Receiving Data Hidden in Music*

Manuel Eichelberger
ETH Zurich
manuelei@ethz.ch

Simon Tanner
ETH Zurich
simtanner@ethz.ch

Gabriel Voirol
ETH Zurich
voirolg@ethz.ch

Roger Wattenhofer
ETH Zurich
wattenhofer@ethz.ch

ABSTRACT

This paper presents a method for transmitting data within music played from loudspeakers. The data is hidden in the music by leveraging the psychoacoustic masking effect, so that humans are not disturbed by the data transmission. The system achieves data rates of over 900 bits per second. The client side of the system could be implemented as a smartphone app, which receives data wherever users are without requiring any setup, making the system user-friendly.

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile devices**; • **Hardware** → **Sound-based input / output**;

KEYWORDS

audio, data transmission, masking effect, psycho-acoustics

ACM Reference Format:
Manuel Eichelberger, Simon Tanner, Gabriel Voirol, and Roger Wattenhofer. 2019. Receiving Data Hidden in Music. In *The 20th International Workshop on Mobile Computing Systems and Applications (HotMobile '19), February 27–28, 2019, Santa Cruz, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3301293.3302360

*The authors of this paper are alphabetically ordered.

HotMobile '19, February 27–28, 2019, Santa Cruz, CA, USA

ACM ISBN 978-1-4503-6273-3/19/02…$15.00
https://doi.org/10.1145/3301293.3302360

1 INTRODUCTION

Wireless communication standards such as Wi-Fi or Bluetooth demand a setup process. A smartphone user must explicitly allow the protocol, and potentially pair the device, which may entail entering passwords or pressing buttons in the right order. While this is acceptable when setting up a permanent communication channel, it is a hassle in countless ad hoc communication situations.[1]

In situations with (background) music, a user simply starts an app, and immediately receives data hidden within the music. Such a zero-setup scheme may have many interesting applications in situations where music is present anyway, or where music can be played without being disturbing. The transmitter hides the data inside the music and the smartphone can receive the data, as the microphone can be accessed without any setup. For example, a speaker in a store or a museum can piggyback information about close-by products. Also, attendees of a large sports event may suffer from a congested LTE connection, but interesting game statistics can be broadcast during musical jingles and breaks. When entering a hotel room, a guest's phone will directly receive the Wi-Fi login and password, maybe even immediately display a slider to control the room temperature.

While a smartphone is able to receive the information carried by the song, a person should not notice any degradation in sound quality. We strive for achieving high bit rates and robust data transmission while preserving imperceptibility independent of the nature of the chosen audio file. Since the frequency range has to be shared between the data and the hearable music signal, psycho-acoustic phenomena are exploited to embed the data into the cover signal. Using the frequency masking effect, our system transmits data in OFDM subcarriers next to frequencies of high amplitude.

Our prototype achieves bit rates of over 900 bit/s for a variety of music styles. In a hallway, our system can transmit data up to 24 meters with bit error ratios (BERs) below 5 %. In a big auditorium, the BER can be kept at 10 % at a distance of 15 meters.

[1] For a funny motivation, we refer to a recent xkcd comic: https://xkcd.com/2055/.

2 RELATED WORK

Several methods for hiding information in audio have been proposed, mostly in the context of *steganography*, aiming at data secrecy and robustness to compression and signal manipulation [4]. Several time domain methods are only applicable to file-based data transport and lead to high error rates when transmitting the audio over the air. These methods include least signification bit (LSB) encoding, echo hiding and hiding in silence intervals. In the frequency domain, LSB encoding in discrete wavelet transforms, phase coding and amplitude coding are employed.

Our focus is not on secrecy, but on embedding data in a way that does not disturb users, while achieving high data rates. Human auditory effects have been exploited before for data hiding using phase modulation, with simulation results achieving data rates up to 243 bit/s [9]. Our method transmits the data over the air, instead of hiding the information in a sound file and achieves data rates of up to 900 bit/s. While other approaches generate pleasant music from the information bits and may not achieve high data rates [8], we hide data in existing music.

