Single-sensor multispeaker listening with acoustic metamaterials

Yangbo Xie, Tsung-Han Tsai, Adam Konneker, Bogdan-ioan Popa, David J. Brady, and Steven A. Cummer

Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708

Edited by Ping Sheng, Hong Kong University of Science and Technology, Kowloon, China, and accepted by the Editorial Board June 29, 2015 (received for review February 3, 2015)

Designing a “cocktail party listener” that functionally mimics the selective perception of a human auditory system has been pursued over the past decades. By exploiting acoustic metamaterials and compressive sensing, we present here a single-sensor listening device that separates simultaneous overlapping sounds from different sources. The device with a compact array of resonant metamaterials is demonstrated to distinguish three overlapping and independent sources with 96.67% correct audio recognition. Segregation of the audio signals is achieved using physical layer encoding without relying on source characteristics. This hardware approach to multichannel source separation can be applied to robust speech recognition and hearing aids and may be extended to other acoustic imaging and sensing applications.

Significance

Combining acoustic metamaterials and compressive sensing, we demonstrate here a single-sensor multispeaker listening system that functionally mimics the selective listening and sound separation capabilities of human auditory systems. Different from previous research efforts that generally rely on signal and speech processing techniques to solve the “cocktail party” listening problem, our proposed method is a unique hardware-based approach by exploiting carefully designed acoustic metamaterials. We not only believe that the results of this work are significant for communities of various disciplines that have been pursuing the understanding and engineering of cocktail party listening over the past decades, but also that the system design approach of combining physical layer design and computational sensing will impact on traditional acoustic sensing and imaging modalities.

Author contributions: Y.X., D.J.B., and S.A.C. designed research; Y.X. and T.-H.T. performed research; Y.X., A.K., and B.-I.P. analyzed data; and Y.X. and S.A.C. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. P.S. is a guest editor invited by the Editorial Board.

1To whom correspondence should be addressed. Email: cummer@e.ee.duke.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1502276112/-/DCSupplemental.
indexed frequencies. This matrix is randomized by the physical properties of the metamaterials to generate highly uncorrelated information channels for sound wave from different azimuths and ranges. The level of randomization of the matrix determines the supported resolution and the multiplexing capability of the sensing system.

To quantify the signal encoding capacities of the modulation channels, here we have chosen the average mutual coherence \( \mu_{av} \) as the metric of the sensing performance (36). \( \mu_{av} \) ranges from a desirable 0 (indicating perfectly orthogonal modulation channels) to a useless 1 (indicating identical and thus indistinguishable modulation channels). Average mutual coherence is directly related to the mean-squared error (MSE) of the reconstruction (36). The frequency responses of the modulation channels that are used for calculating the average mutual coherences are obtained from the Fourier transform of the measured impulse responses of each spatial location. For our experiment presented here, the metamaterials are shown to provide to the sensing task an average mutual coherence of 0.198. In contrast, an omnidirectional sensor without the metamaterial coating exhibits an average mutual coherence of 0.929. (Details concerning the calculation of average mutual coherence and other quantitative characterization of the measurement matrix can be found in the Supporting Information.)

Fig. 1. (A) The 0.2-m radius structure of the metamaterial multispeaker listener and (B) one fan-like waveguide. (C) The frequency modulation of two fan-like waveguides obtained from simulation. (D) Fabricated prototype of the listener. (E) Measured amplitude patterns (normalized) of the measurement modes at 3,120, 4,320, and 5,240 Hz, respectively.

A multispeaker listening system should provide information about “who” is saying “what.” We thus design our sensing experiment as follows: Multiple sound sources simultaneously emit a sequence of independent audio messages (acting as a “conversation”). Each component of the conversation consists of 40 “words” randomly selected from a library containing 100 distinct but broadband synthesized pulses. The sound waves emitted from the sources first propagate in the free space and then are modulated by the encoding channels offered by the metamaterials, before they are collected as a single mixed waveform. In the data processing stage, the inversion algorithm segregates the mixed waveform and reconstructs the audio content of each source. The concept schematic of the measurement and reconstruction process is shown in Fig. 2. A Fourier component of the collected signal can be expressed as the superposition of the responses from all of the waveguides at this frequency:

\[
P_{s}(\omega) = \sum_{i=1}^{K} P_{i}(\omega),
\]

