INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Synthesis of Device-Independent Noise Corpora for Realistic ASR Evaluation

Hannes Gamper[1], *Mark R. P. Thomas*[1], *Lyle Corbin*[2], *Ivan Tashev*[1]

[1]Microsoft Research Redmond
[2]Microsoft Corporation

{hagamper,lylec,ivantash}@microsoft.com, mark.r.thomas@ieee.org

Abstract

In order to effectively evaluate the accuracy of automatic speech recognition (ASR) with a novel capture device, it is important to create a realistic test data corpus that is representative of real-world noise conditions. Typically, this involves either recording the output of a device under test (DUT) in a noisy environment, or synthesizing an environment over loudspeakers in a way that simulates realistic signal-to-noise ratios (SNRs), reverberation times, and spatial noise distributions. Here we propose a method that aims at combining the realism of in-situ recordings with the convenience and repeatability of synthetic corpora. A device-independent spatial recording containing noise and speech is combined with the measured directivity pattern of a DUT to generate a synthetic test corpus for evaluating the performance of an ASR system. This is achieved by a spherical harmonic decomposition of both the sound field and the DUT's directivity patterns. Experimental results suggest that the proposed method can be a viable alternative to costly and cumbersome device-dependent measurements. The proposed simulation method predicted the SNR of the DUT response to within about 3 dB and the word error rate (WER) to within about 20%, across a range of test SNRs, target source directions, and noise types.

Index Terms: automatic speech recognition, device characterization, device-related transfer function, spherical harmonics

1. Introduction

Automatic speech recognition (ASR) is an integral part of many hardware devices, including mobile phones, game consoles and smart televisions, to enable hands-free operation and voice control. When evaluating how robust a device's ASR engine is to noise and reverberation in real world settings, one must account for the device hardware characteristics as well as typical usage scenarios. This is normally achieved by exposing the device under test (DUT) to realistic conditions in terms of environmental noise and reverberation while evaluating the performance of the ASR engine. Such in-situ tests are extremely valuable when tuning hardware and software parameters to maximize ASR performance, especially if the DUT has multiple microphones and the effect of microphone placement and (spatial) speech enhancement algorithms needs to be assessed.

However, in-situ tests are lengthy and cumbersome, requiring hours of recordings made on the DUT that have to be redone whenever hardware changes are made to the DUT. Furthermore, the exact test conditions are difficult or impossible to recreate when attempting to evaluate the effect of a hardware change or comparing the performance of an ASR engine across devices.

In order to overcome the limitations of in-situ tests, a pre-recorded test corpus can be used. An ASR test corpus typically consists of a variety of scenarios–differing in the level, type and spatial quality of the background noise–that are representative of the conditions to which a DUT might be subjected in real-world use. During testing, the corpus is rendered over a loudspeaker setup and recorded through the DUT. Two examples of pre-recorded test corpus methodologies are encoded as part of specifications by the European Telecommunications Standards Institute (ETSI) [1, 2]. Both techniques utilize multichannel recordings that are played back over rigorously calibrated speaker systems. The systems attempt to recreate the original sound field of the real world environment for a device placed at the center of the playback system.

Song et al. describe a method for simulating realistic background noise to test telecommunication devices based on spatial sound field recordings from a spherical microphone array [3]. The authors compare various methods to derive the input signals to a circular loudspeaker array delivering the spatial noise recording to the DUT. While using a pre-recorded test corpus has the advantage of repeatability and simpler logistics compared to in-situ tests, it requires a highly specialized test environment and hardware setup for playback and recording. In addition, emulating the complexity of a real, noisy environment, with potentially hundreds of spatially distributed noise sources, can be challenging.

Here we propose a method that combines the realism of in-situ tests with the convenience and repeatability of a pre-recorded corpus without the requirement of a specialized playback setup. The approach is based on a device-independent spatial in-situ recording that is combined with the directivity characteristics of a DUT to create a synthetic test corpus for ASR performance and robustness evaluation. The DUT directivity characteristics can be obtained through measurements or via acoustic simulation. In a similar fashion to the work by Song et al. [3], the proposed method is based on a spherical harmonics decomposition of a spatial noise recording obtained using a spherical microphone array. However, by also using a spherical harmonic decomposition of the DUT's directivity pattern, we show that it is possible to simulate the DUT response directly without performing actual recordings on the DUT.

