Single-sensor multispeaker listening with acoustic metamaterials

Yangbo Xie, Tsung-Han Tsai, Adam Konneker, Bogdan-Ioan Popa, David J. Brady, and Steven A. Cummer[1]

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.

metamaterials | cocktail party problem | compressive sensing

The "cocktail party" or multispeaker listening problem is inspired by the remarkable ability of the human's auditory system in selectively attending to one speaker or audio signal in a multiple-speaker noisy environment (1, 2). Over the past half a century (3), the quest to understand the underlying mechanism (4–6) and build functionally similar devices has motivated significant research efforts (4–8).

Previously proposed engineered multispeaker listening systems generally fall into two categories. The first kind is based on audio features and linguistic models of speech. For example, harmonic characteristics, temporal continuity, onset/offset of speech units combined with hidden Markov language models can be used to group overlapping audio signals into different sources (7, 9, 10). The drawback of such an approach is that certain audio characteristics have to be assumed (e.g., nonoverlapping in spectrogram) and linguistic model-based estimation can be very computationally intensive. The second kind relies on multisensor arrays to spatially filter sources (11). The need for multiple transducers and system complexity are the major disadvantages of the second approach.

In this work, we demonstrate a multispeaker listening system that separates overlapping simultaneous conversations by leveraging the wave modulation capabilities of acoustic metamaterials. Acoustic metamaterials are a broad family of engineered materials which can be designed to possess flexible and unusual effective properties (12, 13). In the past, acoustic metamaterials with high anisotropy (14, 15), extreme nonlinearity (16), or negative dynamic parameters (density, bulk modulus, refractive index) (17–20) have been realized. Applications such as scattering reducing sound cloak (21, 22), beam steering metasurface (23), and other wave manipulating devices (24–27) have been proposed and demonstrated. We demonstrate here that acoustic metamaterials can also be useful for encoding independent acoustic signals coming from different spatial locations by creating highly frequency-dependent and spatially complex measurement modes (28), and aid the solution finding for the inverse problem. Such physical layer encoding scheme exploits the spatiotemporal degrees of freedom of complex media, which contribute to a variety of random scattering-based sensing and wave-controlling techniques (29–32) and a recently demonstrated radiofrequency metamaterial-based imager (33). The listening system we demonstrate here provides a hardware-based computational sensing method for functionally mimicking cocktail party listening.

Inspired by the frequency-dependent filtering mechanism of the human cochlea system (1), we designed our multispeaker listening system with carefully engineered metamaterials to perform dispersive frequency modulation. This modulation is produced by an array of Helmholtz resonators, whose heights determine their resonating frequencies. The sensing system is shown in Fig. 1. The single sensor at the center is surrounded by 36 fan-like waveguides that cover 360° of azimuth. Each waveguide possesses a unique and highly frequency-dependent response (two examples are plotted in Fig. 1*C*), which is generated by the resonators with randomly selected resonant dispersion. The randomized modulation from all of the waveguides "scrambles" the original omnidirectional measurement modes of the single sensor. As a result, the measurement modes are complex in both the spatial and spectral dimensions. For example, in Fig. 1*E*, three modes measured at different frequencies are shown. Such location-dependent frequency modulation provides both spatial and spectral resolution to the inversion task (34).

We can describe our sensing system with a general sampling model as $\boldsymbol{g} = \boldsymbol{H} \boldsymbol{f}$, where $\boldsymbol{g}$ is the vector form of the measured data (measurement vector); $\boldsymbol{f}$ is the object vector to be estimated. The measurement matrix $\boldsymbol{H}$, which represents the forward model of the sensing system, is formed by stacking rows of linear sampling vectors [also known as test functions (35)] at sequentially

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.

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

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1502276112/-/DCSupplemental.

