording, it is more noticeable as the microphone is moved further from the instrument.

The initial observations based on the cello tone are confirmed when a larger database of tones is analyzed. The recording database includes 60 notes (18 cello and 42 bassoon) recorded in each of the three conditions. A summary of these results is presented in table. 1. There is a consistent reduction in internal correlation as the amount of reverberation in the recording is increased. While the effects of reverberation are consistent across the database, the cello tends to exhibit lower correlations than the bassoon, even in the anechoic environment. Although this initially seems inconsistent with the earlier instrument model, it could potentially be explained by the resonances of the cello body. Unlike the bassoon, the body of the cello acts as an external resonator to enhance the sound. Therefore, it is not necessarily surprising that the sound reflections inside the cello's body could produce a similar decorrelation as those in the room. This result also agrees with earlier vibrato analysis by both Brown and Dubnov [5][6].

Instrument	Anechoic	Close Mic	Far Mic
Bassoon	0.8131	0.7486	0.5519
Cello	0.6829	0.6160	0.4648

Table 1: Average overtone correlation vs instrument and acoustic condition

4. SPEECH

Although musical signals provide a controlled set of modulated signals to test various rooms, the negative effects of reverberation are more apparent in speech applications. For this reason, the previous analysis of frequency correlations has been extended to include spoken vowels. Similar to musical tones, spoken vowels are strongly pitched and contain semantically important inflections. In fact, the frequency modulation in speech is often wider, faster, and less regular than that found in music. Unlike the musical tones which were recorded in multiple acoustic settings, the speech samples were all taken from a single anechoic recording. The original anechoic recording was processed with a Schroeder reverberator and both versions were combined with additive white noise. In contrast to a live recording setting, the use of synthetic manipulation allows the effects of reverberation and noise to be studied independently. Additionally, since the original speech is the same for all four cases, this procedure eliminates variations that might occur between takes in a studio. As is the case for the musical tones,

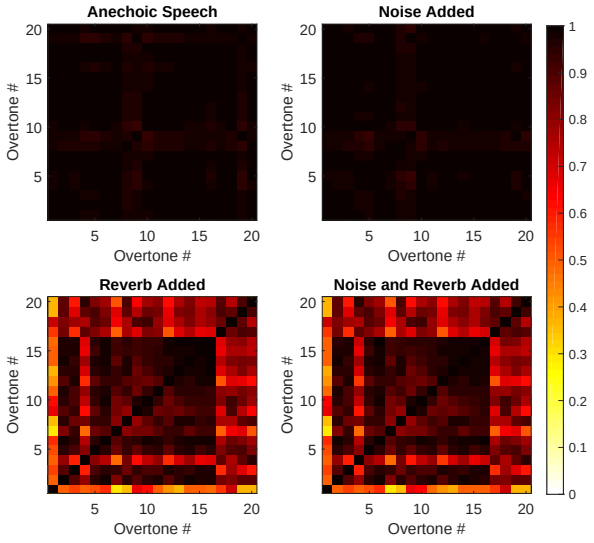

Fig. 6: A reduction in correlation is observed for reverberation but not additive noise

the anechoic speech displays a high degree of correlation among the overtone modulations. However, the correlations are reduced when the reverberation is added, as shown in the lower portion of Fig. 6. Interestingly, the addition of white noise does not produce the same effect, even at significant audible levels (20 dB SNR). This data, combined with our initial calibration of the tracking algorithm suggests that the observed reduction in correlation is in fact due to the reverberation, and not to other sources of distortion.

5. COMPARISON TO SPEECH TRANSMISSION INDEX

While the presence of reverberation seems to consistently diminish the overtone correlations of many sounds, the previous results do not provide insight into the perceptual relevance of the correlation metric. In order to evaluate the perceptual relevance of this metric, it was compared with the speech transmission index using a set of simple filters [2]. These tests were performed on a variety of synthetic vibrato tones processed with a one pole filter of varying feedback gain. For each case, the speech transmission index can be calculated from the impulse response of the specified filter. The results of this analysis are plotted in Fig. 7. While both the overtone correlation and STI decrease monotonically as the filter gain is increased, the correlation metric is more sensitive to small amounts of distortion and decreases quite rapidly for higher filter gain. This makes sense since the STI is designed to measure the reduction in comprehension of words in speech. Therefore it is less sensitive to interference that causes only minor artifacts while the overtone correlation is more sensitive to these effects. When comparing these two measures of reverberant distortion, it is important to recognize that they represent two distinct measurements. While the speech transmission index is a characteristic of the room or acoustic system, the overtone correlation is derived only from the recorded signal. Their correspondence therefore highlights the effect that rooms have on the signal itself and suggests that overtone correlations may provide a useful measure of reverberant interference in applications where a measurement of the impulse response is unavailable or not feasible.