2. Proposed approach

The goal of the proposed method for generating a synthetic test corpus is to simulate the response of a DUT to a pre-recorded noisy environment. Section 2.1 describes the device-*independent* capture and representation of a sound field, while section 2.2 discusses obtaining the device-*dependent* directivity characteristics of the DUT. The proposed approach for combining device-independent recordings with device-dependent directivity characteristics in the spherical harmonics domain is presented in Section 2.3.

http://dx.doi.org/10.21437/Interspeech.2016-978

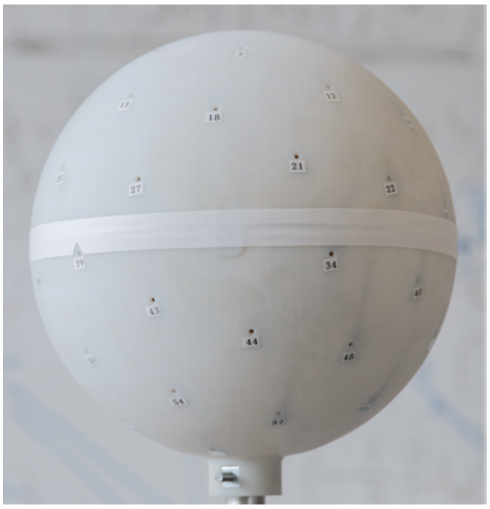

Figure 1: 64-channel spherical microphone array, allowing a 7-th order spherical harmonic approximation to the recorded sound field.

2.1. Sound field capture and decomposition

Real noisy environments typically contain a multitude of spatially distributed noise sources. In order to evaluate the ASR performance under realistic conditions it is important to subject the DUT to spatially diverse noise environments. Therefore, the spatial quality of the noise environment is preserved in recordings used for ASR evaluation. A common way to capture a sound field spatially is to use an array of microphones placed on the surface of a sphere. A 64-channel example with radius 100 mm is shown in Figure 1.

Spherical harmonics provide a convenient way to describe a sound field captured using a spherical microphone array. By removing the scattering effect of the microphone baffle, the free-field decomposition of the recorded sound field can be estimated. Given the microphone signals $P(r_0, \theta, \phi, \omega)$, where r_0 is the array radius, θ and ϕ are the microphone colatitude and azimuth angles, respectively, and ω is the angular frequency, the plane wave decomposition of the sound field captured with a spherical array of M microphones, distributed uniformly on the surface of the sphere [4], is given in the spherical harmonics domain by [5, 6]

$$\breve{S}_{nm}(\omega) = \frac{1}{b_n(kr_0)} \frac{4\pi}{M} \sum_{i=1}^{M} P(r_0, \theta_i, \phi_i, \omega) Y_n^{-m}(\theta_i, \phi_i), \tag{1}$$

where $k = \omega/c$ and c is the speed of sound. The spherical harmonic of order n and degree m is defined as

$$Y_n^m(\theta, \phi) = (-1)^m \sqrt{\frac{2n+1}{4\pi} \frac{(n-|m|)!}{(n+|m|)!}} P_n^{|m|}(\cos\theta) e^{im\phi}, \tag{2}$$

where the associated Legendre function P_n^m represents standing waves in θ and $e^{im\phi}$ represents travelling waves in ϕ. Note that Condon-Shortley phase convention is used such that $Y_n^m(\theta, \phi)^* = Y_n^{-m}(\theta, \phi)$ [7].

