

Room fingerprinting - by Stephen Tarzia

From Casa

Room fingerprinting is a phrase I invented to describe a combination of elements of *room response measurement* and *audio fingerprinting*. The problem is to determine in which of a given set of rooms (i.e. acoustic environments) a recording was made. This is an example of the type of problem that I am interested in solving.

Actually, this survey covers more than just room fingerprinting. The full topic is Laptop-computer Acoustic Sensing (LAS). In the first section, LAS is introduced along with my motivation for studying it. Next, several types of acoustic sensing systems are described followed by a survey of the current state-of-the-art in mobile acoustic sensing systems. Finally, I conclude by speculating about what new LAS systems might be built based on the described systems.

Contents

- 1 Motivation
- 2 Acoustic sensing systems overview
- 3 Active sensing
 - 3.1 Sonar
 - 3.2 Room Response
- 4 Passive sensing
 - 4.1 Audio Fingerprinting
 - 4.2 Other passive systems
- 5 Acoustics in mobile computing
- 6 Conclusions and Future Work

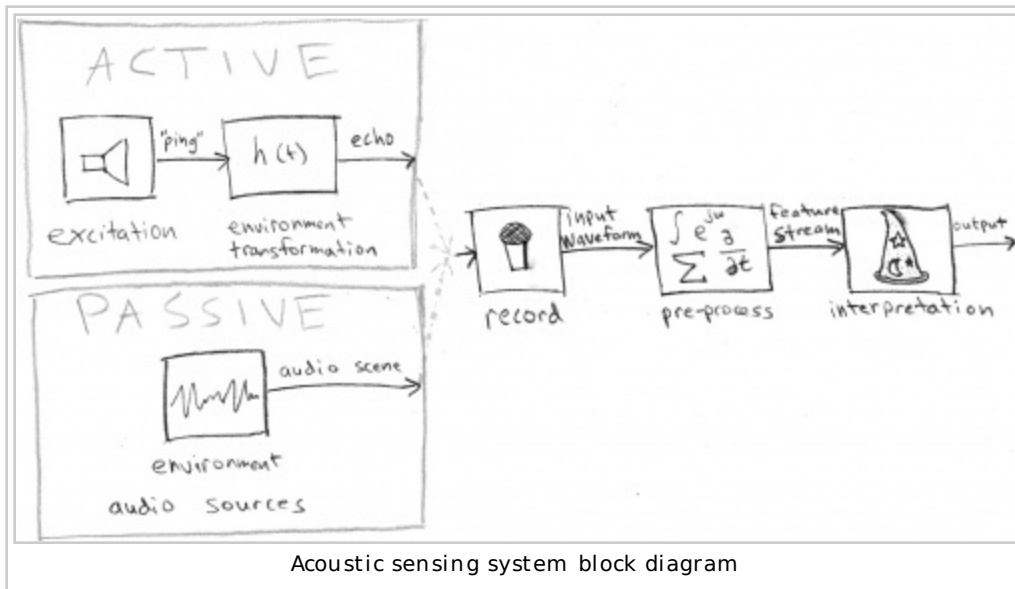
Motivation

This study is motivated by an ongoing research project in which I have attempted to incorporate new kinds of information on the computer user's state in various Operating System policy decisions. For example, I am currently working on a power-management system which can sleep the display immediately after the user leaves the computer. So far, I have been able to distinguish between two sets of audio recordings made on a laptop computer; in one set, a user is sitting in front of the computer watching the screen and in the other set there is no user present. More generally, I am interested in discovering what characteristics of the computer's physical environment can be discerned using audio recordings from its built-in microphone.

The goal is to discover the what types of computational acoustic sensing systems, if any, can be built using a typical laptop computer's built-in audio hardware. To put it another way, I am asking *if we let the computer eavesdrop on you, what can it find out?* I call this Laptop Acoustic Sensing (LAS) and it seems to me that very little has been done addressing this problem. The limitations of that audio hardware will define this work's scope together with the broad directive to improve the user's computing experience. Since LAS is a new research problem, no prior publications directly address it. This literature survey will highlight relevant aspects of several lines of acoustic sensing research.

