REAL-TIME HIGH-RESOLUTION DELAY ESTIMATION IN AUDIO COMMUNICATION USING INAUDIBLE PILOT SIGNALS

Vaggelis Alexiou and Alexandros Eleftheriadis

National and Kapodistrian University of Athens
Department of Informatics and Telecommunications
University of Athens, Ilisia, 15784, Athens, Greece

ABSTRACT

We describe the construction of inaudible pilot signals and associated receiver processing techniques that can be used in audio communication, so that an accurate estimation of the signal's end-to-end delay can be obtained. Considering the properties of the human auditory system, we structure the pilot signal so that it is acoustically untraceable and maximizes the time accuracy of the delay estimation. We provide simulation results verifying the effectiveness of the techniques, and also demonstrate its superiority compared to systems with widely used pulses in similar schemes.

1. INTRODUCTION

In many cases of network communication, it is desirable to measure the end-to-end delay between two remote computer systems. In the case of real-time audiovisual communication, it is useful to be able to calculate the delay through the entire system, including digital-to-analog conversion and playback at the receiver, in order to obtain the real delay experienced by an end-user. There are applications where high-resolution delay estimation is essential. One example is Network Music Performance (NMP) systems, where the end-to-end delay tolerance is reported to be in the order of 20 msec. With such stringent requirements, it is important to be capable of quantifying with high accuracy the delay across the entire path from capture, to encoding and transmission, decoding, as well as playback. In addition to performance analysis, the exact value of the delay that each participant is subject to can be used by the system in order to improve his/her experience. One example is tempo adaptation [1]. In various experiments that have been conducted, a general inverse relationship between tempo and delay was reported. For instance, slowing down the tempo of a particular musician who experiences a bigger delay than the rest of his group, can be beneficial and can enhance an ensemble's ability to play synchronously and with comfort.

We describe a novel system that can be used in applications where audio data is exchanged between two or more end points, such as in an NMP. The proposed system is a high-resolution delay estimator that involves inaudible pilot signals. The objective is to measure the end-to-end delay in transmitting audio between two systems. The structure provides "orthogonality" in that it allows multiple independent measurements between a number of users across a network. A pilot signal is added to the audio during capture (e.g., at the computer used as a capture and transmitting device). At the receiver, the received "complex" signal is processed in order to obtain a high-resolution (in the order of one sample) estimate of the position of the pilot signal.

Assuming that a second time reference is provided, this can be used to obtain a highly accurate delay estimate. For example, in a laboratory setting, the audio signal together with the pilot may be captured without being transmitted, so that it can be used as a reference. Monophonic outputs of the audio signal plus the pilot from the transmitting and receiving devices can be recorded as a stereo pair. The two channels can then be processed, independently, to measure the exact delay. The proposed technique is robust to a number of signal processing operations. Embedding an inaudible pilot signal is similar to audio watermarking [2, 3]; the objective of identifying the position of the watermark, rather than its content, is, however, different.

This system can also be used in order to observe dynamically varying time-related parameters such as network jitter. This can be accomplished using periodic sequential transmission of the pilot signal. A benefit of the proposed technique is that it can measure computer hardware and operating system latency (playout buffer management delay, etc.). The proposed system has been designed as a performance analysis tool in an NMP system under development ("MusiNet").

2. SYSTEM MODEL

Fig. 1 shows the different processing steps for pilot injection and analysis. At the transmitter, the captured audio signal m is mixed with the pilot signal s. The exact time for the "injection" is not relevant. The mixed signal is potentially encoded, and transmitted to the receiver. We make no assumptions about the delivery channel. While a typical path would be packet-based digital distribution (e.g., over the Internet), the proposed scheme works for both analog as well as non-packetized communication. At the receiver, the mixed signal is optionally decoded. If we do not want to measure delays associated with receiver processing such as playout buffer management and operating system delays, the delay analysis can be performed at this stage ("channel delay estimation"). If, however, we want to include such delays in the measurement, then the endpoint will have to play the received audio, and we will need another device to capture the played back, mixed audio signal. The analysis of this signal will indicate the precise position where the pilot was injected ("total delay estimation"). Note that this position does not necessarily relate to any particular clock, except that of the device where the analysis is performed.