APPLIED PHYSICAL SCIENCES

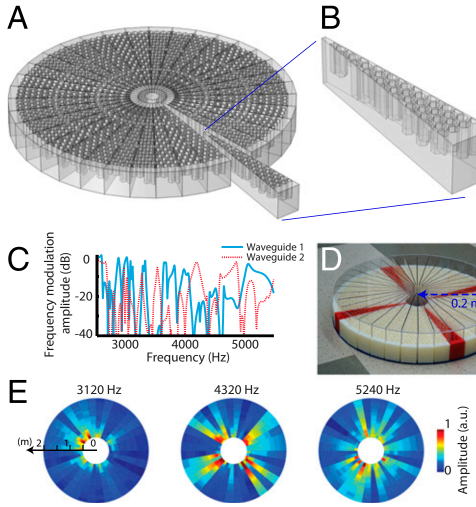

Fig. 1. (*A*) The 0.2-m radius structure of the metamaterial multispeaker listener and (*B*) one fan-like waveguide. (*C*) The frequency modulation of two fan-like waveguides obtained from simulation. (*D*) Fabricated prototype of the listener. (*E*) Measured amplitude patterns (normalized) of the measurement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

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 μ_{av} as the metric of the sensing performance (36). μ_{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*.

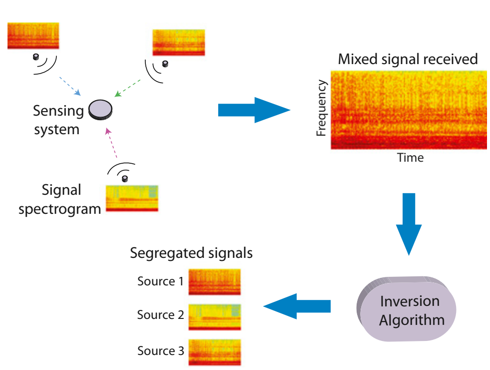

Fig. 2. Schematic of the measurement and reconstruction process.

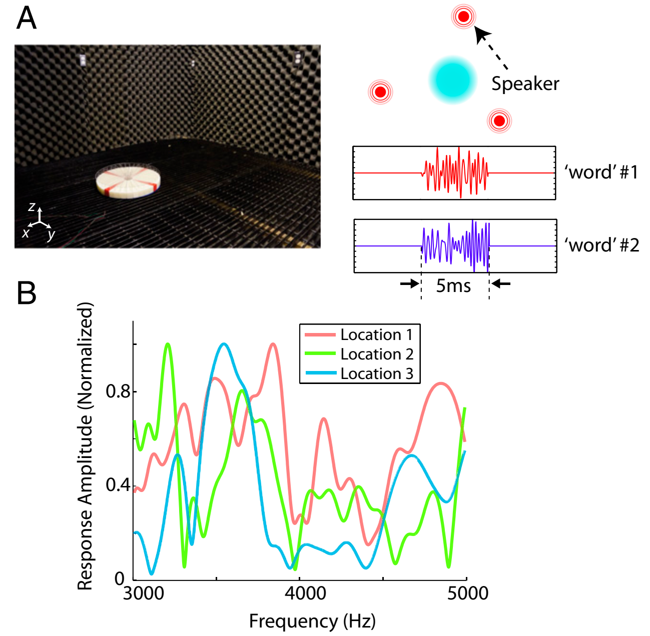

Fig. 3. (*A*) Measurement performed in an anechoic chamber. (*Left*) Photo of the metamaterial listener in the chamber. (*Right*) Schematic of the setup and two examples of synthesized word. (*B*) Measured transfer functions for the locations of three speakers.

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: $P_c(\omega) = \sum_{i=1}^{36} P_i(\omega)$, where $P_i(\omega)$ is the response from the ith waveguide. The measured data vector used for reconstruction is $\boldsymbol{g} = [P_c(\omega_1)\ P_c(\omega_2)\ \cdots\ P_c(\omega_M)]^T$, and the object vector $\boldsymbol{f}$ 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 $\boldsymbol{f}$ (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 $\boldsymbol{g}$ 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