Data can be transmitted hidden from listeners in ultrasound frequencies. This approach is only suitable for transmitting data over short distances [12], since the absorption of sound in air as well as the directivity of audio speakers increase with the frequency [5]. In the experiments with our system, which embeds data in lower frequencies, we show that data transmission is feasible over distances up to 24 m. Spread spectrum techniques focus on robustness while sacrificing data rate and achieve less than 100 bit/s [1]. Tone insertion is one approach which also uses the masking effect, like our technique, to maintain good audio quality. It has a better data capacity than spread spectrum methods, reaching for instance 250 bit/s [7]. Our method enables data transmissions with over 900 bit/s. An interesting technique is *modulated complex lapped transform (MCLT)*, which causes minimal distortions to the cover audio [17]. However, also this method achieves acceptable data rates only for short distances.

For this work, *orthogonal frequency-division multiplexing (OFDM)* is used. Other methods employing OFDM can transmit data either at several hundred bits per second over up to three meters [16] or send data at a lower data rate, but up to 5 or 7 meters [2, 11]. Using a high sound volume of 80 dB, the distance can be increased to 10 m [15]. Different from the existing work, our work takes into account the tonal harmonics of the used audio signal to find the most suitable frequencies for the OFDM subcarriers. This allows us to send data using lower sound volumes and still transmit over longer distances. Also, a non-data-aided time synchronization algorithm is implemented to increase the bit rate.

3 BACKGROUND

3.1 Human Auditory System

A so-called *psychoacoustic model* describes how audio frequencies are perceived by the *human auditory system (HAS)*. Acoustic waves are transformed into motion by the ear drum followed by ossicles in the middle ear. The spatial motion excites the fluids in the cochlea and thus the sensory cells of the basilar membrane which transfer the input to the nervous system. Excitations of similar frequencies or similar onset time instants are seen as one excitation by the basilar membrane. "Similar" frequencies are ones that fall into the same *critical bandwidth* which is a function of the frequency f and can be approximated by $BW_{cr}(f) = 0.2f$ for $f > 500$ Hz [6].

If a signal of high amplitude is present at a certain frequency f_m, a weak signal inside the critical bandwidth is imperceptible to a human. The signal at f_m is then called *masker* and the relation between its loudness and the one of the masked signal is described by a masking threshold which depends on their frequency difference [14].

3.2 OFDM

Orthogonal frequency-division multiplexing (OFDM) is a data modulation method where one OFDM symbol consists of a set of symbols sent on orthogonal subcarriers (SC) which are modulated with a method such as BPSK or QAM. The collection of subcarriers are transformed into a time domain OFDM symbol by an *inverse fast Fourier transform (IFFT)*.

4 SYSTEM DESCRIPTION

We propose a system for transmitting data hidden in music from a speaker to a smartphone microphone. Masking frequencies of a cover audio signal are detected and OFDM subcarriers are inserted close to them without humans being able to notice the modification. Still, a receiving smartphone can read out the information hidden in the music, – without knowing the music being transmitted – since microphones are not affected by the masking effect.

4.1 Transmission

The processing steps at the transmitting side are depicted in Figure 1. The cover audio signal is divided into segments H_i that are multiplied with a window function and then analyzed to find suitable frequencies for data insertion. OFDM subcarriers are inserted in a frequency range from 500 Hz to 10 kHz due to the sensitivity of smartphone microphones which decreases rapidly with higher frequencies.

The spectrum is split into an upper and a lower frequency region. For the first, a fixed bandwidth for data transmission is defined according to the properties of the HAS and smartphone microphones whereas for the latter, dominant

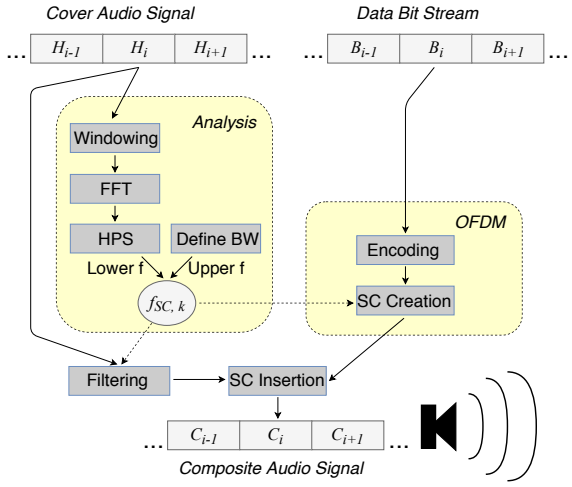

Figure 1: The processing steps to embed data into a cover audio file.

