# Supporting Information for *"Robot Hearing Through Optical Channel in a Cocktail Party Environment"*

Xiao Guo[1], Siyi Ding[1], peng ti[1], Kenan Li[1], and Xiaoping Hong[1]

[1]Southern university of science and technology

June 8, 2022

**Abstract**

This Supporting Information includes: a comparison of the REAL (Robot Ear Accomplished by Laser) with a typical vibration measuring system (Laser Doppler Vibrometers, LDV), frequency response of various materials on REAL and real-time analysis of REAL audio neural network model.

Xiaoping Hong Email: *hongxp@sustech.edu.cn*

## Comparison of Robot Ear Accomplished by Laser (REAL) and Laser doppler vibrometers (LDV)

A typical scheme of LDV is shown in the literature.[1] LDV is based on the principle of laser interference, which requires the laser (He-Ne) with good monochromaticity and phase stability. The detection scheme of LDV is optical heterodyne detection, which requires the reference light with an MHz-level frequency shift produced through an acoustic-optic modulator (AOM). As a result, LDV, composed of high-performance phase-stable lasers, interference optics, AOMs, high-power drivers and complex high-frequency modulation and demodulation circuits, is much more delicate and expensive than REAL, which needs only a stable laser pointer, optical lenses and signal amplification circuits.

## Analysis on LDV with rough surfaces

For the LDV, the reference laser and signal laser must be temporally and spatially coherent across the detecting surface. In many application scenarios, the signal laser is reflected from optically rough surfaces or passes through optically translucent media leading to the destruction of the coherent wavefront. This is known as speckle,[2] which significantly damages the quality of the LDV signal. Noise contribution from speckles is analyzed,[3] and it is demonstrated that the speckle is indeed the largest noise across different frequency bands.[4-6]

Here we analyse the influence of the LDV signal caused by the rough surface such as masks or throat surfaces. According to the principle of LDV, the total electric field entering the LDV detector is

$$E = E_1 \cos[?](2\pi f_1 t + f_1) + E_2 \cos[?](2\pi f_2 t + f_2(x, y))$$

Where $E$ is the total electric field entering the detector, $E_1$ and $E_2$ are the amplitudes of electric fields of signal light and reference light, $f_1$ is the frequency of the signal, which is the sum of AOM modulation frequency, the laser frequency and the Doppler shift caused by the movement of the surface. $f_2$ is the frequency of the reference, which equals the laser frequency. Therefore $f_1 - f_2 = f_{AOM} + f_{Doppler}$. $f_1$ is the phase of the signal, and $f_2(x, y)$ is the phase of the reference. The spatial dependence of $f_2(x, y)$ is related to the roughness of the measured surface. Because $f_1$ and $f_2$ are at optical frequency band and beyond the detector's response, only the cross-term is detected. The output of the detector has the following expression

$$S(t) = kI(t) = kE^2[?][?]E_1E_2\cos[2\pi(f_s1 - f_s2)t + f_1 - f_2(x, y)]dxdy$$

where k is a coefficient reflecting the response of the detector. We use the profile roughness parameter Ra[7] to randomly simulate a rough surface to get a typical behaviour of $f_2(x, y)$ and

$S(t)$. The signal $I(t)$ is calculated when $f_{AOM}$ is 10 MHz, $f_{Doppler}$ is 0.7 MHz (the speed of surface movement is 0.2 m/s), $f_1$ is zero without loss of generality, wavelength of laser is 532 nm,

$dx$ and $dy$ are 1 um, and the diameter of the laser spot is 2 mm. The relationship between the signal to noise ratio (SNR) of $S(t)$ and the roughness Ra can be simulated, as shown in **Figure S1**. The SNR of the LDV is considerably diminished for common surface roughness ($> 5$ um). The average SNR is roughly 1.4 dB, which is very low considering only phase noise is counted. Shot noises from both the reference and the signal will further reduce the SNR of LDV.

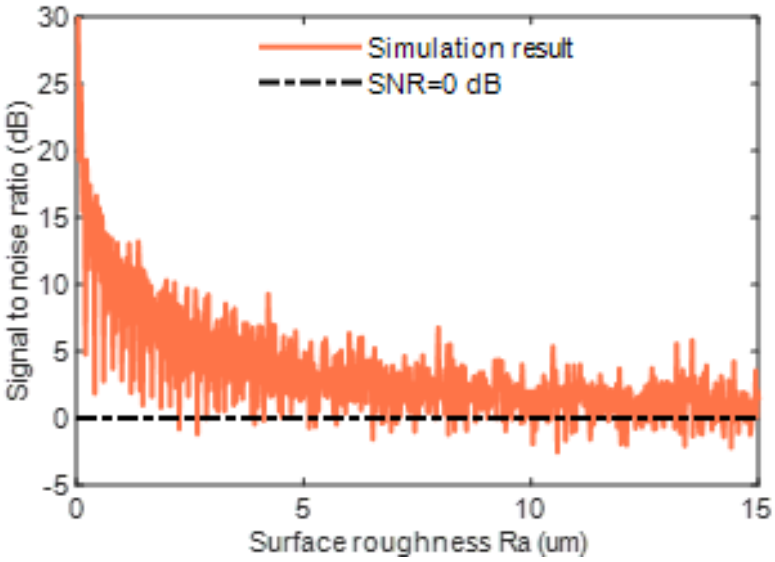



Figure 1:

**Figure S1**. Signal to noise ratio of LDV on different roughness surfaces considering speckle noises only.

In summary, compared with the sophisticated structure and principle, high cost, inevitable speckle noise of LDV, REAL has the advantages of simpler construction, lower cost, high SNR for rough surfaces and miniaturization readiness to become an ideal audio capturing device in a cocktail party environment.