Acoustic sensing systems overview

Generally speaking, there are two types of acoustic sensing systems: passive and active. In active systems, audio is emitted from the system and the echos are recorded and processed. Passive systems do not emit any sound, but instead process sounds produced by the environment:



Examples of each type of system are:

system	type	transducers	acoustic features	output/purpose
Biosonar	active (usually)	vocal/nasal organs, ears	delay, FM	ranging and mapping, target identification
Artificial sonar	active	magnetic speaker/horn, directional microphone	delay, FM	ranging and mapping
Room acoustic response	active	speakers, microphones	impulse response	musical performance reproduction, music amplification calibration
Ultrasonic rangars	active	speaker, microphone	delay	distance; used in robotics, machinery
Human auditory system	passive	ears	spectrum, FM, delay, direction, temporal info, etc.	varied
Automatic Speech Recognition	passive	microphone	MFCCs, spectrogram	speech transcription
F0 detectors	passive	microphone	spectrum, correlogram	pitch, music transcription

Processing in active systems is much simpler because they are concerned only with signals coming from a single well-known source (the *ping*). On the other hand, passive systems are "blind" in the sense that they have no prior knowledge of the signal sources and must, in general, consider many signal sources as potentially important. This survey covers several types of both active and passive systems. We begin with active systems, due to their relative simplicity.

Active sensing

Audio in the 15 to 20 kilohertz range can be produced and recorded by a laptop computer but is inaudible to most adults. Thus, by using these audio frequencies, one can program an active

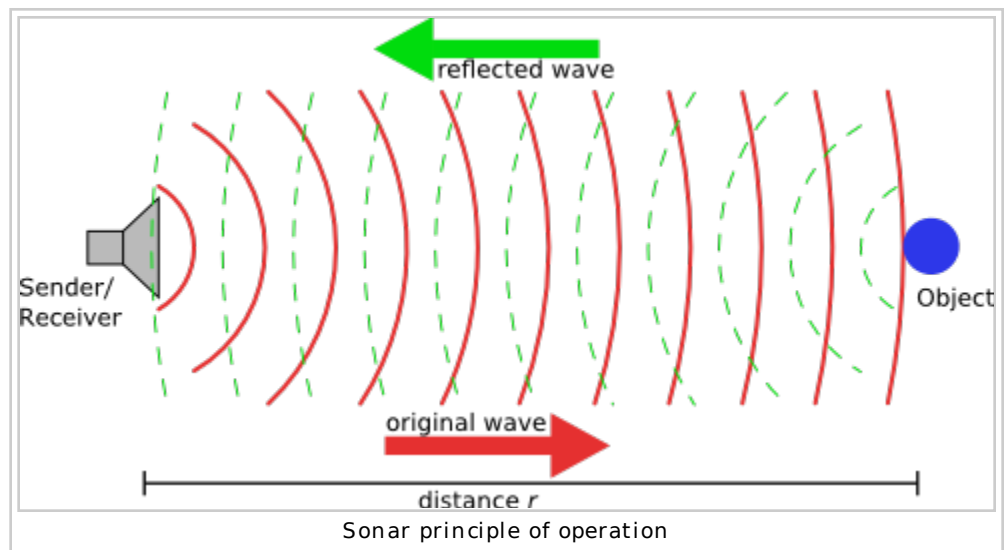
acoustic sensing system that is silent to the user. A practical active system would be limited to a relatively narrow band of ultrasonic audio frequencies. This limitation prevents most of the active approaches discussed below from being used directly. Hence, we must consider adopting passive approaches as well; the latter half of this survey covers passive audio identification techniques.

