IMPERCEPTIBLE AUDIO COMMUNICATION

Manuel Eichelberger, Simon Tanner, Gabriel Voirol, Roger Wattenhofer

ETH Zurich

ABSTRACT

A differential acoustic OFDM technique is presented to embed data imperceptibly in existing music. The method allows playing back music containing the data with a speaker without users noticing the embedded data channel. Using a microphone, the data can be recovered from the recording. Experiments with smartphone microphones show that transmission distances of 24 meters are possible, while achieving bit error ratios of less than 10 percent, depending on the environment. Furthermore, we present a user study which shows that many people do not recognize the added data channel in music, even when being informed about the experiment and therefore actively listening for the data transmission. Depending on the source music, data rates of 300 to 400 bits per second are achieved.

Index Terms— acoustic, data hiding, OFDM, music, user study

1. INTRODUCTION

We present an acoustic data transmission technique which embeds data into music. When a modified tune is played back by a speaker, a person listening to it cannot notice any degradation in sound quality but still, a smartphone is able to read out the information carried by the song. Challenges are achieving high data rates and robust data transmission while preserving imperceptibility independent of the nature of the chosen audio file. To tackle these challenges, the bandwidth used for data transmission is maximized by exploiting psycho-acoustic phenomena like frequency maskers and the tonal harmonics in music. By using OFDM (orthogonal frequency-division multiplexing) and adapting the subcarriers to the source music dynamically over time, the hearable frequency spectrum is maximally utilized for the data transmission. Our method achieves data rates up to 412 bit/s and can transmit up to 24 meters with bit error ratios below 10 %. A subjective audio quality study with 40 participants shows that humans cannot discern our modified and the original music, even when actively listening for inserted data.

By default, smartphones, laptops and tablets are equipped with microphones. At the same time, at many public places such as stores, stadiums, train stations and restaurants, speakers play background music. Our technique opens the potential for an easy communication path from speakers to microphones without the requirement of additional hardware or any setup. The data rate of several hundred bits per second is sufficient for various applications.

2. RELATED WORK

Our method is similar to audio steganography, whose goal is also to embed and conceal information in source data. Many time domain methods such as LSB encoding, echo hiding and hiding in silence intervals, and several frequency domain methods like discrete wavelet transform are only suitable for hiding data in audio files, but not for over-the-air transmissions [1]. A simple method to hide data from humans is to embed it in the ultrasonic spectrum. However, smartphone microphones have low sensitivity at high frequencies, and in air, ultrasound is damped stronger than hearable audio [2, 3].

When the data is hidden in the hearable frequency range, subjective audio quality tests are necessary to evaluate the impact of the method on the sound quality [4, 5]. Our work is also validated by an audio quality study.

Distributing the data in the frequency domain reduces the required carrier amplitude. This idea is used by spread spectrum techniques, which achieve data rates of up to 80 bit/s [6, 7, 8]. Using a *Modulated Complex Lapped Transform (MCLT)*, data rates of several hundred bit/s can be achieved, although only for distances of 3 or 7 m [5, 4]. We employ OFDM, which also embeds the data in different frequencies.

One OFDM data hiding technique adapts to the energy distribution of the host signal [9]. *Energy difference keying* adjusts the energy distribution around the subcarriers for source data with high energy spectrum density, while for low energy data, on-off *Amplitude-Shift Keying (ASK)* is implemented with subcarriers either being present or not. That technique allows reliable data transmission at a distance of 10 m when the music signal is played back at a level of 77 dB but sacrifices the imperceptibility of the modifications. In a similar corridor environment, our system achieves a BER of about 3 % with a sound level of 66 dB. Another work similar to our method also employs OFDM to hide data in music, but focuses more on use cases of transmitting hidden data and trades imperceptibility for higher data rates [10]. With *Quadrature Phase-Shift Keying*, data rates of up to 1.8 $\mathrm{kbit/s}$ can be achieved up to a distance of 6 m [11]. That work does however not analyze the subjective audio quality perception.