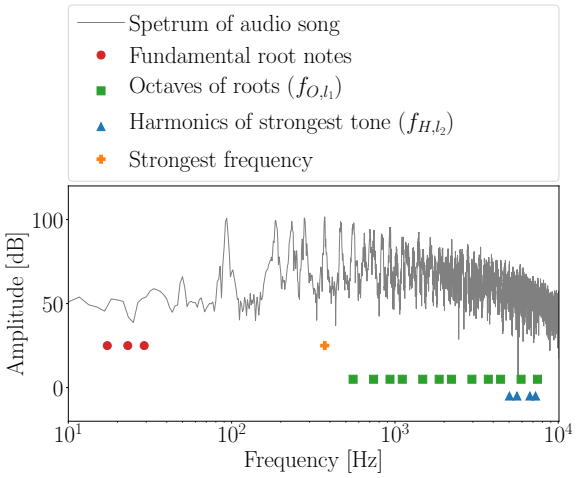

Figure 2: The higher octaves f_{O,l_1} of the fundamental root notes as well as the harmonics f_{H,l_2} of the strongest frequency represent the masking frequencies $f_{M,l}$.

frequencies of the cover signal are found with the method of the *harmonic product spectrum (HPS)* [3]. The masking frequencies $f_{M,l}$ for the audio segment H_i are obtained by this analysis. From these, the frequencies $f_{SC,k}$ of the OFDM subcarriers are derived. A filter is used to clear the spectrum of the audio segment H_i at these frequencies. The subcarriers are then modulated with information bits to generate the composite segment C_i.

The length of one analyzed segment of the cover audio signal is set to $L = 8820$ samples $\hat{=}$ 200 ms. The length of an OFDM symbol is equal to L and contains a cyclic prefix of 2940 samples $\hat{=}$ 66.6 ms. Therefore, the difference in distance of the first and the last arriving echo can be up to 22.4 m without degrading the reception quality.

Upper Frequency Region. Frequencies just below 10 kHz carry information such as drum cymbals and voice consonants. They also contain higher harmonics of tonal instruments and therefore shape their tone. However, in [10] it was noticed that frequencies of music songs in a band from 5 kHz to 10 kHz can be replaced by OFDM subcarriers of the same amplitude without a severe degradation in sound quality. To reduce the sound degradation even more, our system only uses the frequency band of 8 kHz – 10 kHz. The location and size of the upper frequency region is therefore a trade-off between bit rate and audio quality degradation.

The whole upper frequency region of the audio signal is filtered and replaced by OFDM subcarriers to obtain the same spectrum magnitude as the original signal.

Lower Frequency Region. The lower frequency region used for data insertion reaches from 500 Hz to 8 kHz. In this frequency range, the frequency masking is taken advantage of to hide data in subcarriers close to strong frequency components.

For the segment H_i, the fundamental *root notes* between the keys $C_0 = 16.35$ Hz and $B_0 = 30.87$ Hz of the three most dominant tones are found with the method of the harmonic product spectrum. From these three fundamental root notes which are too low to use for data transmission, the higher octaves in the desired frequency range are taken as masking frequencies f_{O,l_1}.

Large gaps between these masking frequencies are filled with the frequencies f_{H,l_2} which are the harmonics of the strongest frequency. Figure 2 shows the lower frequency region. The frequencies f_{O,l_1} and f_{H,l_2} are used as the masking frequencies $f_{M,l}$ from which the subcarrier frequencies $f_{SC,k}$ of an OFDM symbol are derived: The locations in the spectrum just below and above these masking frequencies are used to insert the subcarriers to profit from the masking effect. Two subcarriers are inserted above and two below each masking frequency.