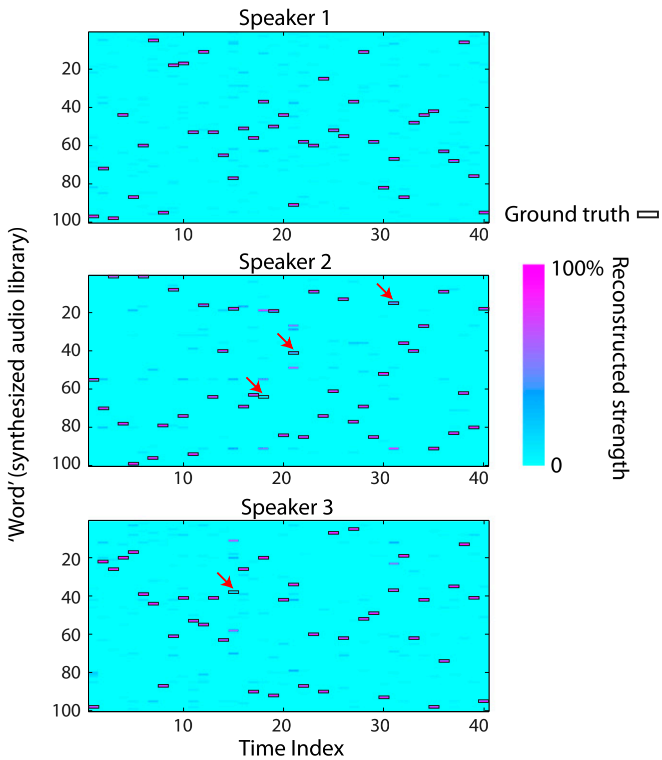

Fig. 4. Reconstruction results for three-speaker conversation. The black rectangles are ground truth and the purple color patches indicate the strength of reconstruction. Out of 40 time indices, on average 38.67 audios are correctly recognized (averaged over three sources) by comparing the audio of the maximum reconstructed strength with the ground truth. The incorrect recognitions are marked with red arrows.

300 source location–audio pair possibilities (possible combinations of 3 source locations and 100 broadband signals), a compression factor of about 6:1 is achieved.

The results shown in Fig. 4 exhibit the reconstruction for each source location–audio combination, where the more purple color indicates higher signal strength. The ground truth marked with black rectangular boxes indicates three overlapping simultaneous speeches in a conversation. The metamaterial listener provides a faithful reconstruction with an average MSE of about 0.08. In contrast, when the metamaterial coating is removed and only an omnidirectional sensor is used to collect the overlapping audio signals, the reconstruction is too poor to provide separated information about the sources (MSE = 1.99; see the *Supporting Information* for the results of the controlled experiment without metamaterials), which is expected as the transfer functions from the source locations to the sensor are less different (or more mutually coherent) from each other in the case without metamaterials. If 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.

1. Bregman AS (1994) *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press, Cambridge, MA).
2. Blauert J (1997) *Spatial Hearing: The Psychophysics of Human Sound Localization* (MIT Press, Cambridge, MA).
3. Cherry EC (1953) Some experiments on the recognition of speech, with one and with two ears. *J Acoust Soc Am* 25(5):975–979.
4. Formisano E, De Martino F, Bonte M, Goebel R (2008) "Who" is saying "what"? Brain-based decoding of human voice and speech. *Science* 322(5903):970–973.
5. Mesgarani N, Chang EF (2012) Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* 485(7397):233–236.
6. Bizley JK, Cohen YE (2013) The what, where and how of auditory-object perception. *Nat Rev Neurosci* 14(10):693–707.
7. Wang D, Brown GJ (2006) *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications* (Wiley-IEEE, Hoboken, NJ).
8. Rosenthal DF, Okuno HG (1998) *Computational Auditory Scene Analysis* (Lawrence Erlbaum Associates, Mahwah, NJ).
9. Cooke M (2005) *Modelling Auditory Processing and Organisation* (Cambridge Univ. Press, New York).
10. Rennie SJ, Hershey JR, Olsen PA (2010) Single-channel multitalker speech recognition. *IEEE Signal Process Mag* 27(6):66–80.
11. Brandstein M, Ward D (2001) *Microphone Arrays: Signal Processing Techniques and Applications* (Springer, New York).
12. Deymier PA (2013) *Acoustic Metamaterials and Phononic Crystals* (Springer, New York).
13. Craster RV, Guenneau S (2012) *Acoustic Metamaterials: Negative Refraction, Imaging, Lensing and Cloaking* (Springer, New York).
14. Li J, Fok L, Yin X, Bartal G, Zhang X (2009) Experimental demonstration of an acoustic magnifying hyperlens. *Nat Mater* 8(12):931–934.
15. García-Chocano VM, Christensen J, Sánchez-Dehesa J (2014) Negative refraction and energy funneling by hyperbolic materials: An experimental demonstration in acoustics. *Phys Rev Lett* 112(14):144301.
16. Popa B-I, Cummer SA (2014) Non-reciprocal and highly nonlinear active acoustic metamaterials. *Nat Commun* 5:3398.