In order to measure the delay, we need to obtain a time reference. Fig. 2 depicts a diagram of an example delay measurement configuration such as the one that would be used in laboratory-based measurements. The pilot signal is transmitted mixed with the audio/music signal. For our purposes, the music signal is treated as

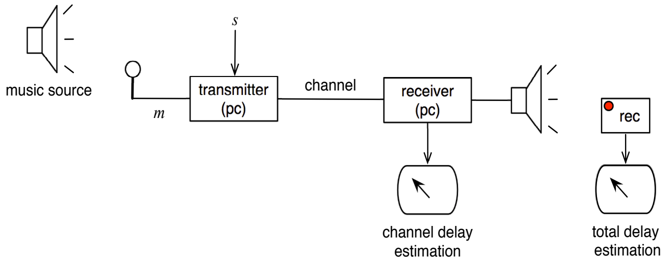

Fig. 1. System diagram.

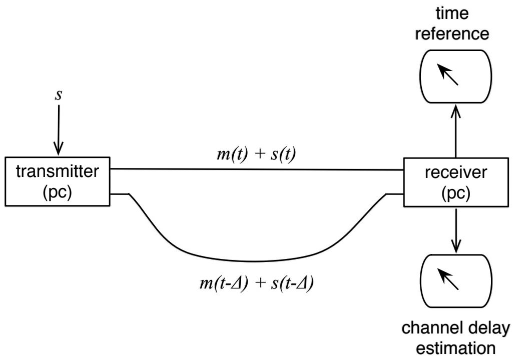

Fig. 2. Example configuration for laboratory-based delay measurement.

"noise". The mixed signal is transmitted over a delay-prone path to a receiver. In order to create a reference, we can assume that the mixed signal without delay is also made available to the receiver. This can be accomplished in a laboratory setup by having a direct, analog or digital audio path between the sender and the receiver. The two signals - with, and without delay - can be captured as a stereo signal at the receiver. Note that the capture may be performed in the analog domain if we wish to include in the delay measurement the delays associated with receiver processing (operating system, play-out buffer management, D/A conversion).

The construction of the pilot signal must be carefully considered, in order to ensure that it both inaudible and that it allows high-resolution delay estimation. Furthermore, assuming that the pilot signal is added in the digital domain, its bandwidth is limited to that provided by the underlying digital representation. In this study we assume a sampling frequency of 44.1 kHz, and thus the pilot signal is limited to 22.05 kHz.

2.1. Pulse Design

A pilot signal with its main power located in a band where human sensitivity to sounds is quite low should be used. In Fig. 3 we show the human Absolute Threshold of Hearing (ATH) as a function of frequency based on the well-known Fletcher-Munson equal loudness contours. Human hearing is found to be insensitive at very low ($<$ 100 Hz) and high frequencies ($>$ 12 kHz). Representative music samples of instruments that are generally used in orchestras were used in order to observe the general properties of the music spectrum [4]. Although there is no mathematical model that can provide a concise and general stochastic description of music, the general tendency of exhibiting lower power in high frequencies was expected and observed in all tested music samples.

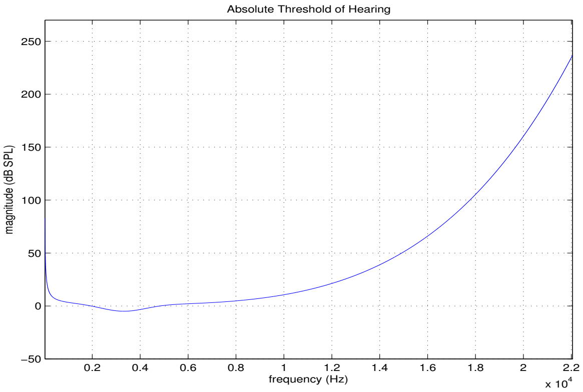

Fig. 3. Absolute threshold of hearing.

It is best, therefore, to limit the pilot signal's spectrum in a high-frequency band, to ensure better performance and robustness. These properties can be achieved from the choice of the used pulse, by including a high-pass filter in the system architecture, or by modulation of band-limited pulses. In this work the first option is followed. The high resolution requirement translates to the need for short pulses which, in turn, implies the use of pulses that have as wide bandwidth as possible (within the limits of the underlying digital representation).