No method above has been tested with signal transmissions distances over 10 meters while our system achieves longer transmission distances using a single receiver.

As in our work, some spread-spectrum methods take advantage of the frequency masking effect to hide data [8, 12]. In addition to adapting the amplitudes at different frequencies, we account for the tonal harmonics of music to find the most suitable frequencies for the OFDM subcarriers. We believe that the combination of OFDM with the frequency masking effect achieves better transmission distances and rates than previous work while maintaining imperceptibility.

3. HUMAN AUDITORY SYSTEM

Depending on factors like age, health and mental state, humans can sense sound waves with frequencies in a range of 20 Hz up to 20 kHz [13]. How the frequencies of this bandwidth are perceived by the *human auditory system (HAS)* can be described by a *psycho-acoustic model*.

The authors of this paper are alphabetically ordered.

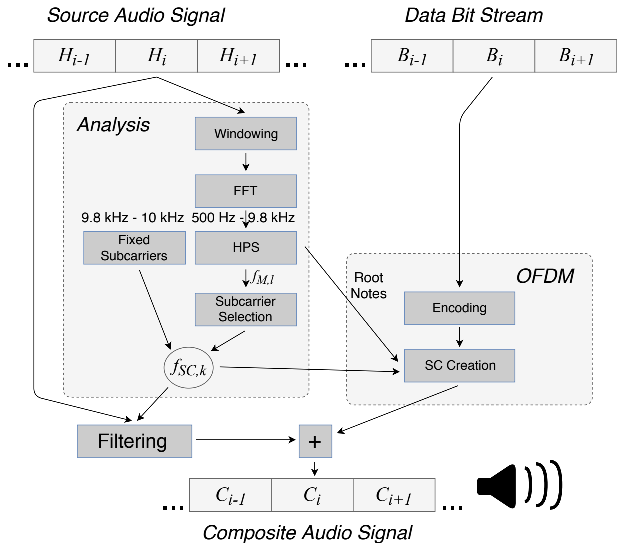

Fig. 1: The processing steps to embed data into a source audio file.

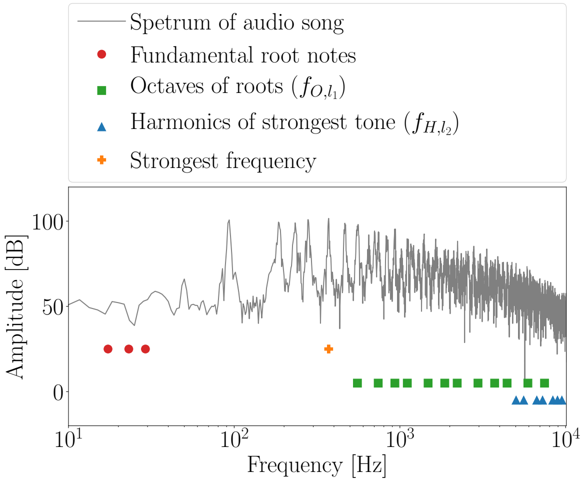

Fig. 2: The octaves f_{O,l_1} and the harmonics f_{H,l_2} are used as masking frequencies $f_{M,l}$.

Acoustic signals with similar frequencies or similar onset instants are perceived as one by the human ear. A range of similar frequencies is called *critical bandwidth* and its width depends on the frequency [14, 15]. The critical bandwidth can be approximated with $BW_{cr}(f) = 0.2f$ for $f > 500$ Hz [13]. If a high amplitude signal is present at a certain frequency f_m, weak signals in $BW_{cr}(f_m)$ are not heard by humans. The signal at f_m is then called *masker*. The loudness threshold for masked signals depends on f_m, the frequency difference to the masker and the masker's loudness [16]. Also in the time domain, a masker can cover a weaker signal shortly before and after. However, the masking threshold decreases rapidly with increasing time difference to the masker. Thus, the temporal masking effect can hardly be exploited [13].