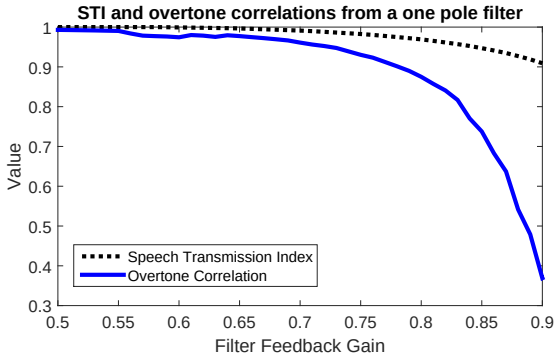

Fig. 7: Comparison of overtone correlation and speech transmission index

6. DISCUSSION

Based on the analysis of both speech and music samples recorded in multiple settings, the presence of reverberation in a sound can significantly affect the resulting instantaneous overtones of the sound, causing them to become less correlated with one another. When applied to vibrato analysis, these results emphasize the importance of the recording environ-

ment. Clearly, it is not generally feasible to record in a completely anechoic environment, and thus the impacts of reverberation must be considered when analyzing the recorded data. The same changes are also found for the frequency tracks of spoken vowels. Although this analysis was performed on recordings of isolated notes and vowels, it should be possible to extend the principles to examples with multiple sound sources. In these applications, it would be necessary to adapt the tracking algorithms for multiple sources. This could be accomplished by substituting a multi-pitch tracker in place of the YIN algorithm in order to initialize all of the tracks. Furthermore, the correlation among the overtones in a harmonic sound provides a potential metric to evaluate the amount of reverberation present in a recording. Depending on the application, it may be desirable to remove the reverberant effects prior to processing, in which case the correlation value could be used as a parameter for optimization in a dereverberation algorithm.

7. REFERENCES

[1] Miyoshi, Masato, and Yutaka Kaneda. "Inverse filtering of room acoustics." *Acoustics, Speech and Signal Processing, IEEE Transactions on* 36.2 (1988): 145-152.

[2] Houtgast, T. A. M. M. O., H. J. M. Steeneken, and R. Plomp. "Predicting speech intelligibility in rooms from the modulation transfer function. I. General room acoustics." *Acta Acustica united with Acustica* 46.1 (1980): 60-72.

[3] Fletcher, Neville Horner, and Thomas D. Rossing. *The physics of musical instruments*. Springer Science & Business Media, 1998.

[4] Sundberg, Johan. *The Science of the Singing Voice*. Northern Illinois University Press, 1989.

[5] Brown, Judith C. "Frequency ratios of spectral components of musical sounds." *The Journal of the Acoustical Society of America* 99.2 (1996): 1210-1218.

[6] Dubnov, Shlomo, and Xavier Rodet. "Investigation of phase coupling phenomena in sustained portion of musical instruments sound." *The Journal of the Acoustical Society of America* 113.1 (2003): 348-359.

[7] McAulay, Robert, and Thomas F. Quatieri. "Speech analysis/synthesis based on a sinusoidal representation." *Acoustics, Speech and Signal Processing, IEEE Transactions on* 34.4 (1986): 744-754.

[8] De Cheveign, Alain, and Hideki Kawahara. "YIN, a fundamental frequency estimator for speech and music." *The Journal of the Acoustical Society of America* 111.4 (2002): 1917-1930.

[9] Brown, Judith C., and Miller S. Puckette. "A high resolution fundamental frequency determination based on phase changes of the Fourier transform." *The Journal of the Acoustical Society of America* 94.2 (1993): 662-667.