After extensive experimentation on various candidate pulses, we concluded that the use of the 15th derivative of a Gaussian pulse is an appropriate choice. The main advantage of using derivatives of Gaussian pulses is that their spectrum is bandlimited (resembles the frequency response of a bandpass filter) with the main spectrum shifted towards higher frequencies as the order of the derivative increases. A typical implementation should involve pulse durations of 0.30 msec, which has a sufficient resolution in terms of samples (13 samples) while the spectrum is broad enough.

We selected this specific pulse using the following process. The duration of each candidate pulse was kept fixed (0.3 msec) and the corresponding spectrum was derived. For the aforementioned pulse width, the spectrum of 15th derivative of Gaussian pulse was observed to be located in the highest frequencies, without exceeding the available bandwidth. Finally, it was found that each pulse should be normalized with energy $\leq 4 \cdot 10^{-13}$ in order to be totally inaudible. In Fig. 4 we show the selected pulse in both time and frequency domains.

2.2. Modulation Technique

A spread spectrum modulation technique was chosen in order to extend the pilot signal's bandwidth over the entire available band. Spread spectrum techniques combine a pseudo-random nature as well as wide bandwidth. The methods that are primarily used are Time Hopping (TH) and Direct Sequence (DS). They offer resistance to interference, hide the signal from reception, and improve multiple access capability. The two last properties are particularly useful in our case.

We chose DS modulation over TH for the following reasons. Firstly, the minimum duration of pilot signal is desired, in order to minimize the probability of: (a) being perceived by human, and (b)

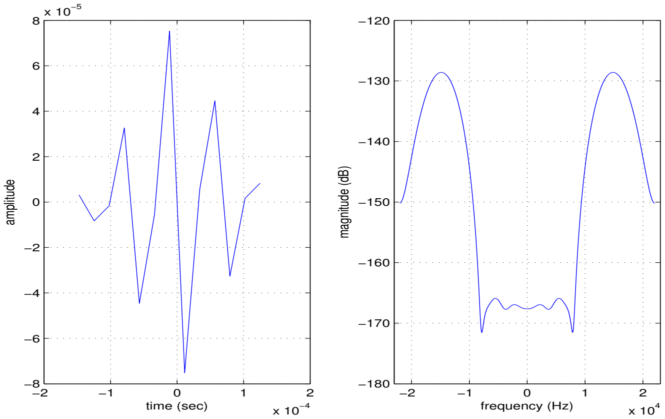

Fig. 4. Using pulse in time and frequency domain.

receiving a part of transmitting signal due to packet loss or jitter. As stated in Section 2.4, a cross-correlation technique is applied at the receiver and the delay estimation is based on detecting the peak at the output. Thus, the maximum possible signal energy is desirable. In addition, as previously mentioned, we have a practical constraint of energy per pulse, to ensure that it remains inaudible. We therefore need the maximum possible number of pulses, given a fixed value of signal's duration. This requirement is met with the DS method.

The general structure of the DS modulation technique is based on the application of a spreading code to each transmitted bit pulse. The waveform for a set of transmitted information bits b_i is:

$$b(t) = \sum_{i=0}^{N_b-1} b_i \, p_b(t - iT_b) \tag{1}$$

where $b_i \in \{-1, +1\}$ corresponding to bit 0 and 1, respectively, N_b denotes the number of transmitted bits, T_b is the bit duration, and $p_b(t)$ is the pulse used to represent each bit. To apply a spreading code to each bit, split the bit duration to N_c non-overlapping intervals called chips. Each chip has a duration equal to $T_c = T_b/N_c$. N_c, the processing gain, is assumed to be much larger than one. The spreading waveform for a single bit will be given by:

$$p_b(t) = \sum_{j=0}^{N_c-1} c_j \, p_c(t - jT_c) \tag{2}$$

where c_j is the spreading code and its value is selected equiprobably between ± 1. $p_c(t)$ is the chip pulse shape. The bandwidth of the spreading waveform is much greater than the bandwidth of the information signal and the spectral characteristics of the transmit signal are dominated by the spreading signal. Specifically, the chip rate and chip pulse shape along with the autocorrelation properties of the spreading sequence will determine the transmit signals spectral properties.

With the above analysis, the transmitted pilot signal can be written:

$$s(t) = \sum_{i=0}^{N_b-1} b_i \sum_{j=0}^{N_c-1} c_j \; p_c(t - iT_b - jT_c) \tag{3}$$

Summarizing, the total time duration of the pilot signal is defined as $T = N_b \cdot T_b$ and in each bit duration, T_b, the chosen pulses are placed sequentially over $N_c = T_b/T_c$ non-overlapping chips. The sign of each pulse placed in the pilot is defined by the DS code c_j in combination with the corresponding bit value b_i.