where \( P_{i}(\omega) \) is the response from the \( i \)th waveguide. The measured data vector used for reconstruction is

\[
g = [P_{1}(\omega_{1}) \ P_{2}(\omega_{2}) \ \cdots \ P_{K}(\omega_{M})]^T,
\]

and the object vector \( f \) is a scalar vector containing \( N = K \times P \) elements (\( K \) is the number of the possible locations and \( P \) is the size of the finite audio library). Because of the sparsity of \( f \) (only several elements are nonzero, corresponding to the activated sources), the sensing process is an ideal fit for the framework of compressive sensing. L1-norm regularization is performed with the Two-step Iterative Shrinkage/Thresholding (TwIST) algorithm (37) to solve the ill-posed inverse problem.

To examine the capability of the metamaterial sensing system in audio segregation, and ruling out other factors such as complex background [which may aid the reconstruction if they are well-characterized (38)], the tests were performed in an anechoic chamber as shown in Fig. 3. Three independent speakers were used as sources to emit words randomly selected from the predefined synthesized library. The measurement vector \( g \) used as the input for the algorithm contains 51 complex elements corresponding to the discretized frequency responses between 3,000 and 5,000 Hz with an interval of 40 Hz. Compared with the
prior knowledge that each source sends out one audio message at each time index is applied, we can define a recognized audio by selecting the message with the highest strength for each source at every time index. The recognition ratio can thus be calculated as the number of the recognized audio over the total number of the audio messages. For the case with metamaterials, the average recognition ratio for the three sources is 96.67%, whereas that for the case without metamaterials is close to zero. The results indicate that metamaterials contribute significantly in creating a forward model that aids the inversion of the sensing task.

Our proposed multispeaker listening system functionally mimics the selective listening capability of human auditory systems. The system employs only a single sensor, yet it can reconstruct the segregated signals with high fidelity. The device is also very simple and robust, as the passive metamaterial structure modulates the signal and, other than the microphone, no electronic or active components are used. The system proposed here does not rely on linguistic models or data mining algorithms (although it could be combined with such to extend its functionality) and has the advantages of low cost and low computational complexity. We also want to note that our demonstrated design does not reflect the mechanism of the cocktail party listening of human auditory systems, which is far more complicated and involves acoustic, cognitive, visual, as well as psychological factors (1–9).

In conclusion, we have demonstrated here an acoustic metamaterial-based multispeaker listener. Results of multiple-source audio segregation are demonstrated. We envision that it can be useful for multisource speech recognition and segregation, which are desired in many handheld, tabletop interactive devices. Besides, by extending such physical layer modulation approach to other applicable frequency ranges, we may expect other acoustic sensing and imaging applications such as hearing aid or ultrasound imaging.

Materials and Methods

The metamaterial listener prototype was fabricated with acrylonitrile butadiene styrene plastics using fused filament fabrication 3D printing technology. The design process was aided with a commercial full-wave simulation package COMSOL Multiphysics. Three-dimensional simulations with Pressure Acoustics Module were conducted to extract the frequency responses of all of the waveguides. The multispeaker listening experiment was performed in an anechoic chamber and multiple speakers used as audio sources were deployed on the floor of the chamber. Detailed discussions concerning the forward model derivation, the quality metric of the reconstruction, measurement matrix evaluation, the spatiotemporal degrees of freedom of the measurement modes, as well as the advantages of using metamaterials, can be found in the Supporting Information. The results of the controlled experiment without metamaterials and the multispeaker listening experiments with different configurations of sources can also be found in the Supporting Information.

ACKNOWLEDGMENTS. The authors thank Prof. Michael Gehm and Prof. Donald B. Bliss for their help. This work was supported by a Multidisciplinary University Research Initiative under Grant N00014-13-1-0631 from the Office of Naval Research.

Supporting Information

Xie et al. 10.1073/pnas.1502276112

SI Materials and Methods

Metamaterial-Based Resonances Design. The resonating dispersion of each waveguide is contributed by an array of Helmholtz resonators distributed within the fan-shaped area. Such side branch resonators are commonly used as acoustic filters (39). The resonating frequency can be approximated as $\omega_0 = c_v/\sqrt{h(h + \Delta h)}$, where $c_v$ is the speed of sound, $h$ is the height of the resonator, and $\Delta h$ is the effective height increase due to the opening. Fig. S1A (Inset) shows the dispersion of the transmission of a resonator numerically retrieved from simulation.