4. SYSTEM DESCRIPTION

Our method uses the frequency masking effect to embed data. In time slices, masking frequencies are identified and OFDM subcarriers close to these maskers are filled with data. Figure 1 sketches the processing steps from the source file to the composite signal which is transmitted over loudspeakers.

4.1. Source Signal Analysis

The source audio signal is divided into consecutive segments for analysis. Each audio segment H_i of $L = 8820$ samples $\hat{=}$ 200 ms is multiplied with a window function to minimize boundary effects. In the frequency range from 500 Hz to 9.8 kHz, the dominant frequencies of the source signal are found, yielding the masking frequencies $f_{M,l}$ for the current segment. Additionally, in a small frequency band from 9.8 to 10 kHz, information is transmitted for the reconstruction of the subcarrier locations at the receiver. The upper limit of the used frequency region is set to 10 kHz due to the low sensitivity of smartphone microphones at high frequencies.

The masking frequencies are determined individually for each analyzed segment H_i. The three most dominant frequencies are found with the *Harmonic Product Spectrum (HPS)* method [17] and rounded to the closest notes of the harmonic chromatic scale. The fundamental *root notes* $f_{F,i=1\ldots3}$ lying between the keys $C_0 = 16.35$ Hz and $B_0 = 30.87$ Hz are determined this way. From the fundamental root notes which are too low to use for data transmission, their higher octaves $2^k f_{F,i}$ are calculated in the frequency range from 500 Hz to 9.8 kHz. Many of these frequencies f_{O,l_1} are strong in a given song due to the nature of the HPS. To fill large frequency gaps between the octaves, the strongest frequency of all the root notes and octaves is chosen and its multiples, the harmonics f_{H,l_2}, are used. The harmonics usually also have high amplitudes due to the overtones of instruments. Figure 2 shows the calculated octaves f_{O,l_1} of the root notes and the harmonics f_{H,l_2} of the strongest tone. Together, the octaves and the harmonics are used as the masking frequencies $f_{M,l}$. Based on $f_{M,l}$, the OFDM subcarrier frequencies $f_{SC,k}$ are derived. Below and above each masking frequency $f_{M,l}$, two subcarriers are inserted. The frequency difference of the subcarriers to the masker is kept small to allow for high subcarrier magnitudes as the masking threshold is highest in the immediate vicinity of the masking frequency.

This calculation of the maskers and the subcarriers has the advantage that only few information bits must be shared between the transmitter and the receiver to reconstruct the location of the subcarriers. Under the assumption that most common songs use the same chamber pitch, there are only 12 possibilities for a valid root note. Songs with notes being slightly off-tune or containing effects such as vibrato reduce the masking threshold, resulting in weaker embedded data signals. In the presence of broadband sounds, the data capacity might be higher than what we effectively use.

Previous work proposed that music frequencies between 5 and 10 kHz can be replaced by OFDM subcarriers without a severe degradation in sound quality as long as the subcarriers are of the same magnitude as the original signal [18]. We only use the band from 9.8 kHz to 10 kHz to transmit the information necessary to reconstruct the location of the used subcarriers. Our system can transmit 13 bits per OFDM symbol in this frequency range. These bits can be used to transmit the information about the subcarrier locations in the lower frequency range. This is enough for the 3 out of 12 root notes, the most dominant key out of the 3 chosen ones and the octave of the most dominant key. If 6 octaves are considered, there are $12 \cdot 11 \cdot 10 + 3 + 6 = 1329$ possibilities and therefore $\log_2(1329) \approx 11$ bits are sufficient.