In the case of a spherical scatterer, the mode strength

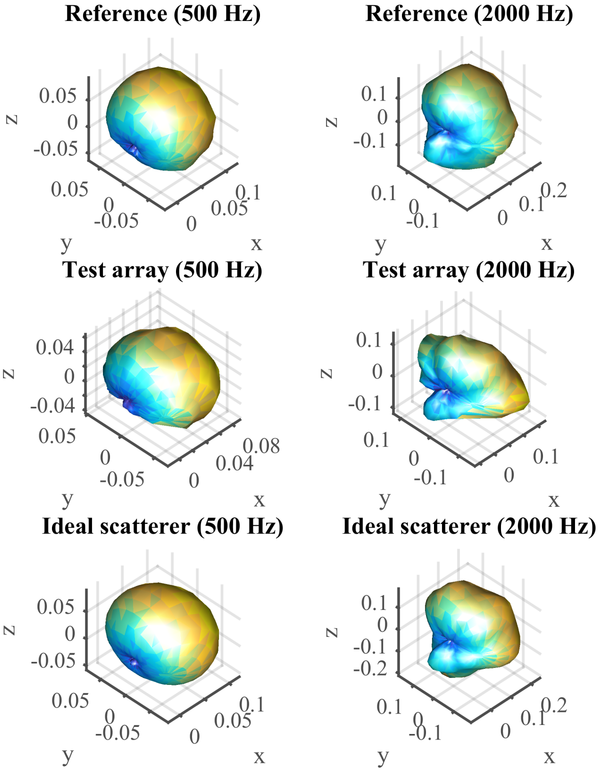
Reference (500 Hz)
Reference (2000 Hz)
Test array (500 Hz)
Test array (2000 Hz)
Ideal scatterer (500 Hz)
Ideal scatterer (2000 Hz)

Figure 2: DUT directivity patterns: measured (top) and simulated through 7th-order spherical microphone array (middle) and ideal 7th-order rigid spherical scatterer (bottom).

$b_n(kr_0)$ is defined for an incident plane wave as

$$b_n(kr_0) = 4\pi i^n \left(j_n(kr_0) - \frac{j'_n(kr_0)}{h'^{(2)}_n(kr_0)} h^{(2)}_n(kr_0) \right), \tag{3}$$

where $j_n(kr_0)$ is the spherical Bessel function of degree n, $h^{(2)}_n(kr_0)$ is the spherical Hankel function of the second kind of degree n, and $(\cdot)'$ denotes differentiation with respect to the argument. The mode strength term in (1) is necessary to account for the scattering effect of the spherical baffle in order to obtain a plane-wave decomposition of the sound field.

2.2. Characterising the device under test (DUT)

Under the assumption of linearity and time invariance, the response of the DUT to an input signal is given by a transfer function describing the acoustic path from the sound source to the microphone. In far field conditions, that is, when the source is further than approximately one meter from the DUT, this transfer function varies spectrally with source azimuth and elevation, whereas the effect of source distance is mostly limited to the signal gain. Therefore, the directivity characteristics of the DUT can be approximated through transfer function measurements at a single distance, spanning the range of azimuth and elevation angles of interest. Figure 2 (top) shows the directivity patterns of one microphone of a Kinect device [8]. Alternatively, acoustic simulation can be used to estimate these transfer functions. Due to the similarity of these direction-dependent transfer functions to the concept of head-related transfer functions (HRTFs) in the field of spatial audio rendering [9], we

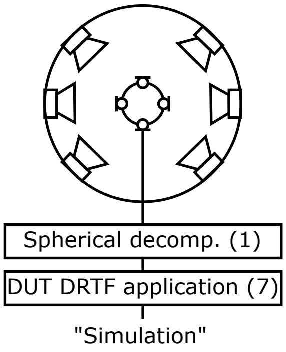

Figure 3: Experimental setup.

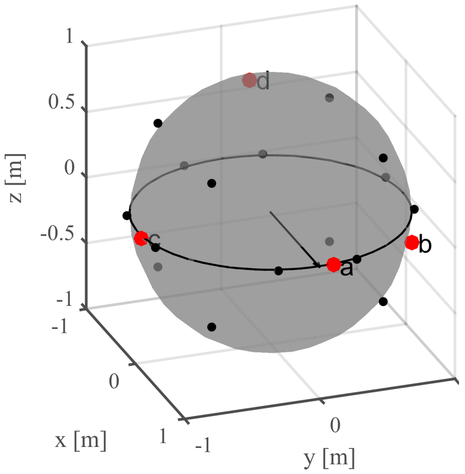

Figure 4: Geometric layout of noise sources (black dots) and speech sources (red dots) at 5.6 degrees azimuth and 0 degrees elevation (a), 63.7 degrees azimuth and -10.4 degrees elevation (b), -84.4 degrees azimuth and 0 degrees elevation (c), and 172.1 degrees azimuth and 44.7 degrees elevation (d).