The information which three out of the twelve root notes are used gets transmitted to the receiver in the upper frequency region to allow calculation of the subcarrier frequencies at the receiver. As depicted in Figure 3, a filter is used to remove the audio data at the subcarrier frequencies $f_{SC,k}$.

The spectrum inside the critical bandwidth of each subcarrier frequency $f_{SC,k}$ in the lower frequency region is analyzed. The amplitude of each inserted subcarrier is chosen such that the subcarrier is detectable by the receiver and at the same time does not exceed the masking threshold level.

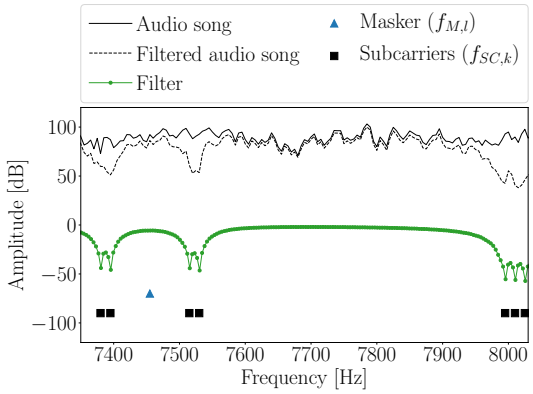

Figure 3: The cover song is filtered and subcarriers are inserted at those frequencies.

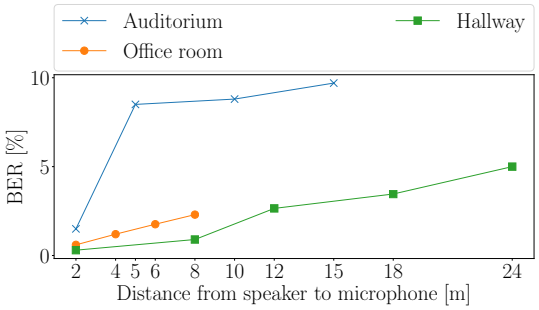

Figure 4: The BERs measured at different distances for the song *Viol* by *Gesaffelstein*.

Since the amplitudes of the subcarriers are determined by the cover audio, the information bits are encoded in the phase difference of two consecutive subcarriers at the same frequency using differential BPSK.

4.2 Reception

After the audio signal with the hidden data has been played back by a loudspeaker, the microphone at the receiving side records the signal. The onset of the OFDM symbols can be detected using the cyclic prefix as described in [13], even in the presence of multipath signals. The information about the most dominant notes is then obtained by decoding the upper frequency region and the set of subcarriers in the lower frequency band can be reproduced.

In our prototype, the signal processing is carried out offline on a computer. However, the computational effort is low and can be done in real-time on smartphones. For the implementation in an app, *forward error correction (FEC)* algorithms will allow reliable transmission.

Note that the receiver does not need to know the music that is transmitted by the loudspeaker. If multiple music sources use our method, the unwanted signal is additional noise in the overlapping OFDM frequencies. Also, we would like to point out that keeping a smartphone's microphone continuously active consumes only little power, as services such as voice search with "Ok Google" already do this today.

5 DATA HIDING AND TRANSMISSION EXPERIMENTS

The data transmission robustness is tested with different cover songs and under several conditions including varying distances from the speaker to the microphone and different amplification levels. The modified audio signals are played back on a *KRK Rokit 8* speaker and a *Nexus 5X* smartphone serves as receiver.

BER vs. Distance. The BER is measured at different distances in three environments: A narrow hallway with carpet, an office without carpet and a large auditorium with a wooden floor. The office contains desks with computer screens and chairs, the auditorium is equipped with rows of bench seats. Figure 4 depicts the BERs with increasing distance for the song *Viol* by *Gesaffelstein*. In the hallway and the office, the BER increases linearly with the distance. Since the hallway is narrow, the direct sound waves and the reflections move more unidirectionally towards the microphone. The BER at a distance of 24 meters does not rise above 5 % in the hallway. In the auditorium, the BER rises abruptly at 4 meters. A possible explanation is the layout of the auditorium where the vertical walls of the wooden benches block the echoes from the floor at distances larger than 2 meters.