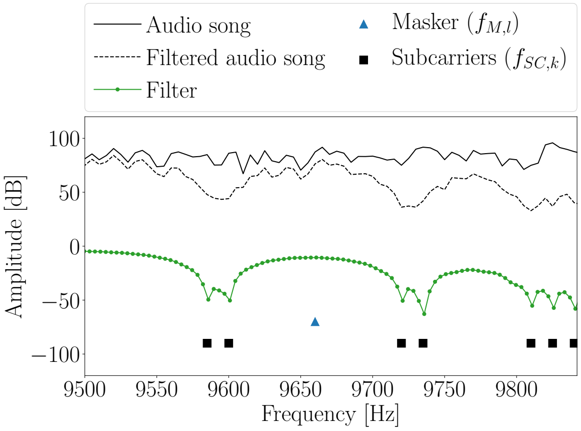

Fig. 3: The source song is filtered to insert the subcarriers close to the maskers and in the upper frequency region.

4.2. Subcarrier Modulation

A filter clears the spectrum of the audio segment H_i at the subcarrier frequencies $f_{SC,k}$. The OFDM symbol is then created based on the information bits in B_i and the composite segment C_i can be transmitted via loudspeaker. The magnitudes and phases of the subcarriers have to be chosen such that the receiver can extract the transmitted data while a person listening to the music does not notice the modification. Figure 3 depicts a spectrum excerpt and the subcarrier frequencies $f_{SC,k}$ of a segment H_i of a source audio song.

For the filtered audio segment, the spectrum inside the critical bandwidth $BW_{cr}(f_{SC,k})$ of each subcarrier frequency $f_{SC,k}$ is analyzed. The average $A(f_{SC,k})$ as well as the maximum value $M(f_{SC,k})$ inside $BW_{cr}(f_{SC,k})$ are calculated. The latter acts as a masker for the subcarrier $f_{SC,k}$. The magnitude of the inserted subcarrier is chosen such that the value $V(f_{SC,k})$ of the composite signal fulfills two conditions:

$$\begin{aligned} |V(f_{SC,k})| &\leq \gamma_M |M(f_{SC,k})|, \gamma_M < 1 \\ |V(f_{SC,k})| &\geq \gamma_A |A(f_{SC,k})| \end{aligned} \tag{1}$$

The first condition states that the magnitude of the composite signal $V(f_{SC,k})$ should not exceed the masking threshold which is defined by the magnitude of the strongest masker $M(f_{SC,k})$, scaled by a factor γ_M. It ensures imperceptibility if the masker does not change faster than an OFDM symbol duration. The second condition asserts the detectability of the hidden data at the receiver: If γ_A is increased, the magnitudes of the subcarriers increase and thus the BER decreases. However, at the same time the presence of the subcarriers can more easily be heard. Ideally, the first condition always holds, but in cases with no strong maskers in BW_{cr}, the subcarriers should still be decodable at the receiver. Therefore, the masking condition is ignored if it is not possible to fulfill both conditions. Because the amplitude is defined by those conditions, the data has to be transmitted in the phase of the subcarriers.

The upper frequency region from 9.8 kHz to 10 kHz is filtered and populated by the subcarriers as a whole. The same magnitudes are used as in the original audio signal.

In both frequency regions, differential BPSK is used to transmit the information and avoids the insertion of pilots. The length of the cyclic prefix is set to 2940 samples $\hat{=}$ 66.6 ms, which allows a distance difference of the first and the last arriving echo of up to 22.4 m.

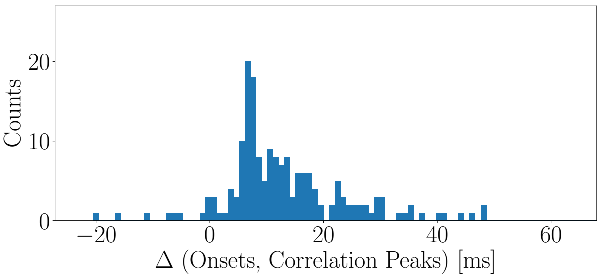

Fig. 4: Histogram of the differences between the real OFDM onsets and the measured correlation peaks measured in an auditorium at a distance of 5 meters from the speaker to the microphone.