In our system, there is no need to transmit information bits. What is of interest is the exact temporal position of the information, and not its content. In fact, the information is already known at the receiver. This is obviously very different from a typical communication system, where the goal is to exchange information between the two ends. In Eq. 3, bits are selected in order to eliminate spectral lines that would be present in frequency domain, spaced by $1/T_b$. An equivalent system can be obtained if we assume that our total pilot duration corresponds to one bit duration (i.e., $T = T_b$, $N_b = 1$, and $b_0 = 1$). In this case, our system is a corner case of equivalent to a 2-PAM system, without using any carrier.

2.3. Power Spectral Density

The Power Spectral Density (PSD) of the transmitted signal, based on Eq. 3, is calculated in [5]. Our pilot is assumed as a cyclostationary process and, thus, its PSD is given by:

$$S_s(f) = \frac{\sigma_b^2 \cdot \sigma_c^2}{T_c} \left|P_c(f)\right|^2 = \frac{1}{T_c} \left|P_c(f)\right|^2 \tag{4}$$

2.4. Channel estimation

We assume that, as the signal propagates trough the system, it is subject to scaling, delay, as well as the addition of noise. The signal at the receiver is thus given by:

$$r(t) = a\, s(t - \Delta) \, + \, a\, m(t - \Delta) \, + \, n(t) \tag{5}$$

where a is a scalar, $n(t)$ is AWGN, and $m(t)$ is the actual audio/music signal.

The proposed delay estimation algorithm calculates the cross-correlation between the received waveform and the pilot signal, which is known at the receiver. Noting that $s(t)$ is a real signal, we have:

$$R_{rs}(t) = r(t) \star s(t) = \int_{-\infty}^{+\infty} r(\tau + t) s(\tau) d\tau$$

where $\star$ denotes cross-correlation. Equivalently, assuming that $a = 1$, we get the correlations of the pilot with itself and both types of "noise" (real noise and the audio/music signal):

$$R_{rs}(t) = R_{ss}(t - \Delta) + R_{ms}(t - \Delta) + R_{ns}(t) \tag{6}$$

where $R_{ss}(\tau)$ denotes the auto-correlation function of pilot signal. $R_{ms}(\tau)$ and $R_{ns}(\tau)$ indicate the cross-correlation between music and noise products with the pilot, respectively.

3. EXPERIMENTAL RESULTS

We define $SNR = P_s/(P_n + P_m)$ where P_s, P_n, P_m is the power of signal, white noise and audio/music respectively. Note that, in our application, both the actual noise and the audio/music signal are considered noise, since the actual information we are interested in is the pilot signal. Each of these values can be calculated as the integral of the corresponding signal's PSD. Given that the power of the pilot signal is known, the evaluation of interfering music track's power can be calculated directly by setting the SNR's value.

In Fig. 5 we compare the performance of our system for different selections of the pilot's pulse. We compare Dirac, rectangular, Gaussian, and the 15th derivative of the Gaussian. The effectiveness

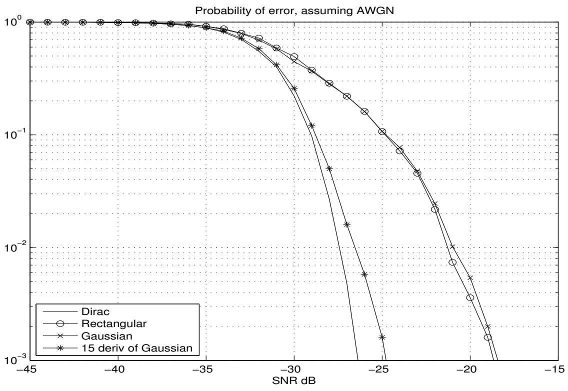

Fig. 5. AWGN.