To aid the design of the sensing system, we model the resonator using lumped elements. For a side branch resonator indexed as $j$ in the $i$th waveguide, the complex transmission can be expressed as $T_{ji} = 1 - (R_0/(R_0 + r_j + j\omega_L + 1/j\omega_C))$. The $i$th waveguide is essentially a band-stop frequency filter network composed of an array of resonators, as shown in Fig. S1B. The resonances are randomized and contribute to a measurement matrix that favors compressive sensing, as will be explained in SI Text, Forward Model Derivation, below.

Experimental Setup. The layout of the experimental setup of the system is shown in Fig. S2. The metamaterial multispeaker listener is deployed on the floor of an anechoic chamber. The sensor at the center is an omnidirectional surface mount microelectromechanical microphone (Analog Devices ADMP401). The signal acquired by the microphone is sent to a preamplifier (Stanford Research SR560) before being digitized by a data acquisition device (NI PCI-6251). Multiple speakers are used as audio sources. They are placed on the floor of the chamber and mounted toward the ceiling so that the radiation of each is symmetrical with respect to the z axis (perpendicular to the floor).

A library of 100 broadband pulses were used as words for the multispeaker listening experiment. Each word has a length of 5 ms and the adjacent words are separated by a pause of 1 s to avoid overlapping of signals of different time indices. All of the words are synthesized by superposing randomized truncated cosine waves whose center frequencies span the spectral range of interest. The time-domain expression of a synthesized word can be expressed as $s_p(t) = \sum_{\alpha,m} a_{\alpha m} \cos(\alpha \omega_0 + \theta_{\alpha m})$, where $a_{\alpha m}$ and $\theta_{\alpha m}$ are computer-generated pseudorandom numbers within the range $a_{\alpha m} \in [0, 1]$ and $\theta_{\alpha m} \in [0, 2\pi]$, respectively. Two examples are shown in Fig. 3A (Inset) in the main text.

SI Text

Forward Model Derivation. Assuming the resonances in each waveguide are distributed sparsely over the interested frequency range and only first-order filtering responses dominate, the overall frequency modulation of a waveguide can be approximated by the multiplication of the individual responses of the resonators: $T_i(\omega) = \prod T_{ji}(\omega)$. For a source located at $\mathbf{r}_k$, the frequency response can be derived by propagating the waveguide responses from each waveguide aperture $\mathbf{r}_j$ to the source location $\mathbf{r}_k$: $P_i(\omega, \mathbf{r}_i, \mathbf{r}_k) = a(\omega) S_0(\omega) \sum_{\alpha,m} G(\omega, \mathbf{r}_i, \mathbf{r}_k) T_i(\omega) R(\omega, \mathbf{r}_i, \mathbf{r}_k)$, where $S_0(\omega)$ is the spectrum of the audio signal from the source, $R(\omega, \mathbf{r}_i, \mathbf{r}_k)$ is the waveguide radiation pattern which is mostly determined by the shape of the waveguide aperture, and $G(\omega, \mathbf{r}_i, \mathbf{r}_k) = e^{-j k |\mathbf{r}_i - \mathbf{r}_k|/|\mathbf{r}_i - \mathbf{r}_k|}$ is the Green’s function from the location $\mathbf{r}_i$ of the aperture of the $i$th waveguide to the location $\mathbf{r}_k$. The coefficient $a(\omega)$ includes all other factors such as sensor and speaker responses that are uniform for different source locations and audio signals.

Each column of the measurement matrix $H$ represents the discretized Fourier components of a source emitting an audio message from the predefined library from one of the possible locations in the scene. The number of columns of the matrix is $N = K \times P$, where $K$ is the number of possible locations and $P$ is the size of the audio library. An element in the measurement matrix can be expressed as

$$h_{mn} = P_i(\omega_m, \mathbf{r}_k, S_p) \langle a(\omega_m) S_p(\omega_m) \rangle \times \sum_{i=1}^{36} G(\omega_m, \mathbf{r}_i, \mathbf{r}_k) T_i(\omega_m) R(\omega_m, \mathbf{r}_i, \mathbf{r}_k).$$

Each row $H_m$ of the measurement matrix represents a test function for the object vector at one frequency, because a measurement value in the measured data vector is sampled in the way defined by the test function as $g_m = \langle f, H_m \rangle$, where the angle bracket indicates the inner product. The randomization of the measurement matrix for this sensing system is contributed by the carefully designed waveguide responses $T_i(\omega)$. Discussions on the quantified characterization of the measurement matrix are detailed in SI Text, Measurement Matrix Evaluation, below.