BER vs. Volume. We compare two different speakers in a silent environment as well as the system's performance under the influence of various noise sources. Speaker 1 is the same *KRK Rokit 8* speaker as in the previous section, whereas Speaker 2 is a *Logitech X-530* hi-fi speaker. These experiments are conducted in an office room with carpet at a distance of 2 meters between the speaker and the smartphone. The BERs obtained in the different conditions are shown in Figure 5.

For high volumes, the BERs for both speakers are similar and remain at 1 % which seems to be a lower limit due to the residues of the filtered music signal which interfere with the OFDM subcarriers. At reduced volumes, the BER of Speaker 2 increases. Speaker 1 maintains an acceptable error ratio, probably since studio monitors are designed to show a flat frequency response independent of the volume.

To evaluate the impact of ambient noise on the system, noise has been recorded in three public areas: at a gym

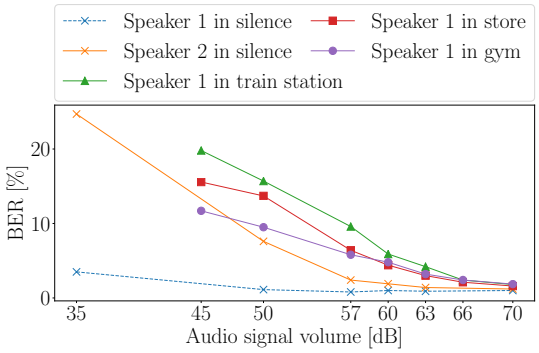

Figure 5: The BERs for varying audio signal volumes at a distance of 2 meters between the transmitter and the receiver.

(53 dB), in a store (55 dB) and at a train station (60 dB). Again, Speaker 1 plays the audio with the hidden information at different volumes while Speaker 2, placed next to Speaker 1, simultaneously replays the recorded noise, such that at the smartphone the level of the noise alone matches the levels in the public areas. As expected, the BER increases in louder environments and with decreasing audio signal volume.

Different Cover Songs. To test the applicability to various music types, our method has been applied to 9 pieces of 5 different music genres.[2] Song snippets of 12 seconds are modified for which the achieved bit rates are listed in Table 1. Depending on the song, a data rate of nearly 1 kbit/s can be achieved. Table 1 also lists the error ratios measured in the auditorium with a level of 63 dB at a distance of 2 meters. The exact bit rate depends on the cover song since the chosen set of fundamental root notes decides on the number of valid subcarrier locations inside the lower frequency region.

The greatest challenge for the algorithm are songs with high dynamic range, for instance classical music. Since the magnitudes of the data-carrying subcarriers are adapted to the spectrum of the host audio signal, they can sink below the minimum level needed by the receiver for correct decoding.

Continuously loud and busy Hard Rock or Electronic songs are optimal for hiding data as masking frequencies are strongly present over a wide frequency range.

5.1 Subjective Audio Quality Test

As objective metrics like the signal-to-distortion ratio do not necessarily correlate with the perceptibility of audio changes, we conducted our own experiments with test subjects.

[2]Song snippets are available at *https://doi.org/10.3929/ethz-b-000294476* - WAVE Files used in Audio Experiment for "Receiving Data Hidden in Music"

Table 1: List of analyzed songs sorted by the BERs.

Artist – Song	Bit Rate [bit/s]	BER [%]	Genre
J. S. Bach – Suite No. 2 in B minor BWV 1067: IV. Bourree I, II	971	14	Classical
Schola Romana Lucernensis & John Voirol – In Tempore Pacis	930	13	Classical
Bisco – Nothing's Left	922	4.3	Electronic
Fliptrix – Patterns Of Escapism	945	3.8	Hip Hop
Art Blakey – Moanin'	964	3.5	Jazz
Rümbold – Balcony	984	2.7	Hard Rock
John Coltrane – In A Sentimental Mood	953	2.5	Jazz
Sir Donkey's Revenge – Gawky Talky	982	2.1	Hard Rock
Gesaffelstein – Viol	964	1.5	Electronic

Test Environment. To gather data about the degree of perceptibility of the modifications by our method, the same song snippets used for the experiments in the previous section (listed in Table 1) were published on a website designed for this audio quality test.

