Acoustic Data Transmission to Collaborating Smartphones – An Experimental Study

Roman Frigg
Disney Research & ETH Zurich
Zurich, Switzerland

Giorgio Corbellini
Disney Research
Zurich, Switzerland

Stefan Mangold
Disney Research
Zurich, Switzerland

Thomas R. Gross
Dept. of Computer Science
ETH Zurich, Switzerland

***Abstract*—The acoustic capabilities (i.e. microphone) and the fast processors of modern smartphones allow for the transmission of data to groups of such devices through the audio channel. We discuss an acoustic data transmission system for broadcast communication to a multitude of smartphones without the need of a radio access point. Acoustic data transmission is particularly attractive in scenarios that involve sound systems (e.g., movie theaters or open-air film festivals). We discuss different techniques to hide data in sound tracks and how to form a microphone array from a collection of smartphones in the same location. Collaborating smartphones share (using their radio interfaces to form an ad hoc network) the received data streams to jointly correct errors. With a testbed of up to four smartphones, we demonstrate how the robustness and reliability of a downlink broadcast via an acoustic communication system can be improved by collaboration between spatially distributed devices. With field tests in different scenarios, we investigate the potential gain of the collaboration in a real environment.**

I. Introduction

Modern smart- and feature phones are equipped with many different sensors such as cameras, microphones, accelerometers, or GPS. While connecting a smartphone to the internet is common, such functionality always requires a network infrastructure, e.g., Wireless LAN (WLAN) or cellular base stations. However, there are many situations where it would be desirable to send (push) information to smartphones without relying on any dedicated infrastructure. One example application is content dissemination for what is referred to as a "second screen" application in a cinema, theater, or with TV broadcast: During a show, additional information (beyond the movie) is provided to viewers. This information may include links to movie- or cinema-specific web sites, games and questionnaires, or coupons for repeated attendances. A communication channel for such a setup can be one-way (from show to visitor), have moderate capacity requirements, operate in broadcast mode (all visitors receive the same content) and be opportunistic (visitors without smartphones are left out, there is no guarantee that all visitors receive the extra content). For a show provider, it is attractive to provide such additional content without relying on local infrastructure (not to depend on WLAN in a cinema), to control experience and revenue share.

Acoustic data transmission provides an attractive path to reach an audience's smartphones. It enables communication through sound from loudspeakers to devices equipped with microphones (the smartphones). The additional (second screen) content is directly transmitted over the acoustic channel. Current smartphones employ rather simple microphones, therefore, embedding the second screen information so that it does not perturb the (human) owner's listening experience creates a number of challenges. (Section IV discusses in depth techniques and trade-offs for audio data hiding.) Fortunately the scenarios that can benefit from acoustic data transmission also usually include many smartphones, and these smartphones support WLAN communication. So if some smartphones suffer from poor audio reception, others in the same room or at the same event may have experienced better reception (or may have received the parts that are missing). The smartphones can form an ad hoc network through their radio interfaces (WLAN or Bluetooth) to jointly decode the second screen information.

978-1-4799-4937-3/14/$31.00 ©2014 IEEE

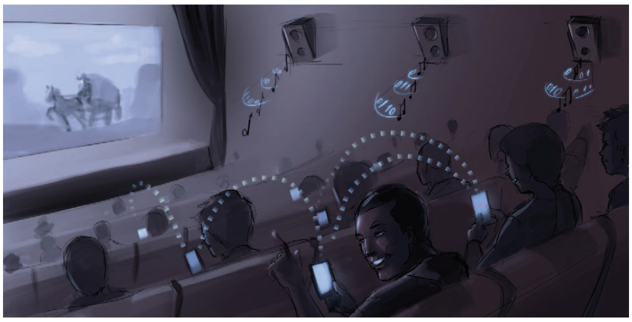

Fig. 1. Concept art (© Disney): Acoustic data transmission in a movie theater [1]. The smartphones are grouped together and jointly decode the received data. The system does not rely on any local infrastructure such as WLAN access points in the cinema, or cellular base stations.

This novel approach enables an improved reliability and performance. This paper describes the design, implementation, and practical experiments of an acoustic data transmission system for movie theaters that allows for the distribution of second screen content and uses collaboration between multiple receivers to improve link quality and robustness. Practical experiments include a thorough evaluation of our system, including a test campaign in a movie theater. The system was first demonstrated in [1]. The present study explores diversity, i.e. optimizing the reliability of the transmission link by combining the input data coming from multiple receivers. Figure 1 illustrates the use case.

Figure 2 illustrates the system architecture. The first step is embedding data into audio files (real-time or offline). The resulting audio is played back over a speaker. Multiple smartphone receivers, each with a microphone, process in real-time the received audio track to decode the hidden message. The receivers opportunistically form an ad hoc network that enables

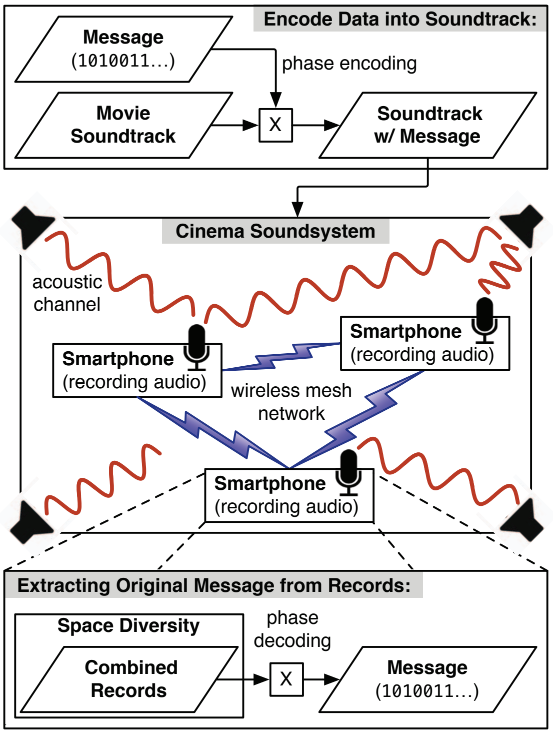

Fig. 2. System architecture: Multi-channel audio signals with data embedded are played over a speaker system and received by a group of wirelessly connected smartphones. The phones form an ad hoc radio network that enables them to jointly extract and decode the data.

them to collaboratively extract and decode the embedded messages.