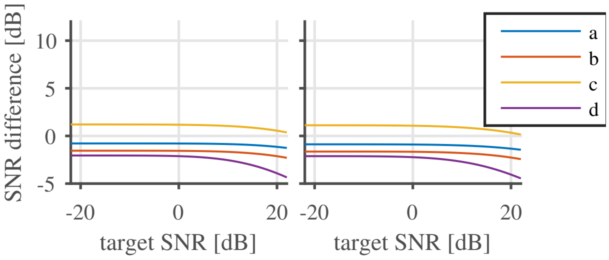

Figure 5: SNR difference between simulation and reference, for brown noise (left) and market noise (right). Labels a–d indicate the speech source locations labelled a–d in Figure 4.

refer to the direction-dependent DUT transfer functions as the *device-related transfer functions (DRTFs)*.

In analogy to spherical microphone array recordings, DRTFs measured at points uniformly distributed over the sphere can be decomposed using spherical harmonics:

$$\breve{\mathcal{D}}_{nm}(\omega) = \frac{4\pi}{N}\sum_{i=1}^{N} D(\theta_i,\phi_i,\omega)Y_n^{-m}(\theta_i,\phi_i), \qquad (4)$$

where N is the number of DRTFs and $D(r,\theta_i,\phi_i,\omega)$ are the DRTFs as a function of the measurement colatitude and azimuth angles of arrival, θ and ϕ. In cases where the DRTF measurement points do not cover the whole sphere or are not uniformly distributed, a least-squares decomposition can be used [10].

2.3. Combining sound field recordings and DUT directivity

To simulate the DUT behavior in the recorded noise environment, the device-related transfer functions (DRTFs) of the DUT are applied to the spherical array recording. This can be conveniently performed in the spherical harmonics domain, in analogy to applying head-related transfer functions to a recording for binaural rendering [11]. An aperture weighting function derived from the DRTFs is applied to the estimated free-field decomposition of the recorded sound field. The sound pressure at each microphone of the DUT is then found by integrating the DRTF-weighted pressure over the sphere:

$$\begin{aligned} P(\omega) &= \int_{\Omega\in S^2} S(\Omega,\omega)\mathcal{D}(\Omega,\omega)d\Omega && (5)\\ &= \sum_{n=-\infty}^{\infty}\sum_{n'=-\infty}^{\infty}\sum_{m=-n}^{n}\sum_{m'=-n'}^{n'} \breve{S}_{nm}(\omega)\breve{\mathcal{D}}_{n'm'}(\omega) \\ &\qquad \int_S Y_n^m(\Omega)Y_{n'}^{m'}(\Omega)d\Omega && (6)\\ &= \sum_{n=-\infty}^{\infty}\sum_{m=-n}^{n} \breve{S}_{nm}(\omega)\breve{\mathcal{D}}_{n,-m}(\omega). && (7) \end{aligned}$$

3. Experimental evaluation

The experimental setup (see Figure 3) consisted of a spherical microphone array (see Figure 1) and a Kinect device as the DUT. Impulse responses of both the spherical array and the DUT were measured in an anechoic chamber for 400 directions at a radius of one meter, using a setup described by Bilinski et al. [12]. For the resulting DUT DRTFs, extrapolation was used to cover the whole sphere [10]. Figure 2 (top) illustrates the directivity patterns of one microphone of the DUT. Figure 2 (middle) depicts the directivity patterns equivalent to applying the DUT DRTFs to a spherically isotropic sound field recorded via the spherical array. Each point in the directivity patterns in Figure 2 (middle) is obtained by decomposing the N spherical array impulse responses for that direction using (1) and applying the DUT DRTF via (7). As the spherical array shown in Figure 1 does not behave like an ideal scatterer with ideal microphones, the sound field decomposition is imperfect and the resulting equivalent DUT directivity slightly distorted compared to the actual, measured DUT DRTF. As shown in Figure 2 (bottom), this distortion is largely corrected when replacing the real array impulse responses in the simulation with those of an ideal scatterer [6]. A follow-up study [15] addresses the discrepancy between real and ideal scatterer through calibration of the array and by deriving optimal scatterer removal functions [13, 14].