4.3. Reception

When the audio signal with the hidden data is played back by a loudspeaker and transmitted through the air, a microphone at the receiving side records the signal. To find the start positions of the embedded OFDM symbols, the recordings are first band-pass filtered to extract the upper frequency range where no interfering music signals are present between the subcarriers. The onsets of the OFDM symbols can be found with the help of the cyclic prefix [19]. Once the onsets have been found, the information about the most dominant notes is obtained by decoding the upper frequency region. From this, the masking frequencies $f_{M,l}$ and therefore also the subcarrier frequencies $f_{SC,k}$ can be reproduced. OFDM is robust in the presence of narrow-band interfering sound sources, since these affect only some of the subcarriers.

5. EXPERIMENTS

The data transmission robustness is tested under several conditions. A *KRK Rokit 8* speaker plays back the modified songs while a *Nexus 5X* smartphone records them.

5.1. Time Synchronization

The histogram in Figure 4 shows the time differences between the real OFDM starting points and the calculated onsets in an auditorium at a distance of 5 meters between speaker and microphone. The majority lies between 0 and 25 ms, so a valid onset inside the cyclic prefix of 66.6 ms can be found. Note that the receiver takes into account that the OFDM symbol onsets should occur periodically, which makes the detection of the symbol starts even more robust.

5.2. BER vs. Distance

The impact of the distance on the bit error ratio (BER) is evaluated in three different environments: a hallway with carpet, an office with linoleum floor and an auditorium with wooden floor. The classic rock song *And The Cradle Will Rock* by *Van Halen* is used as source song for this experiment. The volume is adjusted such that the sound levels measured by the smartphone at a distance of 2 m to the speaker is 63 dB. Figure 5 shows the BERs for increasing distances up to the maximum sizes of the rooms. In the narrow hallway, the sound signal loses less power than in open areas and the BER is at most 5 %. In the auditorium, the BER increases significantly at 5 m. At

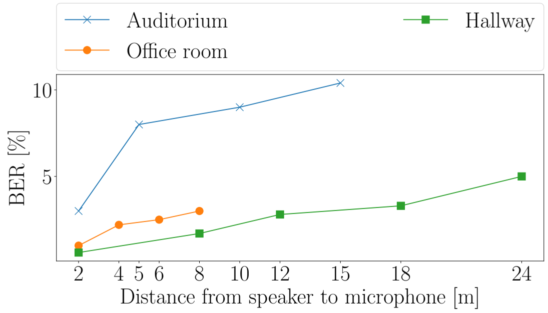

Fig. 5: The BERs measured at different distances to the speaker, up to the maximum size of the rooms.

Artist – Song	Data Rate [bit/s]	BER [%]	Genre
Munstrous – When Am I (SourOne Remix)	400	4.2	Electronic
Van Halen – And The Cradle Will Rock	359	3.0	Classic Rock
Pink Floyd – Breathe	309	2.9	Ballade
Scorpions – Can't Live Without You	315	2.8	Classic Rock
Queen – The Show Must Go On	412	2.6	Ballade
Gorillaz – All Alone	324	2.2	Electronic

Table 1: The list of songs analyzed in this project and their BERs.

this distance, wooden benches prevent echoes from the floor from reaching the microphone.

In the hallway, a sound level of 40 dB is measured by the smartphone at the maximum distance of 24 meters to the speaker. The levels at the maximum distances of 15 meters in the auditorium and 8 meters in the office are 55 dB and 57 dB, respectively. Apart from the decreasing level, the inter-symbol interference of echoes traveling for a longer time than the duration of the cyclic prefix plays an important role, too. Since the auditorium and the office are more reverberant, late echoes of OFDM symbols exceed the length of the cyclic prefix of $66.6\ \mathrm{ms} \widehat{=} 22.4\ \mathrm{m}$ and increase the BER.