Reconstruction Quality Metric. The average MSE is used as a metric to quantify the reconstruction quality. It is defined as $1/K \sum_k \sum_\omega |f_k(\omega) - f_{\mathbf{g}}(\omega)|^2$, where $K$ is the number of all the possible source locations, $T$ is the maximum time index, $f_k(\omega)$ is the measurement function for the $k$th source at time index $\omega$, and $f_{\mathbf{g}}(\omega)$ is the ground-truth function for the same source at the time index. In our experiment, the nonzero values in the object vector (ground truth) are set to unity. However, nonunity relative amplitudes in the object vector in principle can still be reconstructed by the L1-norm regularization inverse algorithm (37).

Measurement Matrix Evaluation. The measurement matrix reflects the physical properties of the sampling process (such as the efficiency of phase and amplitude modulations and the spectral responses of the measurement). Here we discuss the performance of measurement matrix, including the resolution and the stability of the inverse estimation.

The resolution of the inverse problem is associated with the number of measurement $M$ and the sampling efficiency of those measurement modes. In compressive sensing, $M$ is less than the dimensionality of the object vector $N$, which means that without any prior information, the measurement cannot fully recover the scene. To overcome such limitation, enough independent sampling provided by the measurement matrix is required. A quality measurement matrix reduces overlapping of measurement modes and thus increases sampling efficiency, providing an opportunity for a reconstruction to outperform an undersampled measurement. With a reconstruction algorithm that uses proper regularization, an ill-conditioned compressive sensing might still be able to recover the scene.

Here we use the average mutual coherence to evaluate the independence of the modulation channels (36, 40). The mutual
coherence between two modulation channels \( \hat{h}_i \) and \( \hat{h}_j \) is \( W_{ij} = h_i h_j^*/|h_i||h_j| \). The average mutual coherence can be expressed as \( \mu_{ij} = \sum_{m=1}^{M} W_{ij}/m(m-1) \), which represents the average overlap between all \( m \) modulation channels. The average mutual coherence for the case with metamaterials is about 0.20, whereas that without metamaterials is about 0.93. The significantly reduced mutual coherence generally indicates high reconstruction fidelity and lower error (36).

The noise sensitivity of the measurement matrix can be considered as its contribution of noise amplification during the inverse process. Singular value decomposition can be used to analyze the measurement matrix to estimate the performance of the ill-conditioned inverse problem (34). Because small singular values cause the inverse process sensitive to the noise, a stable inverse problem should reduce the number of small singular values. One way to evaluate the performance has been inspired by the covariance matrix. The variance will be amplified by the sum of inverse square of the singular values \( \sum_j j^{-2} \). However, in an ill-posed inverse problem, such estimation will be dominated by those close-to-zero elements, which reduce the dynamic range of the evaluation. Practically, a threshold level can be set for singular value spectrum with the measured noise floor. The singular values above this threshold will be considered as effective and they dominate the sampling process. The number of effective singular values can thus provide a metric of the performance of the inverse problem.

The Spatiotemporal Degrees of Freedom of the Measurement Modes. The quality of the measurement modes can be quantitatively described by the number of the degrees of freedom (DOF) in both spatial and spectral domains (41). The spatial DOF along a dimension \( x \) can be estimated as \( X/\delta x \), where \( X \) is the total range and \( \delta x \) is the correlation distance along this dimension. Similarly the spectral DOF can be estimated as \( \Delta \omega/\delta \omega \), where \( \Delta \omega \) is the system bandwidth and \( \delta \omega \) is the correlation frequency. Using the physical model described in an earlier section (SI Text, Forward Model Derivation), we calculated the decorrelation for both frequency and the azimuth of the measurement modes. The results are presented in Fig. S3 A and B. If we define the correlation range where the normalized correlation coefficient drops to 0.5, the azimuth correlation angle is about 8°, whereas the correlation frequency is about 80 Hz. This indicates that the number of DOF of the azimuth is about 45 and that within the spectrum of interest is about 38.