Alternatively, an active acoustic sensing system might process echoes of the computer's normal sound output. For example, if the user was listening to music, the echos of that music could be compared to the original waveform. Such a system would have the potential advantage of using a wideband stimulus. However, the stimulus waveform and occurrence would depend on the user's current musical preference, if any.

Sonar

Sonar systems emit sound *pings* and listen for the resulting echoes. Based on the characteristics of the echoes, a rough map of the surrounding physical space can be derived. Sonar is used by animals, such as bats and dolphins, for navigation and hunting. Man-made systems have been invented for fishermen, divers, submarine crews, and robotics; these systems are used both underwater and in air. Radar (<http://en.wikipedia.org/wiki/Radar>) uses electromagnetic rather than sound waves, but its basic principles are the same as sonar. Generally speaking, radar systems are favored in atmospheric applications while sonar systems are favored in underwater applications; this is due to the transmission characteristics of the two media.

The simplest type of sonar (and radar) systems work by emitting a brief, high-intensity, and highly directional sound ping in the direction of interest. The sound wave travels (at constant speed) for some distance and eventually reaches a solid object where it is reflected. Depending on the composition and angle of the reflecting surface, some portion of the ping's sound energy is reflected back to the source where it is recorded by a highly-sensitive directional microphone. In this type of sonar setup, the most important features of the recorded signal is the arrival time and intensity of the echoes. In practice, many echoes will occur, with varying intensities and delays which give a crude map of the physical environment in the direction that the system was pointed. The system can be rotated or multiple emitters and receivers can be used to map the environment in all directions surrounding the system.



recorded by a highly-sensitive directional microphone. In this type of sonar setup, the most important features of the recorded signal is the arrival time and intensity of the echoes. In practice, many echoes will occur, with varying intensities and delays which give a crude map of the physical environment in the direction that the system was pointed. The system can be rotated or multiple emitters and receivers can be used to map the environment in all directions surrounding the system.

Historically, research in sonar has largely been driven by its military applications. Sonar was developed for anti-submarine operations from World War I through the Cold War. Since World War II and Britain's simple but effective Chain home (http://en.wikipedia.org/wiki/Chain_Home) system, radar has been well studied for both military and civilian aircraft as well as weather forecasting applications.

In the 1970s, biologists began study the biosonar used by (microchiropteran) bats and marine mammals such as dolphins. These systems are, in fact, much more sophisticated than any man-made system yet developed. Bats, in particular, are able to measure the precise location, orientation, and trajectory of their insect prey with astonishing accuracy.

The following table compares the sonar systems of echolocating bats and dolphins. Their sonar is equally sophisticated but adapted to different environments and prey:

	bats	dolphins
sonar range	up to 3-4 m	up to hundreds of m
distance resolution	~ 2 cm	~ 2 cm
angular resolution	~ 4 degrees	better than bats due to larger head (better binaurality)
transmission duration	0.3-300 ms	0.04-0.1 ms
propagation speed	~ 330 m/s	~ 1500 m/s
transmission characteristics	narrowband constant freq, or broadband FM sweep	wide or narrowband click (transient)
transmission E	high, due to high sound absorption of air	high or low, due to efficiency of underwater sound transmission
transmission freq.	12-200 kHz	30-150 kHz
dynamics	transmissions change as the animal enters different hunting phases	transmissions remain relatively constant

The omnidirectional (unfocused) and relatively insensitive microphones and speakers built into most laptops are not ideal for building a precise sonar system. Therefore, traditional sonar techniques, which are capable of building a detailed range map, do not apply. However, there is hope that a cruder sonar system could be developed.