II. Why Audio and not Wireless LAN

For show providers and content producers, it is attractive to offer such improved second screen applications that allow an enriched performance in a consistent way, without relying on a local infrastructure. If, e.g., a movie is shown during open air festivals, or in many cinemas across the world, it is desirable to enable the application with only the help of the audio system, but not WLAN or cellular networks: This setup ensures that the new experience can be provided everywhere and during any event, including TV broadcast, video on demand, or with DVD media, without the need to ensure cooperation of a local infrastructure provider (that may demand compensation). Another advantage of acoustic data transmission is that synchronization of additional content with the main content is trivial: e.g., TV broadcast arrives at slightly different times in different homes (around 10-20s deviation is common as different TV systems – analog, digital, cable, satellite – create different delays). It is therefore not possible to distribute in-sync second screen content from a centralized server over the internet. With acoustic data transmission it is however simple, because the additional content uses the same channel as the main content.

III. Related work

Several different methods for audio hiding have been presented previously [6], [7]. According to the low-bit coding technique, the Least Significant Bit (LSB) of each sample is replaced with the message bit to be embedded. This is one of the simplest information hiding methods for audio signals. It allows for high data rates (44100 b/s (bit per second) that are hidden in audio signals with a sampling rate of 44.1 kHz). LSB coding is not robust against most kinds of signal processing operations as well as noise and interference occurring through aerial transmission. Another method, called echo hiding, exploits the fact that the human auditory system is not capable of distinguishing artificially introduced echoes in an audio signal from the echo that a room might introduce naturally caused by its acoustics [8]. Artificially introduced echoes with different offsets can therefore be used to encode data. Echo hiding is more robust against signal processing operations and is less audible than LSB coding. However, the low data rate around 50 b/s does not favor the use of this technique in acoustic data transmission systems. A spread spectrum method can be used to communicate data over an acoustic channel [9]. Spread spectrum hiding relies on pseudo-random sequences embedded as noise in the frequency domain and exploits auditory masking to hide information in audio signals [10]. A different approach is to use a spread spectrum hiding method together with phase coding, a technique that manipulates the phases of several subbands in the frequency domain to encode a message [11]. Phase coding is a suitable method for data embedding in audio signals because the human ear is not sensitive to phase changes and thus, the hidden information remains mostly inaudible. The collaborative audio communication system presented here uses the phase coding technique [11]. Further, the Modulated Complex Lapped Transform (MCLT) [12] is preferred to short-time Fourier transform to compute spectrograms of audio signals because it introduces less audible artifacts. The collaborative audio system investigated in this study exploits space diversity techniques known from radio communication; in the scenario of audio communications, multiple spatially distributed receivers allow to increase the reliability of the acoustic data transmission system. In space diversity for radio communication, multiple receivers and/or transmitters are used to improve the quality and reliability of wireless communication links. The usage of multiple antennas at both the transmitter and the receiver is known as Multiple-Input Multiple-Output (MIMO) [13],

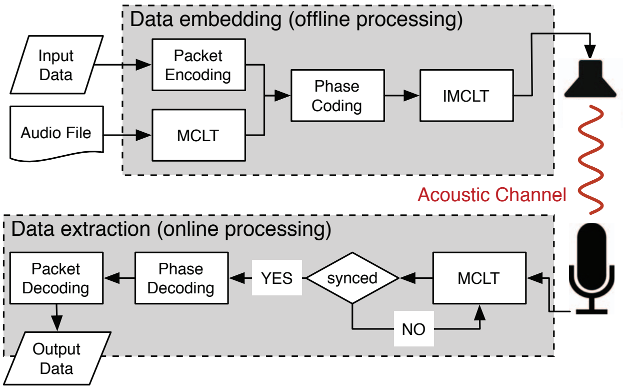

Fig. 3. Audio data hiding. The digital data is encoded into the phase information of the orginal audible sound, at frequencies subbands determined by the (inverse) modulated complex lapped transform, as described in [2]–[5].

which is an important part of modern wireless communication standards such as IEEE 802.11n [14]. All speakers that are used to playback the embedded stream carry the same data. The system aims at having no infrastructure requirements (beyond what is necessary to encode the data stream) and keeps all essential real-time processing to the receiver. This arrangement is flexible and makes no assumptions on the number of loudspeakers.

IV. Efficient audio data hiding

This section describes the acoustic data transmission method along with the data hiding techniques that are combined in the system. As mentioned earlier, the approach is based on the acoustic data transmission technique [11], which uses a phase coding method in the MCLT domain [12]. The key modules of the system are shown in Figure 3. Each of the modules depicted in the figure together with the control flow between them are described in the following paragraphs.