5.3. Different Source Songs

To show the performance of the system for different music types, six songs of three genres are modified. Table 1 lists the music pieces with the obtained data rates and bit error ratios. The measurements are performed in the auditorium with a level of 63 dB at a distance of 2 meters. The data rate differs mostly because differential BPSK works best when the same subcarriers are used subsequently, which is the case when adjacent segments contain the same maskers.

Continuously loud songs provide an optimal base for hiding data, as masking frequencies are strongly present over a wide frequency range. Fast changing music may mask OFDM symbols only partially in time, because of the fixed length of the analysis window.

5.4. Subjective Audio Quality Tests

To evaluate the degree of perceptibility of the modifications, 12 seconds long song snippets[1] (see Table 1) were published on a website designed for this audio quality test. The data hiding method was applied to audio files in the *Free Lossless Audio Codec (FLAC)* format.

Artist – Song	$p(O\|O)$, E1 [%]	$p(O\|M)$, E1 [%]	Δp, E1 [%]	$p(E)$, E2 [%]
Munstrous – When Am I (SourOne Remix)	40.7	71.4	-30.7	71.8
Van Halen – And The Cradle Will Rock	33.3	40.0	-6.7	38.5
Pink Floyd – Breathe	64.0	62.5	1.5	20.5
Scorpions – Can't Live Without You	92.0	87.5	4.5	23.1
Queen – The Show Must Go On	60.0	68.8	-8.8	53.8
Gorillaz – All Alone	33.3	40.9	-7.6	35.9
Average (40 Participants)	53.9	61.8	-7.9	40.6

Table 2: Subjective audio quality test results: The blind test (E1) is described using three columns $p(O|O)$, $p(O|M)$ and Δp whereas the amount of errors in the direct comparison (E2) is described by $p(E)$ (see Section 5.4).

In the first experiment (E1), each participant is given either the modified or the original snippet to listen to and has to decide if the snippet is original or modified. In the second experiment (E2), the participant can listen to both versions of each song snippet arbitrarily often, and then has to decide which one is modified.

Table 2 lists the results of both experiments. For E1, the results are given by two values: $p(O|O)$ is the percentage of participants that labeled the original version correctly and $p(O|M)$ the percentage that labeled the modified version as original. Note that $p(M|O) = 1 - p(O|O)$ and $p(M|M) = 1 - p(O|M)$. $\Delta p = p(O|O) - p(O|M) = p(M|M) - p(M|O)$ is the percentage of people correctly identifying the song modifications minus the percentage of people identifying the song as modified no matter which version is played. Thus, only for the additional fraction Δp of all participants, the added modifications are perceptible when playing the modified song instead of the original. Interestingly, negative numbers occur in our test, which means that some people think the modified songs are even more "original" than the actual original ones. A negative average rate difference of almost $-8\ \%$ is obtained, indicating that the original and the modified version are perceived similarly.

In the direct comparison (E2), $p(E)$ is the percentage of erroneous decisions. Over $40\ \%$ of the participants are not able to identify the modifications. A perfect error rate of $50\ \%$ is approached even though the testers are actively paying attention to suspicious sounds.

Apparently, the uncompressed original versions of some songs already contain sounds which may sound suspicious and are therefore erroneously labeled as being modified.

6. CONCLUSION

In this paper, a system is proposed to imperceptibly transmit data in music. By analyzing the harmonic composition of a source song, suitable frequencies are found to exploit the frequency masking effect. OFDM subcarrier magnitudes are chosen such that robust transmission is possible while being hidden by the masking effect.

The transmission performance is evaluated under different conditions. The system achieves a data rate of up to $400\ \mathrm{bit/s}$ and the bit error ratio measured at a distance of 15 meters in a big auditorium can be kept at $10\ \%$. Additionally, our subjective audio quality test shows the imperceptibility of the source song modifications.

To improve the reliability of the transmission scheme, the spectrum of a song could be slightly modified to allow higher amplitudes of the data carriers. At the same time, the frequency response of the speaker can be selectively compensated to improve the audio quality.