- Simon Tucker and Guy J. Brown, Classification of transient sonar sounds using perceptually motivated features, (http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1593805) Oceanic Engineering, IEEE Journal of , vol.30, no.3, pp.588-600, July 2005
- J. A. Thomas, C. F. Moss, and M. Vater, (editors). Echolocation in Bats and Dolphins (<http://nucatl.library.northwestern.edu/cgi-bin/Pwebrecon.cgi?BBID=3776644>) . University of Chicago Press. 2004.
- Mark Denny. Blip, Ping & Buzz: Making Sense of Radar and Sonar (<http://nucatl.library.northwestern.edu/cgi-bin/Pwebrecon.cgi?BBID=4768330>) . John Hopkins University Press. 2007.

Room Response

In the acoustics community, *room response measurement* essentially is the measurement of the acoustic properties of a room. With respect to a single stationary sound source and a single receiver, the reverberations characteristic of the room can be considered Linear and Time Invariant (LTI). An LTI system is described by the equation

$$\text{output}(t) = \text{input}(t) * h(t)$$

where **h(t)** is called the system's *impulse response* and * is the convolution operation. Once **h(t)** is determined, the output can be precisely calculated for any given input. Room response techniques attempt to measure **h(t)** by emitting excitation signal(s) and measuring the room's acoustic output (the recording):

$$\text{recording}(t) = \text{excitation}(t) * h(t)$$

The challenge is to choose a combination of excitation waveform and deconvolution technique that yields an impulse response $h(t)$ estimate with maximum signal-to-noise ratio and minimal non-linear artifacts.

The simplest approach uses a Dirac delta function (http://en.wikipedia.org/wiki/Dirac_delta_function) (aka unit impulse function) as the excitation. As the name implies, the impulse response $h(t)$ can be measured directly by applying such an excitation to an LTI system. However, a unit impulse is impossible to create in practice. It can be approximated by a short loud transient signal. So, firing a pistol in a room will produce echoes that approximate the room's impulse response; this procedure is actually suggested by the ISO.

More sophisticated approaches to calculate $h(t)$ use tone-sweeps or pseudo-random noise (e.g. maximum length sequences (http://en.wikipedia.org/wiki/Maximum_length_sequence)) as the excitation. These methods produce lower signal-to-noise ratio, assuming some given excitation energy constraint. Certain post-processing is needed to de-convolve the recorded signal into $h(t)$.

The reverberation time (http://en.wikipedia.org/wiki/Reverberation_time) of a room is another room acoustic characteristic which is less useful than the impulse response but which is easier to measure. It indicates the time after which echoes fall below a certain threshold (60 dB below the excitation). To measure reverberation time, one simply plays a long sound to bring the room to steady-state, stops the sound, and records the echoes' decay. Applying a low-pass filter to the recording gives the *decay curve* and the reverberation time.

In contrast with sonar, room response measurement calls for sound sources as omnidirectional as possible. This makes it suited to laptop speakers. However, broadband signals are typically used for excitation which would not fit within the limited silent ultrasonic band. In addition, depending on the room size, a very high energy excitation may be needed; the laptop's hardware may not be able to produce a sufficiently strong excitation. Also, the ISO procedure recommends that microphones not be placed too close to any sound source, though no explanation is given.

- ISO 3382:1997 Acoustics -- Measurement of the reverberation time of rooms with reference to other acoustical parameters (<http://pcfarina.eng.unipr.it/Public/NPL-workshop/ISO3382-1997.pdf>). International Organization for Standardization. Geneva, Switzerland, 1997.
- Gerzon, Michael A. Recording Concert Hall Acoustics for Posterity (http://www.acoustics.net/objects/pdf/review_aes_gerzon01.pdf) JAES Volume 23 Issue 7 pp. 569, 571; September 1975
- A Farina, R Ayalon. Recording Concert Hall Acoustics for Posterity. (http://www.acoustics.net/objects/pdf/article_farina05.pdf) AES 24th International Conference on Multichannel Audio. 2003
- Stan, Guy-Bart; Embrechts, Jean-Jacques; Archambeau, Dominique. Comparison of Different Impulse Response Measurement Techniques. (<http://www.montefiore.ulg.ac.be/~stan/ArticleJAES.pdf>) JAES Volume 50 Issue 4 pp. 249-262; April 2002