A. Time-frequency representation

To alter the phases of an audio carrier signal, it is necessary to compute its time-frequency representation. The MCLT, a $2\times$ oversampled generalized Discrete Fourier Transform (DFT) filter bank [12], is applied for that purpose. The MCLT causes fewer artifacts than standard DFT filter banks when applied for audio processing [12]. Transforming a block of $2D$ real-valued audio samples results in D complex-valued MCLT coefficients. The coefficients represent the amplitude and phase at D subbands, which are equally distributed over the frequency spectrum. To limit the impact on audibility, only a portion of the spectrum is used to embed data (only M subbands out of D are used). The number of subbands M defines how many bits can be embedded in each MCLT block.

The system setup uses a sampling rate of 44.1 kHz. For an MCLT block with 4096 complex samples (subbands), there are 4096 phase coefficients that can be altered to embed data in the entire spectrum between 0 and 22.05 kHz. However, since the encoder typically considers frequencies between 6 and 10 kHz for information hiding, there are 4 kHz to spread over M subbands. The aforementioned frequency range with a block size of $D = 4096$ results in $M = 384$ carrier subbands, meaning that every subband has an approximate width of 10 Hz. After embedding data, the Inverse MCLT (IMCLT) is applied to transform the signal back to the time domain.

B. Packet and bit encoding

Repetition and CRC codes are used to increase the reliability of the data link. Repetition coding is chosen for simplicity and can be replaced by a more efficient forward error correction. For simplicity, the size of all packets is constant (typically 60 bytes). With packets aligned with the audio blocks, there is not need for a packet header.

1) Spread spectrum coding to add redundancy: Each bit is translated into a code sequence of size K consisting of multiple symbols. The translation of each bit has the effect to spread it over multiple carrier subbands thereby decreasing the probability of transmission errors. Spreading codes are often used in radio communication to limit cross-channel interference. For a code length of $K = 4$ and the alphabet of code symbols $C = \{-1, 1\}$, the mapping from bit to binary spreading codes is defined by

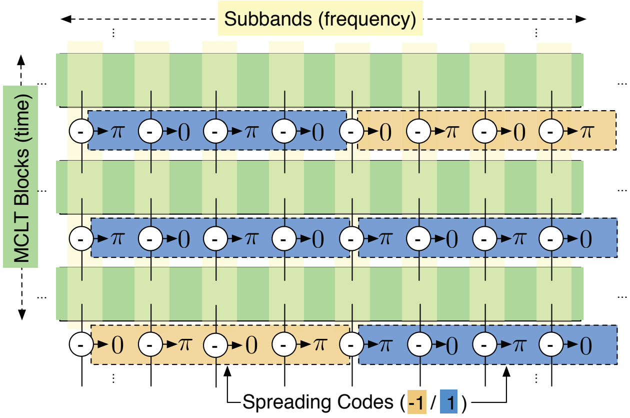

Fig. 4. Differential phase coding: Bits are encoded in the differences between phases of adjacent audio blocks. Differences do not change in case of phase shifts, so that the hidden bits can be decoded correctly without the need for additional synchronization.

$$\begin{aligned} 1 \mapsto \quad & c_{one}(k) = \{-1, 1, -1, 1\} \\ 0 \mapsto \quad & c_{zero}(k) = \{1, -1, 1, -1\}. \end{aligned}$$

This mapping is used to translate the data bit by bit to a binary spreading code sequence. The resulting sequence is a concatenation of the codes $c_{one}(k)$ and $c_{zero}(k)$:

$$\{s(n)\}_{n=0}^{KN-1} = \{s(0), s(1), \ldots, s(KN-1)\}, \tag{1}$$

where $s(n) \in C = \{-1, 1\}$. N is the length of the embedded bit sequence and therefore also the number of spreading codes $c_{one}(k)$ and $c_{zero}(k)$ contained in $s(n)$. KN is the total length of $s(n)$.

2) Phase encoding to embed the data: Differential Binary Phase Shift Keying (DBPSK) is used to encode data into the phase spectrum of the carrier signal. The encoding is differential because bits are encoded into the differences between phases of two consecutive MCLT blocks. It is binary because the phases are modulated to either 0 or π. The phase spectrum is modified as

$$X'_{i,m} = \frac{X_{i-1,m}}{|X_{i-1,m}|}|X_{i,m}|s((i-1)M + m), \tag{2}$$

where $i = 1, 2, \ldots, B-1$ with B being the number of MCLT blocks and $m = 0, 1, \ldots, M-1$ with M equal to the number of used subbands per block. $X_{i,m}$ is the original phase at the m-th subband of the i-th MCLT block and $X'_{i,m}$ represents the modified phase at the m-th subband of the i-th block. The sequence of code symbols $s(n) \in \{-1, 1\}$ is derived from the data do be embedded. Figure 4 illustrates the phase encoding technique: For each MCLT block the phase coefficients in the different subbands are modulated so that their difference to the corresponding coefficients in the previous block becomes either 0 or π. In the shown configuration,

$K = 4$ neighboring subbands represent a spreading code that encodes either a 1 (logical 1) or a -1 (logical 0). The first MCLT block acts as a reference and does not contain any hidden data. Therefore, the index i in Equation 2 starts from 1. Phase alterations cause only rotations of the MCLT coefficients in the complex plane whereas the magnitudes are unchanged.

The system employs phase coding based on DBPSK, whereas previous work uses BPSK without differential coding [11]. A benefit of DBPSK over BPSK is that the encoded bits are not affected by phase shifts introduced by the transmission over an acoustic channel or the digital-to-analog/analog-to-digital converters. When using BPSK, different phase shifts occur in different subbands leading to more complex synchronization and decoding processes. In that case, previously defined synchronization sequences must be transmitted between sequences carrying data so that it is possible to identify the effective phase shift at the receiver and compensate for this shift in the decoding process.