Simulations were performed combining speech recordings with simulated and recorded spatial noise. The speech corpus consisted of 2000 utterances containing 14146 words recorded by 50 male and female speakers in a quiet environment, for a total duration of 2.5 hours. Speech recognition was performed using a DNN based ASR engine [16] with acoustic models trained on the clean speech corpus. Two types of noise were used: 60

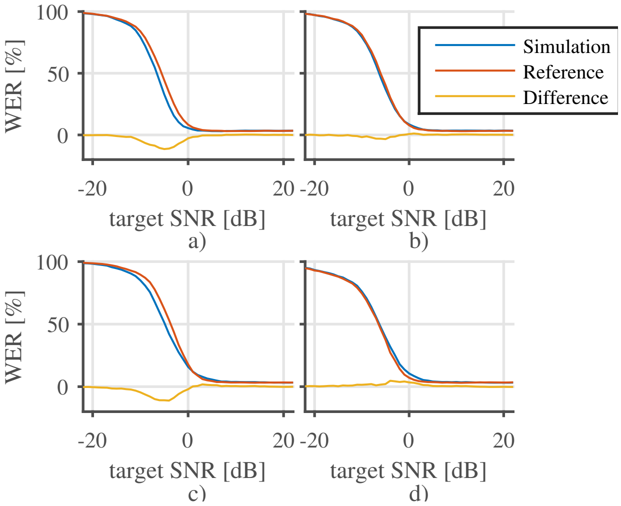

Figure 6: Word error rates (WERs) for brown noise. Labels a–d indicate the speech source locations labelled a–d in Figure 4.

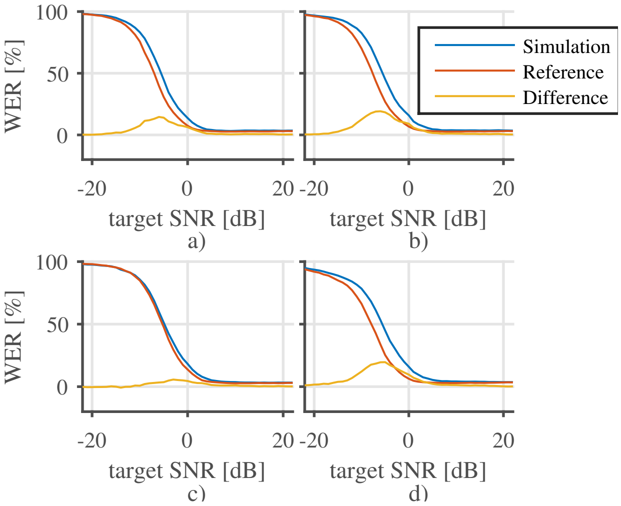

Figure 7: Word error rates (WERs) for market noise. Labels a–d indicate the speech source locations labelled a–d in Figure 4.

seconds of random Gaussian noise with a frequency roll-off of 6 dB per octave (i.e., *brown noise*), and a 60 second recording of ambient noise in a busy outdoor market place, obtained with the spherical microphone array. For the experiments, the spherical harmonics decomposition of the recorded sound field was evaluated at 16 directions, as shown in Figure 4, to emulate playback over 16 spatially distributed virtual loudspeakers. The number of virtual speakers was chosen as a trade-off between spatial fidelity and computational complexity.

The output of both the spherical array and the DUT was simulated by convolving virtual source signals with the corresponding measured impulse responses. The virtual source signals were derived by extracting a pseudo-random segment of the noise data and mapping it to a virtual source direction. Similarly, the speech recordings were mapped to one of four virtual source directions, to simulate a speaker embedded in noise. The noise and speaker locations used are shown in Figure 4. Note that the setup includes locations off the horizontal plane to emulate the spatial diversity found in real environments.