Passive sensing

A passive LAS system would process environment recordings directly. Passive systems are more complex because their task is less well-defined. Active systems simply calculate the transformation that the emitted sound underwent on its echo return path. Since the "ping" waveform is known beforehand, and the variety of possible acoustic transformations are limited (usually to just delay, amplification/attenuation, and frequency modulation), it is relatively easy to identify the important portions of the recorded signal and discard the rest as noise. The noise

filtering and classification problems for active systems are much more complex because they have no knowledge of what *should* be in the signal.

Audio Fingerprinting

In the content-based audio processing community, *audio fingerprinting* (also called *acoustic fingerprinting*) is the hashing of an audio recording (or a set of variations of an audio recording) to a unique identifier. The most common application domain for audio fingerprinting is music. Typically, the goal is match noisy snippets of recording to the full recordings stored in a database and thereby provide metadata such as the song name and artist for the snippet. These systems have been designed to be robust to several types of noise and are sometime explicitly targeted to cell phones, so we can safely say that such a music lookup service could very easily run on a laptop computer. We might consider adapting audio fingerprinting techniques to the room identification problem mentioned in the introduction. To evaluate this possibility, an understanding of audio fingerprinting techniques is needed.

There are two main steps in an audio fingerprinting system: fingerprint extraction and database matching. In the extraction step, the goal is to create a summary of the recording which is robust to content-preserving transformations. In the context of music, content is defined by a human perception so the fingerprint is a kind of perceptual summary. In room fingerprinting, we might want to identify which sounds are characteristic of the room. However, these sounds will not be present in the same way in every recording. For example, my office might be characterized by my telephone's distinctive ring tone, or by the voice of my officemate. However, the occurrence of a telephone ring and my officemate's utterances occur unpredictably. Therefore, a room should be characterized by a more flexible *model* rather than by a precise perceptual window, as is done in audio fingerprinting. To summarize, a room fingerprinting system must be robust to *content-rescheduling* transformations; the direct lookup-based matching scheme used by audio fingerprinting algorithms is not.

- Cano, P.; Batle, E.; Kalker, T.; Haitsma, J., A review of algorithms for audio fingerprinting, (http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1203274) Multimedia Signal Processing, 2002 IEEE Workshop on , vol., no., pp. 169-173, 9-11 Dec. 2002
- Jaap Haitsma and Ton Kalker, A Highly Robust Audio Fingerprinting Algorithm, (<http://ismir2002.ismir.net/proceedings/02-FP04-2.pdf>) Proceedings of ISMIR, 2002
- Burges, C.J.C.; Platt, J.C.; Jana, S., Extracting noise-robust features from audio data, (http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1005916) IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002. Proceedings. (ICASSP '02). pp. 1021-1024, 2002
- Burges, C. J., Platt, J. C., and Goldstein, J. 2003. Identifying audio clips with RARE. (<http://portal.acm.org/citation.cfm?id=957013.957104#>) In Proceedings of the Eleventh ACM international Conference on Multimedia (Berkeley, CA, USA, November 02 - 08, 2003). MULTIMEDIA '03. ACM, New York, NY, 444-445.
- Zmudzinski, S. and Steinebach, M. 2008. Psycho-acoustic model-based message authentication coding for audio data. (<http://portal.acm.org/citation.cfm?id=1411343#>) In Proceedings of the 10th ACM Workshop on Multimedia and Security (Oxford, United Kingdom, September 22 - 23, 2008). MM&Sec '08. ACM,

Other passive systems

Every science fiction future includes computers that respond to voice commands. Currently, such systems do not yet work well enough to be widely used except under controlled conditions and with high-quality microphones. However, researchers are making constant progress in several directions. Such passive systems include speech recognition as well as music identification, characterization and transcription. Research in these systems is beyond the scope of this survey.