3) Interference cancellation: Once the modified phase content incorporates the bits to be embedded, the MCLT blocks can be transformed back to the time domain using an Inverse MCLT. The MCLT has a major drawback: In the MCLT domain there is significant interference among frequency responses of neighboring blocks and subbands [11]. Applying the overlap-add operation of the IMCLT mixes the phase content of adjacent blocks. The result is that the modified phase content of an audio file is not the same anymore, after synthesizing the audio and later transforming it back to the MCLT basis again. A possible way to cancel MCLT interference is to use every other block and subband to embed data. Hence, by subtracting a correction coefficient from the altered phase $X'_{i,m}$, it is possible to compensate in advance for the interference that is introduced by neighboring subbands and the overlap-add operation [11]. This cancellation technique requires that the adjacent blocks and subbands remain unmodified. Because data can only be embedded into every other block and subband, four times less information can be transmitted compared to transmitting without the cancellation techniques. By using a wider frequency range in which data is embedded (larger value of M), it is possible to compensate for that limitation, although a wider frequency range with manipulated phases makes the embedded data more audible.

C. Decoding at receiver

To decode the embedded data, the receiver analyzes the phase content in the MCLT domain: The phases of the subbands are extracted and normalized to obtain the sequence $r(n)$ of the form

$$\{r(n)\}_{n=0}^{KN-1} = \{r(0), r(1), \dots, r(KN-1)\},$$

where $r(n) \in [-1, 1]$. The difference compared to the original sequence of spreading codes before transmission, Equation 1, is that $r(n)$ can contain arbitrary values in the range $[-1, 1]$ instead of only 1 and -1. The magnitude of each value is proportional to the likelihood that it represents the corresponding spreading code symbol. A value close to 0 means that the given value does not correspond to any of the expected spreading code symbols. For MCLT coefficients $X_{i,m}$, $r(n)$ is computed by

$$r((i-1)M + m) = -2\left(\frac{\min\{|d_{i,m}|, 2\pi - |d_{i,m}|\}}{\pi} - 0.5\right), \quad (3)$$

with $d_{i,m} = \text{angle}(X_{i,m}) - \text{angle}(X_{i-1,m})$ using the same index ranges for the variables i and m as in Equation 2. The variable $d_{i,m}$ is the angular difference between the phase of the current (i) and the previous $(i-1)$ MCLT block (given the same subband m). To make sure the smaller of the two possible differences between two phases (inner and outer angle) is used, the minimum of the two is computed. This operation leads to values in the range of $[0, \pi]$. We map the phases into the range of $[-1, 1]$ dividing them by π, subtracting 0.5 and multiplying by -2. The minus sign comes from the fact that the spreading code symbol 1 is embedded with a phase of 0, whereas π is used for the symbol -1. Then, $r(n)$ is mapped back to the logical bit stream with symbols 0 and 1 as follows: Equally sized sub-sequences of length K are cross-correlated with $c_{one}(k)$, the spreading code for the 1-bit. Hereby K is the code length, the number of symbols in $c_{one}(k)$ and $c_{zero}(k)$. The result is a sequence of correlation coefficients $\rho(n)$ that is defined as

$$\rho(n) = \frac{1}{K}\sum_{k=0}^{K-1} r(nK + k)c_{one}(k), \quad (4)$$

where $n = 0, 1, \dots, N-1$. N is equal to the length of the embedded bit sequence representing the original message. The sign of each correlation coefficient $\rho(n)$ indicates the corresponding logical bit (1-bit if positive, 0-bit if negative).

D. Synchronization

Before being able to decode an embedded message in an incoming audio stream, a receiver must find the correct partition into blocks to compute the correct time-frequency representation. If the audio block partition is wrong, the resulting MCLT blocks do not contain the phases that represent the original message. Therefore, each receiver needs to synchronize to the signal before starting to decode the embedded data. The size of the MCLT blocks is a constant parameter. If a receiver figures out the right sample offset at which a new block of audio samples starts, it can correctly partition the audio and can synchronize to the embedded signal. To find the correct offset, the synchronization algorithm tries to decode a single audio block at each possible sample offset. Bits are only encoded into every other block because of the interference cancellation described in Section IV-B. Thus, for a block length of D, there are $2D - 1$ offsets to be tested. To determine the correct offset, the synchronization algorithm decodes a single block for each offset k, resulting in a sequence of correlation coefficients $\rho_k(n)$ as defined in Equation 4 with length N. (N here represents the number of embedded bits in a single block.) To synchronize, the receiver evaluates the mean of the absolute values of this sequence, which gives a measure of signal strength $S(k)$ as a function of the offset k:

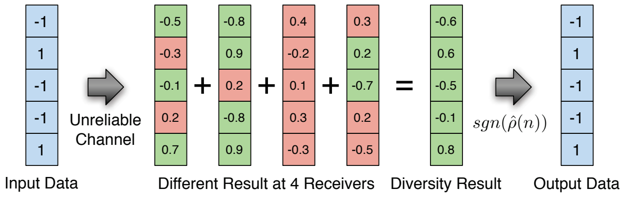

Fig. 5. Space diversity with spreading code correlation sequences. Although none of the four receivers can correctly decode the input data, all errors are corrected through collaboration.

$$S(k) = \frac{1}{N} \sum_{n=0}^{N-1} |\rho_k(n)|.$$

The mean of the absolute values of $\rho_k(n)$ is computed for every offset $k = 0, 1, 2, \ldots, 2D - 1$. The maximum signal strength is reached in correspondence to the offset q: At the optimal offset q, $\rho_q(n)$ mostly consists of values close to the spreading code symbols 1 or -1, which leads to a high signal strength $S(q)$. That is not the case at a sub-optimal offset r. There, $\rho_r(n)$ mainly contains values close to 0, which causes $S(r)$ to be small. The closer k is to the optimal offset, the higher the signal strength. The system achieves synchronization without dedicated synchronization codes which reduces the overhead and represents an improvement with respect to previous work [11]. The process of packet decoding includes the CRC check per each received packet to determine its validity.