In the first experiment, a blind test, for each song snippet the participant is given either the modified or the original version to listen to. For each song the participant has to choose whether the original or the modified song snippet is present.

In the second experiment, the direct comparison, the participant can listen to both versions of each song snippet and compare them directly. The participant then has to decide which of the two versions is the original.

Results. Table 2 summarizes the test results. For Experiment 1 (E1), the amount of visitors who assume that the original version is present is described by two quantities: $p(O|O)$ indicates that the original version is actually present whereas $p(O|M)$ stands for the case of an erroneously labeled modified version.

Δp is the difference between $p(O|O)$ and $p(O|M)$ and represents the benefit of the original version over the modified version to be interpreted as original. A high percentage signifies that many people identify the modified versions while the originals are still recognized. If the percentage is low, a similar amount of both versions are labeled as originals so the modifications are not noticed.

The results of Experiment 2 (E2) are described by $p(E)$, which stand for the probability of an erroneous decision. In average over 30 % of the probands are not able to tell the difference between the two versions in the direct comparison. For the blind test, in only 18 % of the cases the modified versions appeared more suspicious than the original ones. However, the original versions are labeled as such in 68 % of the test cases which indicates a bias towards the modified versions. Such a bias can occur since music pieces of

Table 2: Subjective audio quality test results: Experiment 1 (E1) is described using three columns $p(O|O)$, $p(O|M)$ and $p(E)$ (see Section 5.1) whereas the amount of errors in Experiment 2 is described by $p(E)$.

Artist – Song	$p(O\|O)$, E1 [%]	$p(O\|M)$, E1 [%]	Δp, E1 [%]	$p(E)$, E2 [%]
J. S. Bach – Suite No. 2 in B minor BWV 1067: IV. Bourree I, II	78.6	52.2	26.4	22.4
Schola Romana Lucernensis & John Voirol – In Tempore Pacis	56.2	36.8	19.4	28.6
Bisco – Nothing's Left	62.1	45.5	16.6	32.7
Fliptrix – Patterns Of Escapism	55.6	27.3	28.3	34.7
Art Blakey – Moanin'	79.2	55.6	23.6	42.9
Rümbold – Balcony	65.4	48.0	17.4	28.6
John Coltrane – In A Sentimental Mood	70.4	62.5	7.9	32.7
Sir Donkey's Revenge – Gawky Talky	84.6	68.0	16.6	38.8
Gesaffelstein – Viol	63.6	55.2	8.5	30.6
Average (50 Participants)	68.4	50.1	18.3	32.4

common audio quality are chosen which possibly contain artifacts and noise arising from the recording or compression. Participants paying attention to any abnormalities perceive any unfamiliar sound elements as suspicious and therefore the original versions are often labeled as being modified.

Note that finding a bitrate level which results in $p(E) = 50\ \%$ would require a huge number of experiments and participants, since $p(E)$ depends on the song and each participant can only judge one bitrate per song.

Similar to the findings when analyzing the BERs, the classical songs show a low performance in the direct comparison. A high dynamic range thus not only complicates a robust transmission but also the process of unobtrusive embedding.

6 CONCLUSION

In this paper, a novel system is proposed to hide data in music. The harmonic compositions of the cover songs are analyzed to exploit the masking effect. The system is evaluated under diverse circumstances. The BER measured at a distance of 15 meters in a big auditorium can be kept below 10 %. The method is applied to different cover songs and achieves bit rates of over 900 bit/s while being only slightly noticeable to the human ear. The computational effort for receiving the hidden data is low enough for real-time processing on smartphones. Through a standardized protocol, one universal app could be used in all situations without any user setup.