17. Fang N, et al. (2006) Ultrasonic metamaterials with negative modulus. *Nat Mater* 5(6): 452–456.
18. Liang Z, Li J (2012) Extreme acoustic metamaterial by coiling up space. *Phys Rev Lett* 108(11):114301.
19. Xie Y, Popa B-I, Zigoneanu L, Cummer SA (2013) Measurement of a broadband negative index with space-coiling acoustic metamaterials. *Phys Rev Lett* 110(17): 175501.
20. Hladky-Hennion A-C, et al. (2013) Negative refraction of acoustic waves using a foam-like metallic structure. *Appl Phys Lett* 102(14):144103.
21. Zhang S, Xia C, Fang N (2011) Broadband acoustic cloak for ultrasound waves. *Phys Rev Lett* 106(2):024301.
22. Zigoneanu L, Popa B-I, Cummer SA (2014) Three-dimensional broadband omnidirectional acoustic ground cloak. *Nat Mater* 13(4):352–355.
23. Xie Y, et al. (2014) Wavefront modulation and subwavelength diffractive acoustics with an acoustic metasurface. *Nat Commun* 5:5553.
24. Lemoult F, Fink M, Lerosey G (2011) Acoustic resonators for far-field control of sound on a subwavelength scale. *Phys Rev Lett* 107(6):064301.
25. Lemoult F, Kaina N, Fink M, Lerosey G (2013) Wave propagation control at the deep subwavelength scale in metamaterials. *Nat Phys* 9(1):55–60.
26. Fleury R, Alù A (2013) Extraordinary sound transmission through density-near-zero ultranarrow channels. *Phys Rev Lett* 111(5):055501.
27. Ma G, Yang M, Xiao S, Yang Z, Sheng P (2014) Acoustic metasurface with hybrid resonances. *Nat Mater* 13(9):873–878.
28. Brady DJ (2009) *Optical Imaging and Spectroscopy* (Wiley, Hoboken, NJ).
29. Mosk AP, Lagendijk A, Lerosey G, Fink M (2012) Controlling waves in space and time for imaging and focusing in complex media. *Nat Photonics* 6(5):283–292.
30. Katz O, Small E, Silberberg Y (2012) Looking around corners and through thin turbid layers in real time with scattered incoherent light. *Nat Photonics* 6(8):549–553.
31. Lemoult F, Lerosey G, de Rosny J, Fink M (2009) Manipulating spatiotemporal degrees of freedom of waves in random media. *Phys Rev Lett* 103(17):173902.
32. Fink M, et al. (2000) Time-reversed acoustics. *Rep Prog Phys* 63(12):1933–1995.
33. Hunt J, et al. (2013) Metamaterial apertures for computational imaging. *Science* 339(6117):310–313.
34. Aster RC, Borchers B, Thurber CH (2013) *Parameter Estimation and Inverse Problems* (Academic, Waltham, MA).
35. Duarte MF, et al. (2008) Single-pixel imaging via compressive sampling. *IEEE Signal Process Mag* 25(2):83–91.
36. Duarte-Carvajalino JM, Sapiro G (2009) Learning to sense sparse signals: Simultaneous sensing matrix and sparsifying dictionary optimization. *IEEE Trans Image Process* 18(7):1395–1408.
37. Bioucas-Dias JM, Figueiredo MA (2007) A new twIst: Two-step iterative shrinkage/thresholding algorithms for image restoration. *IEEE Trans Image Process* 16(12): 2992–3004.
38. Carin L, Liu D, Guo B (2008) In situ compressive sensing. *Inverse Probl* 24(1):015023.
39. Kinsler LE, Frey AR, Coppens AB, Sanders JV (1999) *Fundamentals of Acoustics* (Wiley, New York).
40. Elad M (2007) Optimized projections for compressed sensing. *IEEE Trans Signal Process* 55(12):5695–5702.
41. Derode A, Tourin A, Fink M (2001) Random multiple scattering of ultrasound. II. Is time reversal a self-averaging process? *Phys Rev E Stat Nonlin Soft Matter Phys* 64(3): 036606.
42. Lipworth G, et al. (2013) Metamaterial apertures for coherent computational imaging on the physical layer. *J Opt Soc Am A Opt Image Sci Vis* 30(8):1603–1612.