V. Smartphone collaboration

The system described in this paper exploits collaboration and diversity, i.e., joint packet decoding and error handling of the nodes (smartphones) that are in the target setup. Multiple physically separated receivers allow observing the same signal under different conditions. Often, if one receiver experiences a large amount of destructive interference, one of the other receivers has sufficient signal quality. By combining the received signals from all the receivers, link quality and reliability can be improved. Phase decoding results in a sequence of correlation coefficients as defined in Equation 4. Correlation coefficients give a measure for the correctness (or, confidence) of the corresponding bits which can be used for diversity means in an easy way: To exploit spatial diversity, all correlation sequences are combined as

$$\hat{\rho}(n) = \frac{1}{L} \sum_{l=0}^{L-1} \rho_l(n), \qquad (5)$$

where L is the total number of receivers, $\rho_l(n)$ is the correlation sequence at receiver l and $\hat{\rho}(n)$ is the combined spreading code correlation sequence. Equation 5 simply computes the average of the signals received at the different receivers. Bits with high corresponding correlation coefficients are likely to be decoded correctly and therefore have a stronger influence on the combined sequence than the bits with a low corresponding correlation coefficient. The sign of $\hat{\rho}(n)$ determines if the corresponding bit is a -1 (logical 0) or a 1 (logical 1).

In the rest of the paper, this collaborative approach is referred to as space diversity with spreading code correlation sequences. Figure 5 shows a numerical example of the diversity scheme. In the figure the spreading code sequence $s = \{-1, 1, -1, -1, 1\}$ is transmitted over an unreliable channel. Four receivers decode the input signal leading to different spreading code correlation sequences. Correctly decoded values are indicated in green, incorrect values in red. Although no receiver decodes the whole sequence on its own correctly, the combined output of the diversity scheme is correct.

VI. Practical experiments and evaluation

This section provides an evaluation of the acoustic data transmission system conducted in two different indoor scenarios: a movie theater and a lab space. The two test environments differ in terms of acoustic characteristics: The shape and size of the rooms as well as the wall cladding are different, causing distinct frequency responses and reverberation times. In both cases, up to four receivers were grouped together.

For all the tests a popular Reggae song is used as the carrier signal [15]. Due to the song's more or less consistent frequency spectrum over time, many samples can be measured under similar conditions. Data is embedded into the song using a Matlab implementation of the procedure described in Section IV. The receivers are smartphones running a dedicated application that features the diversity scheme of Section V. The application also handles the radio connectivity (Bluetooth or Wi-Fi) among the receivers. The MCLT block size D used in the evaluation experiments is either 2048 or 4096 samples. The choice of the block size is a trade-off between inter-block interference, computational complexity of the receiver, and audibility. On one side, longer blocks limit the relative inter-block interference. In fact, the IMCLT process causes partial overlapping of consecutive blocks. However, the length of the overlapping region is independent from the block size, thus, with longer blocks the inter-block interference becomes smaller relatively to the full length of a block. On the other side, a larger block size increases both computational complexity of the MCLT and impacts the perceived quality of the track [16]. An empirical evaluation found that a block size of 4096 samples is an acceptable compromise, confirming previous work [17]. Note that an MCLT of a signal of length N using a block size of D has a complexity of $\mathrm{O}\left(\frac{N}{D} Dlog(D)\right) = \mathrm{O}\left(Nlog(D)\right)$. The smaller the block size, the faster the MCLT computation executes. Three different performance criteria are considered:

- **Bit error rate:** The ratio of the number of incorrectly received bits to the total number of received bits;
- **Packet loss rate:** The fraction of packets for which the received CRC is invalid;
- **Throughput:** The throughput in b/s counts the number of bits of correctly received (valid CRC) packets.

A. Movie theater experiments

The movie theater used for the experiments has a size of 10 × 20 m and offers 132 seats. The experiment is performed using three speakers, located behind the projection screen at left, center and right positions. The volume level is set to 80 dB on average. The error bars in all the following plots indicate the 95 percent confidence intervals.

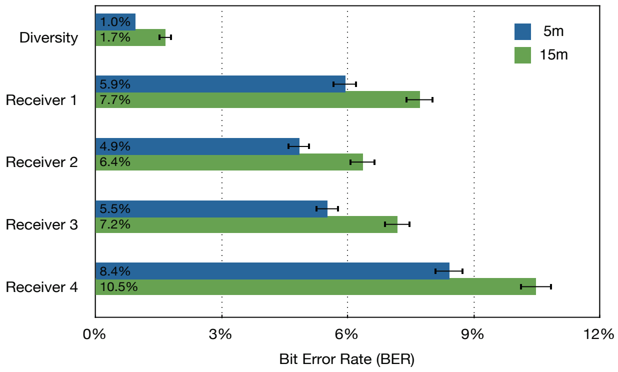

Fig. 7. Movie theater field test. BER of four receivers with diversity arranged at two distances from the audio source. The diversity scheme outperforms the BER at all individual receivers. MCLT block size $D = 4096$ samples.