# Frequency response of various materials on REAL

In addition to the mask and throat surface, REAL can measure the vibration of plastic bags, papers etc. The optical characteristics and vibration characteristics of the surfaces directly determine the REAL measurement results. The scattering characteristics of the surfaces will determine the collected laser power and hence the SNR of REAL. The vibration characteristics including frequency-dependent damping of the surfaces are key to the quality of REAL audios. The frequency response of mask, plastic bag and paper (**Figure S2**b) on REAL is studied (Figure S2a). These three surfaces are excited by a loudspeaker playing chirp audio (0.1˜7 kHz). In Figure S2c, the frequency responses measured by REAL on three respective surfaces is shown. It can be seen that there are sharp peaks and valleys at some frequencies, which are related to the resonances of sound waves on these surfaces. Both plastic bag and paper perform better than mask, possibly due to the more elastic nature.

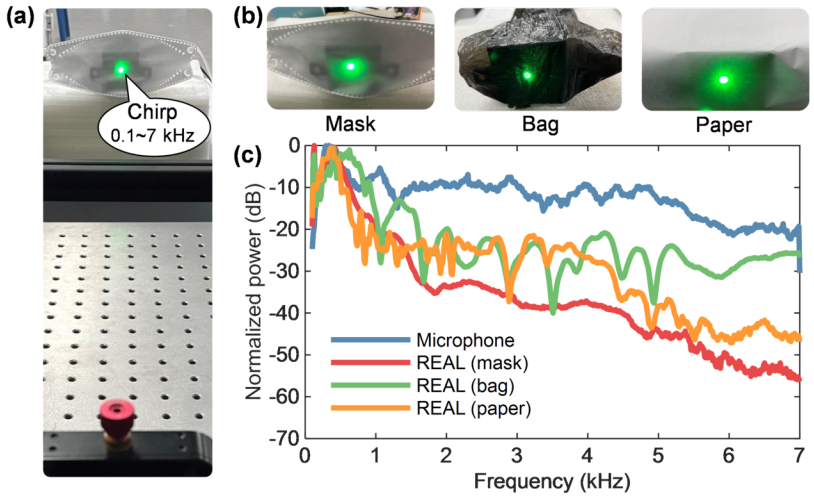


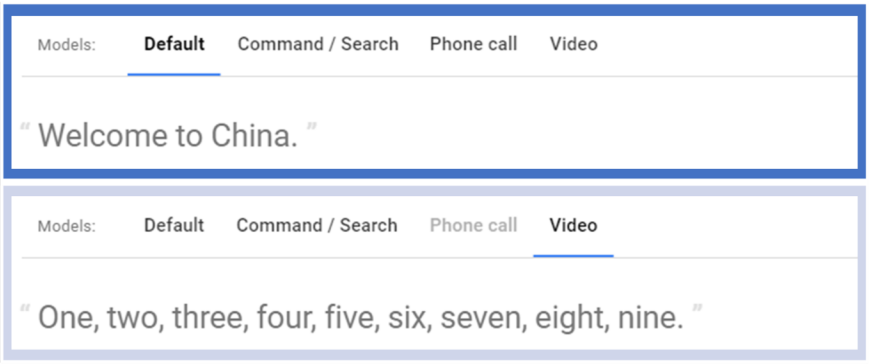

Figure 2:

**Figure S2.** Frequency response of various materials on REAL. a) REAL collects the vibration of the mask, which is on a loudspeaker playing a chirp audio (0.1˜7 kHz). b) Various measured materials: mask, plastic bag and paper. c, Power spectrum of REAL responses from these materials.

# Real-time analysis of REAL audio neural network model

In inference, the average time cost for the model to estimate 3-second enhanced audio is 0.08 seconds using an Nvidia 2080s GPU, and 0.78 seconds if it runs on an i7-10700K CPU. Therefore, the model can be deployed in real scenarios with affordable computing cost and low latency, to perform inference in a streaming fashion using the sliding window strategy, which constantly processes and estimates new incoming frames of REAL audio.

**Figure S3**. Speech recognition results of direct REAL audios from masks (supplementary audios) through Google speech-to-text platform.




## Supplementary Audios

Rich media available at https://youtu.be/hfqAlmqkSY4

**Supplementary Audio 1** 'Welcome to China' recorded by REAL on a speaker's mask

Rich media available at https://youtu.be/Ii2poHel_mk

**Supplementary Audio 2** Numbers recorded by REAL on a speaker's mask

Rich media available at https://youtu.be/JSjwBNGJ_x0

**Supplementary Video 1** Voice collection in a noisy environment (mask)

Rich media available at https://youtu.be/4gBnQVdHu04

**Supplementary Video 2** Voice collection in a noisy environment (throat)

## References

[1] E. Esposito, *Laser Doppler Vibrometry in Handbook of the Use of Lasers in Conservation and Conservation Science*, COST Office, Brussels, GER **2008**.

[2] J. C. Dainty, *Laser speckle and related phenomena.* Springer science & business Media, Berlin, GER **2013**.

[3] T. Pfister, A. Fischer, J. Czarske, *Meas. Sci. Technol.* **2011**, *22*, 055301.

[4] C. A. Hill, M. Harris, K. D. Ridley, E. Jakeman, P. Lutzmann, *Appl. Optics* **2003**, *42*, 1091-1100.

[5] L. A. Jiang, M. A. Albota, R. W. Haupt, J. G. Chen, R, M. Marino, *Appl. Optics* **2011**, *50*, 2263-2273.

[6] T. Lv, X. Han, S. Wu, Y. Li, *Opt. Commun.* **2019**, *440*, 117-125

[7] T. Thomas, *Precis. Eng.* **1981**, *3*, 97-104.