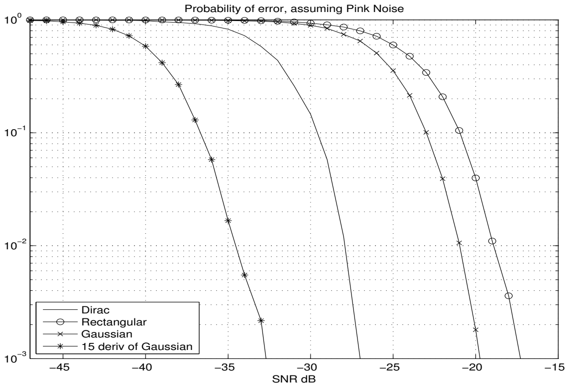

Fig. 6. Pink noise.

of the system is measured using the probability of error. In general, the Mean Squared Error (MSE) is a better measure for the evaluation of an estimator. However, what is of interest in our case, is the assessment of the system performance, requiring accuracy in the order of one sample. Hence, a binary decision for the delay estimation is considered: if the delay estimation equals the true delay of channel, no error occurs. The probability of error is calculated as the ratio of the total number of errors divided by the number of repetitions, for each value of SNR.

The best performance under AWGN is expected to be achieved with Dirac pulses, which indeed is the case. We observe that the proposed pulse achieves very similar performance to Dirac.

Pink noise, where each octave contains the same amount of energy, can be a useful model for rich musical sounds. In Fig. 6 we analyze the performance of different pulses under pink noise. In this case the 15th derivative of Gaussian pulse performs much better than the other pulses. The reason for the substantial difference in performance is that, by design, the spectrum of our pulse is positioned in frequencies where the pink noise's power is decreased. Consequently, a significant increase in SNR at the output of the cross-correlation is observed, in contrast to the other candidate pulses.

Finally, the performance of the system with various music samples is shown in Fig. 7. The selected samples are based on typical orchestra instrumentation involving baritone, bass, cello, clarinet, flute, horn, snare, trombone, trumpet, tuba, viola and violin tracks.

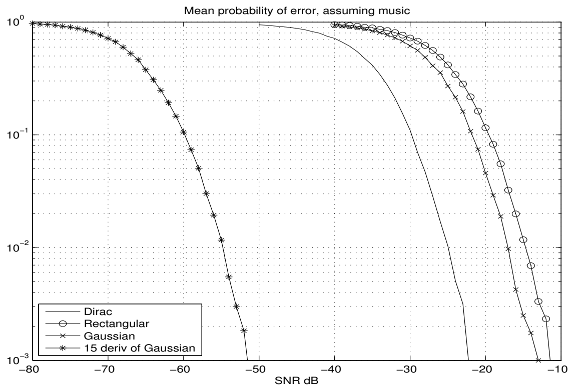

Fig. 7. Music.

Given the pulse that is used, it is obvious that different behavior is obtained for different music samples, as the corresponding PSD differs. Hence, we calculated the average probability of error over all chosen music samples, for each compared system. The results of Fig. 7 show that the proposed pulse outperforms all the rest by a significant margin. It is important to note that, in addition, by setting the specified value for energy per pulse, all the candidate pilot signals were audible except of the proposed.

4. CONCLUDING REMARKS

We have presented a system for accurately measuring the delay in the transmission of an audio signal by injecting an inaudible pilot. We show how to select an appropriate pilot pulse and use spread-spectrum techniques to both make the pilot inaudible as well as increase its ability to be detected at the receiver. Comparison with other candidate pulses such as Dirac, rectangular, and Gaussian, has demonstrated that the proposed 15th derivative of the Gaussian performs as good as Dirac for the AWGN case, and outperforms all other candidates in pink noise and real music cases.

Acknowledgments

This research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program “Education and Lifelong Learning” of the National Strategic Reference Framework (NSRF) - Research Funding Program: THALIS - University of Crete - MUSINET.

5. REFERENCES

[1] A. Barbosa. *“Displaced Soundscapes”*. PhD thesis, Pompeu Fabra University, Barcelona, Spain, 2006.

[2] W. Bender, D. Gruhl, N. Morimoto, and A. Lu. “Techniques for data hiding”. *IBM System Journal*, 35:313–336, 1996.

[3] G. B. Khatri and D. S. Chaudhari. “Digital audio watermarking applications and techniques”. *International Journal of Electronics and Communication Engineering and Technology (ICEJET)*, 4(2):109–115, March-April 2013.

[4] M. Martinson and B. Martinson. http://www.beginband.com.

[5] J. G. Proakis and M. Salehi. *“Communication Systems Engineering”*. Prentice Halls, 2002.