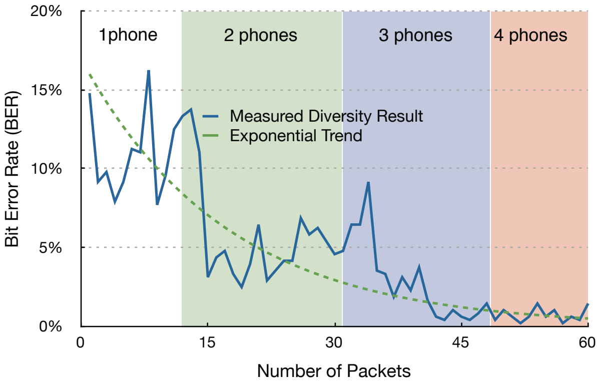

Fig. 8. Movie theater field test. BER for a varying number of receivers at 15 m from the screen using an MCLT block size $D = 2048$ samples. On average, with every additional phone in the network, the BER is reduced.

1) Bit error rate: The network of smartphones contains up to four devices. In every experiment, all devices are deployed in a row, uniformly spaced and at the same distance from the screen. Smartphones number 1 and 4 are close to the walls of the movie theater, number 2 and 3 are placed near the center of the respective row. Figure 7 shows the BER for every individual device along with the BER resulting from the diversity scheme. The most important observation is that the diversity scheme causes a BER drop to about one third of the value measured at the best performing receiver. For a block size of 4096 the BER lies between 1.0 and 1.7 percent, whereas the best performing receiver (Receiver 2) has a BER between 4.9 and 6.3 percent. It is evident that the described diversity algorithm does not simply select the best among the receivers, but combines the results of all available receivers to achieve a much lower BER. Furthermore, Figure 7 shows that the receivers closer to the center of the cinema $(2, 3)$ perform more reliably than receivers near the side walls $(1, 4)$. Especially Receiver 4, which in each arrangement is located only about thirty cm from the left side wall, has a higher BER than the more centered receivers. The receivers close to walls seem to experience stronger interference due to reflections.

Figure 8 shows the dependence of spatial diversity from the number of devices that collaborate. The figure illustrates the result for an experiment for a block size of 2048 samples at a distance of 15 meters from the screen. Increasing the network size over time results in a notable performance gain.

2) Packet loss rate with ASCII text: This test is closer to a real scenario because it evaluates the performance for the transmission of text characters. Figure 9 shows the packet loss rate of the diversity scheme for both single device and network scenarios.

In addition to the data of the four individual receivers and the diversity scheme, the plot includes another data set for a theoretical receiver referred to as logical AND. The logical AND receiver incurs a packet loss if and only if all the four measured receivers lose the packet. Thus, this theoretical receiver provides a reasonable baseline to compare the diversity scheme with. The results show that none of the individual receivers is able to reliably decode the ASCII text. Already at a distance of five meters from the screen, the packet loss rates lie between about 30 and 50 percent. For the same reasons described in the analysis of the previous experiment, receivers closer to the center of the movie theater perform better than the ones near the walls.

The most interesting result shown in Figure 9 is that the logical AND baseline achieves lower packet loss rates than each of the receivers individually, but cannot by any means compete with the performance of the collaboratively computed diversity scheme. The diversity algorithm is often able to decode packets correctly through collaboration for which each receiver on its own does not compute a valid CRC code.

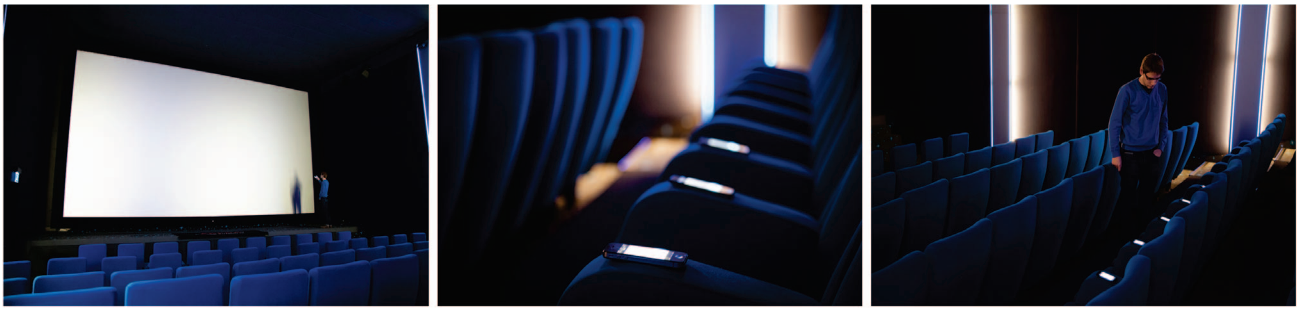

Fig. 6. Impressions from measurement campaign. Commercially available smartphones are used to receive data from the multi-speaker audio system of a modern cinema with 132 seats. The test sound used is music. There is no audience attending.

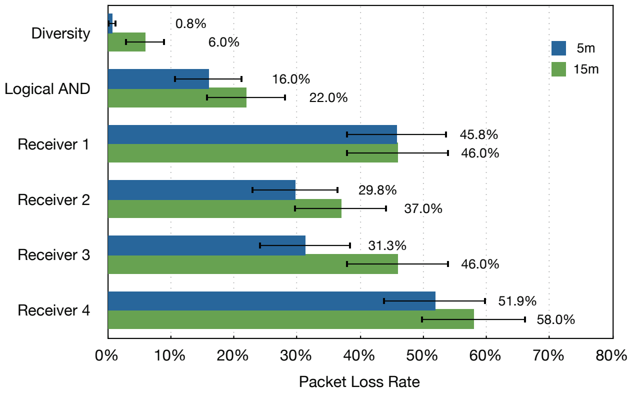

Fig. 9. Movie theater field test. Packet loss rate for the transmission of ASCII packets with 3 × redundancy to four receivers located at two distances from the audio source. MCLT block size $D = 4096$ samples.