[1] The song snippets are available at `https://doi.org/10.3929/ethz-b-000297564`.

7. REFERENCES

[1] Fatiha Djebbar and Beghdad Ayad, "Comparative study of digital audio steganography techniques," *EURASIP J. Audio, Speech and Music Processing*, p. 25, 2012.

[2] H E. Bass, Louis Sutherland, and A J. Zuckerwar, "Atmospheric absorption of sound - update," vol. 88, 1990.

[3] AcousticFrontiers, "Speaker directivity / off axis response: theory and measurement techniques," 2013, [Online; accessed 5-September-2018].

[4] Kiho Cho, Jae Choi, and Nam Soo Kim, "An acoustic data transmission system based on audio data hiding: method and performance evaluation," *EURASIP J. Audio, Speech and Music Processing*, p. 10, 2015.

[5] Kiho Cho, Hwan Sik Yun, and Nam Soo Kim, "Robust data hiding for MCLT based acoustic data transmission," *IEEE Signal Process. Lett.*, vol. 17, no. 7, pp. 679–682, 2010.

[6] Nevena Lazic and Parham Aarabi, "Communication over an acoustic channel using data hiding techniques," *IEEE Trans. Multimedia*, vol. 8, no. 5, pp. 918–924, 2006.

[7] Po-Wei Chen, Chun-Hsiang Huang, Yun-Chung Shen, and Ja-Ling Wu, "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*, 2009, pp. 1421–1424.

[8] Hosei Matsuoka, "Spread spectrum audio steganography using sub-band phase shifting," in *Second International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP 2006), Pasadena, California, USA, December 18-20, 2006, Proceedings*, 2006, pp. 3–6.

[9] Qian Wang, Kui Ren, Man Zhou, Tao Lei, Dimitrios Koutsonikolas, and Lu Su, "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, New York City, NY, USA, October 3-7, 2016*, 2016, pp. 29–41.

[10] Manuel Eichelberger, Simon Tanner, Gabriel Voirol, and Roger Wattenhofer, "Receiving data hidden in music," in *Proceedings of the 20th International Workshop on Mobile Computing Systems & Applications, HotMobile 2019, Santa Cruz, CA, USA*. ACM, 2019.

[11] Y. Nakashima, H. Matsuoka, and T. Yoshimura, "Evaluation and demonstration of acoustic ofdm," in *2006 Fortieth Asilomar Conference on Signals, Systems and Computers*, 2006, pp. 1747–1751.

[12] Xin Li and Hong Heather Yu, "Transparent and robust audio data hiding in subband domain," in *itcc*. IEEE, 2000, p. 74.

[13] Hugo Fastl and Eberhard Zwicker, *Psychoacoustics: Facts and Models*, Springer-Verlag, Berlin, Heidelberg, 3 edition, 2006.

[14] Lloyd A. Jeffress, "Chapter three - masking," in *Foundations of Modern Auditory Theory*, Jerry V. Tobias, Ed., pp. 85 – 114. Academic Press, 1970.

[15] Bertram Schaf, "Chapter five - critical bands," in *Foundations of Modern Auditory Theory*, Jerry V. Tobias, Ed., pp. 157 – 202. Academic Press, 1970.

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

[17] Patricio de la Cuadra, Aaron S. Master, and Craig Sapp, "Efficient pitch detection techniques for interactive music," in *Proceedings of the 2001 International Computer Music Conference, ICMC 2001, Havana, Cuba, September 17-22, 2001*, 2001.

[18] Hosei Matsuoka, Yusuke Nakashima, and Takeshi Yoshimura, "Acoustic communication system using mobile terminal microphones," *NTT DoCoMo Technical Journal*, vol. 8, 2006.

[19] Jan-Jaap van de Beek, Magnus Sandell, and Per Ola Börjesson, "ML estimation of time and frequency offset in OFDM systems," *IEEE Trans. Signal Processing*, vol. 45, no. 7, pp. 1800–1805, 1997.