The next step in making our system accessible is a real-time implementation of the receiver in a smartphone app, maybe tailored for a specific use case. With additional user studies, the number and amplitude of the subcarriers can be optimized to allow higher data rates. Also, data rates might be increased by adapting the music, for instance by amplifying maskers.

REFERENCES

[1] Po-Wei Chen, Chun-Hsiang Huang, Yun-Chung Shen, and Ja-Ling Wu. 2009. Pushing information over acoustic channels. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2009, 19-24 April 2009, Taipei, Taiwan*. 1421–1424.

[2] Kiho Cho, Jae Choi, and Nam Soo Kim. 2015. An acoustic data transmission system based on audio data hiding: method and performance evaluation. *EURASIP J. Audio, Speech and Music Processing* (2015), 10. https://doi.org/10.1186/s13636-015-0053-x

[3] Patricio de la Cuadra, Aaron S. Master, and Craig Sapp. 2001. Efficient Pitch Detection Techniques for Interactive Music. In *Proceedings of the 2001 International Computer Music Conference, ICMC 2001, Havana, Cuba, September 17-22, 2001*.

[4] Fatiha Djebbar, Beghdad Ayad, Karim Abed Meraim, and Habib Hamam. 2012. Comparative study of digital audio steganography techniques. *EURASIP Journal on Audio, Speech, and Music Processing* 2012, 1 (2012), 25.

[5] H E. Bass, Louis Sutherland, and A J. Zuckerwar. 1990. Atmospheric absorption of sound - Update. 88 (1990).

[6] Hugo Fastl and Eberhard Zwicker. 2006. *Psychoacoustics: Facts and Models* (3 ed.). Springer-Verlag, Berlin, Heidelberg.

[7] Kaliappan Gopalan and Stanley Wenndt. 2004. Audio steganography for covert data transmission by imperceptible tone insertion. In *Proc. The IASTED International Conference on Communication Systems And Applications (CSA 2004), Banff, Canada*.

[8] Anil Madhavapeddy, Richard Sharp, David Scott, and Alastair Tse. 2005. Audio Networking: The Forgotten Wireless Technology. *IEEE Pervasive Computing* 4, 3 (2005), 55–60.

[9] Hafiz MA Malik, Rashid Ansari, and Ashfaq A Khokhar. 2007. Robust Data Hiding in Audio Using Allpass Filters. *IEEE Transactions on Audio, Speech, and Language Processing* 15, 4 (2007), 1296–1304.

[10] Hosei Matsuoka, Yusuke Nakashima, and Takeshi Yoshimura. 2006. Acoustic Communication System Using Mobile Terminal Microphones. *NTT DoCoMo Technical Journal* 8 (2006). https://www.nttdocomo.co.jp

[11] Hosei Matsuoka, Yusuke Nakashima, Takeshi Yoshimura, and Toshiro Kawahara. 2008. Acoustic OFDM: Embedding high bit-rate data in audio. In *International Conference on Multimedia Modeling*. Springer, 498–507.

[12] Nirupam Roy, Haitham Hassanieh, and Romit Roy Choudhury. 2017. Backdoor: Making microphones hear inaudible sounds. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 2–14.

[13] Jan-Jaap van de Beek, Magnus Sandell, and Per Ola Börjesson. 1997. ML estimation of time and frequency offset in OFDM systems. *IEEE Trans. Signal Processing* 45, 7 (1997), 1800–1805.

[14] Jesko Lars Verhey. 1999. *Psychoacoustics of spectro-temporal effects in masking and loudness perception*. Bibliotheks- und Informationssystem der Carl von Ossietzky Universitaet Oldenburg.

[15] Qian Wang, Kui Ren, Man Zhou, Tao Lei, Dimitrios Koutsonikolas, and Lu Su. 2016. Messages behind the sound: real-time hidden acoustic signal capture with smartphones. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking (MobiCom) 2016*.

[16] Shuai Wang. 2011. Embedding data in an audio signal, using acoustic OFDM.

[17] Hwan Sik Yun, Kiho Cho, and Nam Soo Kim. 2010. Acoustic data transmission based on modulated complex lapped transform. *IEEE Signal Processing Letters* 17, 1 (2010), 67–70.