B. Lab space measurements

The results of the experiments discussed so far show already that collaboration between receivers can drastically improve the reliability of an acoustic data transmission system in movie theaters. This section shows the achievable throughput at increasing distances, for a different numbers of collaborating phones. The lab space used for measurement is a narrow corridor of length 35 m and width 2 m. All measurements consider a mono setup with a single speaker. There are no obstacles between the speaker and the receivers. In every experiment, all receivers are close to each other at the same distance from the speaker. The MCLT block size is $D = 4096$ samples and the same pop song as in the movie theater scenario is used as the carrier signal. All figures indicate the 95 percent confidence intervals.

1) Throughput: Figure 10 shows the throughput for different network sizes. When the distance between speaker and receivers is almost zero, the achieved throughput comes close to the maximum channel bandwidth of 516 b/s (for the specific configuration used in the evaluation), independently of the number of devices that collaborate. The benefit of increasing the number of devices becomes evident observing the shape of the throughput curve for one receiver; this curve presents a dramatic drop already at distances of 15 m. The addition of a second receiver results in a considerable gain. For example, the throughput at 35 m for one device is 26.4 b/s; as the number of devices is increased the throughput grows to 179.5 b/s, 262.7 b/s and 403.7 b/s for 2, 3 and 4 collaborating devices, respectively. In contrast to the movie theater scenario, all devices are close to each other, thus, there are only minor differences between the performance of different devices.

2) Bit error rate: The general trend is that BER increases with distance from the loudspeaker. As shown in Figure 11, increasing the distance negatively affects the BER, given the same number of collaborating devices. As expected, a single receiver performs worst. Four collaborative receivers have a BER of around 6 percent at 35 m, about the same number of errors an individual receiver experiences at only 10 m distance. The results from Figure 11 and Figure 7 numerically differ from each other, although the same parameters were used for the tests. The reason for this effect is that the BER depends on the acoustic properties of the environment where the audio track is played back. However, in both scenarios the gain of the proposed diversity scheme is evident.

C. Subjective audio quality comparison

We conducted an online study to assess the quality of audio signals processed with our audio hiding method. Audio material with embedded data was evaluated with a MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) test [18] by 93 volunteers. Figure 12 shows the mean scores for each phase coding configuration and type of audio sample.

Figure 12 indicates that the impact is most perceivable for speech. The audio quality comparison test scenario encouraged the participants to carefully listen to the audio files and examining them for the slightest impairments. Cinema visitors are not expected to do that with a soundtrack, as they mainly go to watch a movie for entertainment reasons and not to assess its audio quality. Therefore, we expect the hidden data to be subjectively less audible in a cinema scenario. To

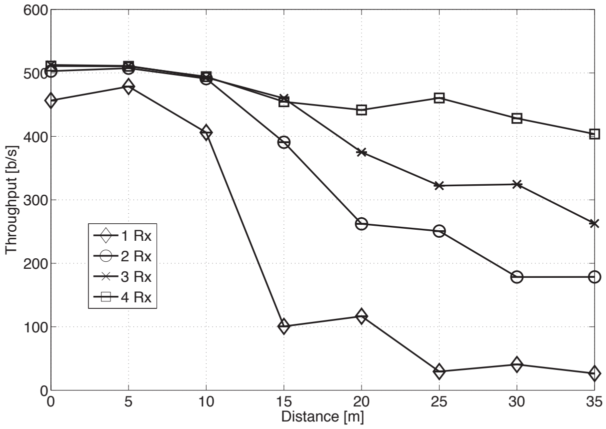

Fig. 10. Lab space field test. Achievable throughput vs. distance from loudspeaker for different network sizes. MCLT block size $D = 4096$ samples. The higher the number of collaborating devices, the better the throughput.

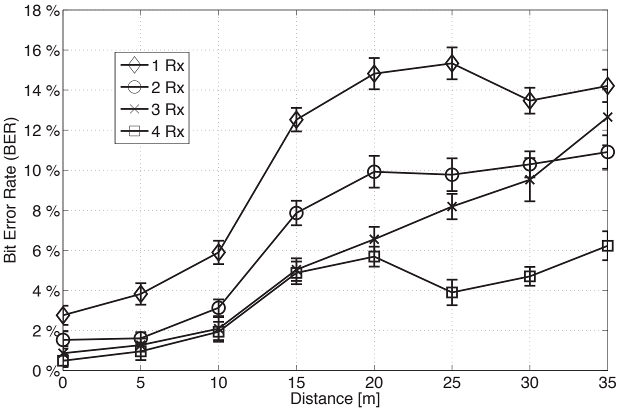

Fig. 11. Lab space field test. BER vs. distance from loudspeaker for different network sizes. MCLT block size $D = 4096$ samples. Increasing the number of collaborating devices affects the maximum achievable distance.

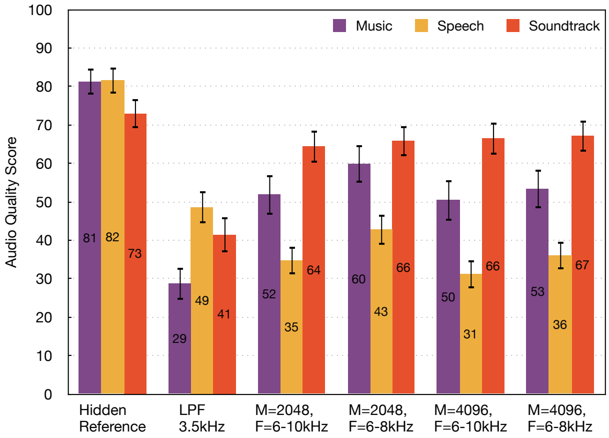

Fig. 12. MUSHRA test scores with 95% confidence intervals. Hidden reference, 3.5 kHz low-pass filtered signal (LPF) and four test configurations with different block sizes (M) and frequency ranges (F)

reduce perceptibility, the frequency range of the embedded data can be narrowed at the expense of a smaller capacity of the acoustic channel. Furthermore, the soundtrack can be analyzed to identify regions where embedded bits remain the least perceptible and the phase encoding can be adapted to embed information only into blocks of audio that comply with this metric.