Acoustics in mobile computing

Recently, there have been a few research projects in the ubiquitous and mobile computing communities which have used audio hardware for various nontraditional purposes. These represent the LAS state-of-the-art. Madhavapeddy et al. have used audio as a short-range wireless data transmission medium. Their system uses either ultrasonics or data-encoding musical melodies to achieve low-bitrate communication (8-14 bits/sec).

Madhavapeddy et al. also developed a room localization system called WALRUS which uses both 802.11 wifi and ultrasonics. WALRUS provides mobile laptop or PDA users with information about the room in which they are currently located. One disadvantage of WALRUS is that it requires some infrastructure; each room must contain a PC (or equivalent) which can occasionally broadcast a wifi room information packet while emitting an ultrasonic tone. The client devices listen for the wifi packets. When a wifi packet is received, the client switches on its microphone. The assumption is that each room is relatively well insulated from exterior sound but not from exterior wifi transmissions; several room information packets may be received from a given location, but only one contains information for the current room. If an ultrasonic tone is heard along with a wifi room information packet then that packet is recognized as originating from the current room.

BeepBeep is a simple acoustic ranging system which can determine the distance between two mobile phones (or PDAs, laptops, etc.). The two devices emit a tone while recording on their microphones; this is done on both devices at roughly the same time. Each device's recording will have both its own beep and that of its neighbor. Using the observed differences in beep times in both devices' recordings the signal propagation time (and thus distance) between the devices can be easily calculated. This approach requires that the devices can communicate through wifi, or some other mechanism to coordinate the semi-synchronous beeping and to exchange their observed beep time differences.

- A. Madhavapeddy, D. Scott, A. Tse, and R. Sharp. Audio Networking: The Forgotten Wireless Technology (http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1495392) . IEEE Pervasive Computing, vol. 4(3). 2005. pp. 55-60.
- A. Madhavapeddy, D. Scott, and R. Sharp. Context-Aware Computing with Sound (<http://www.springerlink.com/content/lft5fwctgbyac61l/>) . In Proceedings of The 5th International Conference on Ubiquitous Computing. 2003.
- G. Borriello, A. Liu, T. Offer, C. Palistrant, and R. Sharp. WALRUS: wireless acoustic location with room-level resolution using ultrasound (<http://portal.acm.org/citation.cfm?id=1067170.1067191>) . MobiSys '05: Proceedings of the 3rd international conference on Mobile systems, applications, and services. 2005.
- C. Peng, G. Shen, Y. Zhang, Y. Li, and K. Tan. BeepBeep: a high accuracy acoustic ranging system using COTS mobile devices (<http://portal.acm.org/citation.cfm?id=1322263.1322265>) . SenSys '07: Proceedings of the 5th international conference on Embedded networked sensor systems. 2007.

Conclusions and Future Work

This survey covered a wide range of topics. In this section I will review by summarizing the prospects for each acoustic sensing system on laptop computer hardware. Each system has some requirements that would be unfulfilled on a laptop, but these shortcomings may be allowed using various adjustments:

system	unfulfilled requirements	adjustments
sonar	mic/speakers are not directional enough to generate a detailed environment map, narrow ultrasonic	measure closest reflection surface rather than generating map, use constant-frequency ping

	bandwidth	
room response	speakers are too weak to reverberate a large room well, mic is too close to speakers, narrow ultrasonic bandwidth	limit to small rooms, mic-speaker proximity may not be a real problem, restrict to reverberation time measurement only
audio fingerprinting	recognizing a room is not a direct matching problem like song lookup	must build a <i>model</i> of the room, based on its common sounds rather than doing a direct database comparison

I believe that the prospects for each of these three types of systems is quite good, although I have not evaluated their potential impact.

Retrieved from "http://music.cs.northwestern.edu/classes/casa/index.php/Room_fingerprinting_-_by_Stephen_Tarzia"

- This page was last modified on 6 April 2009, at 16:48.