There is a fundamental tradeoff between the quality factors of the physical layer modulations, the received signal power, and the DOF. On the one hand, the quality factors of the realized system will be lower than the ideal model and will thus decrease the DOF as suggested by Lipworth et al. (42). On the other hand, the modulations offered by metamaterials are essentially band-stop filtering. The higher the quality factors, the lower the received signals power. The interplay between these aspects needs to be tailored for further optimization of the sensing performance.

The Advantages of Using Metamaterials. Resonator-based metamaterials contribute several unique features to our sensing system: (i) the flexible material parameter dispersions allow us to design and realize the optimal measurement modes for the sensing tasks. In the demonstrated work, the randomized resonating dispersions create sufficient incoherent information channels; (ii) the metamaterials allow us to achieve required spatial complexity and frequency dependency (DOF) within much smaller size, compared with the random scatterers-based methods where the device size typically needs to be orders of magnitude larger than the wavelength; and (iii) the system performance can be estimated relatively easily with the effective parameters of the metamaterials using theoretical forward models and full-wave simulations.

Controlled Experiment Without Metamaterials. The control experiment of attempting to reconstruct three sources with only an omnidirectional microphone has an averaged MSE = 1.99, which is about 25x that of the reconstruction with metamaterial (MSE = 0.08). Both results are obtained using the sample experimental setup and with the same regularization parameters in the reconstruction algorithm. As is obvious from Fig. S4B, the reconstruction without metamaterial is too poor to provide useful information about the sources. The deficient reconstruction results from the more correlated modulations (transfer functions) for the three sources when the metamaterial coating is removed, as shown in Fig. S4A.

Reconstruction of Two Fixed Activated Sources out of Six Possible Locations. The task of this reconstruction is to identify the locations of two activated sources out of six possible fixed locations and at the same time segregate their overlapping signals. The reconstruction for the six locations exhibits two blocks (corresponding to the two activated sources) with dominant signal strength (shown in Fig. S5), whereas four others with negligible reconstructed signal strength are not shown. This designed task is to demonstrate a simple combination of signal segregation and localization of fixed sources. A successful recognition is counted when the content reconstructed with the highest reconstruction strength is above 50% and matches the ground-truth word for each activated source at each time index (whereas the maximum reconstruction is below 10% for each silence source). The overall recognition ratio calculated using such definition is 87.5% for this task.

Reconstruction of Randomly Emitting Sources out of Four Possible Locations. This task is to reconstruct the signals from four possible speakers. Unlike the previous task where the activated sources are fixed, in this task two or three speakers are randomly selected from the four possibilities at each time index to emit random words. The reconstruction results are shown in Fig. S6. Using the same definition of recognition as the previous task, the overall recognition ratio is 95%.
Fig. S1. (A) Fan-like waveguide with one resonator highlighted. (Inset) Simulation result of the amplitude of the transmission of the waveguide with only the contribution from the highlighted resonator. (B) Band-stop equivalent lumped elements network representation of the waveguide.

Fig. S2. (A) Schematic of the system layout. (B) Photograph of the data acquisition platform outside of the anechoic chamber.

Fig. S3. Azimuthal and spectral correlation curves estimated from the forward model. (A) Normalized correlation coefficient with respect to the shifted azimuthal angle. (B) Normalized correlation coefficient with respect to the shifted frequency.
Fig. S4. (A) Frequency responses measured for three sources with only an omnidirectional microphone. (B) Reconstruction results without metamaterials. (Left) Reconstruction. (Right) Ground truth. Out of 40 time indices, on average only 0.33 audios are correctly recognized (averaged over three sources) by comparing the audio of the maximum reconstructed strength with the ground truth.

Fig. S5. Reconstruction of two active resources out of six possible locations. (A) Schematic of the reconstruction task. Two red radiating circles indicate the sound-emitting sources, whereas the gray circles represent silent sources. (B) Reconstruction of two locations with significant reconstruction strength. The overall recognition ratio for this task is 87.5% using the criterion defined in SI Text.
Fig. S6. Reconstruction of two or three randomly selected active sources out of four possibilities. (A) Schematic of the reconstruction task. Four sources have the same possibility of radiating source and only two or three are randomly selected to emit sound at each time index. (B) Reconstruction of the four locations. The overall recognition ratio for this task is 95% using the criterion defined in SI Text.