VII. Concluding Remarks

In many parts of the world, a dedicated network infrastructure is not practical or affordable. Yet smartphones are popular and prevalent. Collaborative audio transmission allows dissemination of content in such settings to implement a "second screen" that may enable new approaches to interactive and enriched story telling or audience engagement.

The novel way of combining the input of multiple receivers of an acoustic data transmission to achieve a reliable transmission channel presented in this paper has many desirable properties. Space diversity with spreading code correlation sequences, as our method is called, makes use of a radio ad hoc network between receivers to exchange information about the received bits: Receivers jointly decode the received input signals, resulting in a lower bit error rate.

The paper presents results from an evaluation in a cinema setting and a lab space but the technique can also be employed in other environments. The acoustic characteristics of a movie theater have a positive influence on the performance of the acoustic data transmission. The cinema allows the system to cover large distances of up to 15 m at bit error rates below 2 percent. In the lab space, four devices can communicate with BER below 7 percent up to 35 m. However, the most important finding of our field test evaluation is that the diversity scheme outperforms each individual receiver by a factor of at least two in every experiment and for each parameter configuration. Results indicate that the kind of collaboration used in our diversity scheme provides a larger improvement of transmission performance and reliability than the one that can be achieved by simply determining the currently best performing receiver and relaying its result to all the other receivers. In many situations, the diversity scheme produces a correct result, although none of the individual receivers can correctly decode a packet. We conclude that applying diversity techniques to acoustic data transmission is a simple and effective way to improve reliability.

VIII. Acknowledgement

The visualization in Figure 1 was created by the artist Maurizio Nitti of Disney Research, Zurich.

References

[1] R. Frigg, T. R. Gross, and S. Mangold, "Multi-channel acoustic data transmission to ad-hoc mobile phone arrays," in *ACM SIGGRAPH 2013 Mobile*, ser. SIGGRAPH '13. New York, NY, USA: ACM, 2013.

[2] H. S. Malvar, "Lapped Transforms for Efficient Transform/Subband Coding," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 38, no. 6, pp. 969–978, 1990.

[3] H. Yun, K. Cho, and N. Kim, "Acoustic Data Transmission Based on Modulated Complex Lapped Transform," *Signal Processing Letters, IEEE*, vol. 17, no. 1, pp. 67–70, 2010.

[4] H. S. Malvar, "Fast algorithm for the modulated complex lapped transform," *Signal Processing Letters, IEEE*, vol. 10, no. 1, pp. 8–10, 2003.

[5] T. Aach, "Fourier, block, and lapped transforms," *Advances in Imaging and Electron Physics*, vol. 128, pp. 1–50, 2003.

[6] C. Lopes and P. Aguiar, "Aerial acoustic communications," in *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*, 2001, pp. 219–222.

[7] A. Madhavapeddy, R. Sharp, D. Scott, and A. Tse, "Audio networking: the forgotten wireless technology," *Pervasive Computing, IEEE*, vol. 4, no. 3, pp. 55–60, 2005.

[8] D. Gruhl, A. Lu, and W. Bender, "Echo hiding," in *Information Hiding*. Springer, 1996, pp. 295–315.

[9] N. Lazic and P. Aarabi, "Communication Over an Acoustic Channel Using Data Hiding Techniques," *Multimedia, IEEE Transactions on*, vol. 8, no. 5, pp. 918–924, 2006.

[10] B. Moore, *An Introduction to the Psychology of Hearing*. Academic Press London, 1982, vol. 4.

[11] K. Cho, H. Yun, and N. Kim, "Robust Data Hiding for MCLT Based Acoustic Data Transmission," *Signal Processing Letters, IEEE*, vol. 17, no. 7, pp. 679–682, 2010.

[12] H. Malvar, "A Modulated Complex Lapped Transform and its Applications to Audio Processing," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, vol. 3. IEEE, 1999, pp. 1421–1424.

[13] V. Tarokh, N. Seshadri, and A. Calderbank, "Space-Time Codes for High Data Rate Wireless Communication: Performance Criterion and Code Construction," *Information Theory, IEEE Transactions on*, vol. 44, no. 2, pp. 744–765, 1998.

[14] IEEE-SA, "IEEE 802.11n-2009 - Amendment 5: Enhancements for higher throughput," 2009.

[15] Bob Marley & the Wailers, "No woman no cry," 1974.

[16] L. Liu, J. He, and G. Palm, "Effects of Phase on the Perception of Intervocalic Stop Consonants," *Speech Communication*, vol. 22, no. 4, pp. 403–417, 1997.

[17] K. Cho, J. Choi, Y. G. Jin, and N. S. Kim, "Quality Enhancement of Audio Watermarking for Data Transmission in Aerial Space Based on Segmental SNR Adjustment," in *Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), 2012 Eighth International Conference on*. IEEE, 2012, pp. 122–125.

[18] I.-R. R. BS, "Method for the Subjective Assessment of Intermediate Sound Quality (MUSHRA)," *International Telecommunications Union, Geneva, Switzerland*, 2001.