Tests were performed by combining the simulated noise and speech responses at a target signal-to-noise ratio (SNR). The SNR was calculated in the frequency band 100–2000 Hz, as the energy ratio between the microphone response during speech activity (*signal*) and the response in absence of speech (*noise*). For each target SNR, appropriate noise and speech gains were derived by simulating the DUT response via (7), i.e., the proposed method (“simulation” in Figure 3). Those same gains were then applied to the noise and speech samples convolved directly with the DUT DRTFs (“reference” in Figure 3). This experiment evaluates how closely the SNR estimated from the simulated DUT response matches the SNR of the reference response. As shown in Figure 5, the mismatch between simulated and reference SNRs is within ±5 dB across the tested target SNRs, source directions, and noise types, with lower errors for low target SNRs and speech directions closer to the front (a and b). For target SNRs below 10 dB, the predicted SNRs are actually within ± 3 dB. Above 10 dB SNR background noise in the raw speech recordings may start to affect results.

The simulation and reference responses of the DUT to the noisy speech samples generated for the SNR experiment described above were fed to the ASR engine. Figures 6 and 7 show the resulting average word error rates (WERs). The simulated corpus predicts the reference WERs fairly accurately across SNRs and source directions, except around -5 dB where the WER is most sensitive to the SNR. For brown noise, the simulation underestimates the WER for source directions a and c, whereas for market noise, the simulation overestimates the WER for directions a, b and d. This estimation bias could be explained by SNR mismatches between simulated and reference responses, as illustrated in Figure 5. However, the WER change as a function of SNR is predicted fairly well by the simulation.

4. Summary and conclusion

The proposed method allows the use of a device-independent spatial noise recording to generate a device-specific synthetic speech corpus for automatic speech recognition performance evaluation under realistic conditions. Experimental results indicate that the proposed method allows predicting the expected signal-to-noise ratio (SNR) of a device under test (DUT) exposed to spatial noise to within about ± 3 dB. The mismatch between simulation and reference SNRs may be reduced by applying a calibration and appropriate optimal scatterer removal functions to the spherical microphone array used for the spatial noise recordings. The prediction of average word error rates (WERs) was accurate to within about 20%. While estimation bias may have affected absolute WER prediction, the proposed method predicted WER change as a function of SNR fairly well. This indicates that the method may be well suited to evaluate the relative effect of hardware changes on ASR performance.

One limitation of the method is the assumption of far-field conditions, i.e., that all sound sources are further than approximately one meter from the DUT. However, in a close-talk situation with a target source in the vicinity of the DUT, the method may still prove useful for evaluating the effect of ambient noise on ASR performance. A major advantage of the proposed method is that it allows running completely simulated experiments. Here, speech recognition was performed on 2.5 hours of speech data for two noise types, four speech directions, and over 40 target SNRs. Collecting this data using live recordings would have taken $2.5 \times 2 \times 4 \times 40 = 800$ hours. Future work includes verification of the method in live noise environments.

5. References

[1] *Speech and multimedia Transmission Quality (STQ); A sound field reproduction method for terminal testing including a background noise database*, ETSI EG 202 396-1 Std., 2015.

[2] *Speech and multimedia Transmission Quality (STQ); Speech quality performance in the presence of background noise; Part 1: Background noise simulation technique and background noise database*, ETSI TS 103 224 Std., 2011.

[3] W. Song, M. Marschall, and J. D. G. Corrales, "Simulation of realistic background noise using multiple loudspeakers," in *Proc. Int. Conf. on Spatial Audio (ICSA)*, Graz, Austria, 2015.

[4] J. Fliege and U. Maier, "A two-stage approach for computing cubature formulae for the sphere," in *Mathematik 139T, Universität Dortmund, Fachbereich Mathematik, 44221*, 1996.

[5] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*, 1st ed. London: Academic Press, 1999.

[6] B. Rafaely, "Analysis and design of spherical microphone arrays," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 1, pp. 135–143, 2005.

[7] N. A. Gumerov and R. Duraiswami, *Fast Multipole Methods for the Helmholtz Equation in Three Dimensions*. Elsevier, 2004.

[8] "Kinect for Xbox 360," http://www.xbox.com/en-US/xbox-360/accessories/kinect.

[9] C. I. Cheng and G. H. Wakefield, "Introduction to head-related transfer functions (HRTFs): Representations of HRTFs in time, frequency, and space," in *Proc. Audio Engineering Society Convention*, New York, NY, USA, 1999.

[10] J. Ahrens, M. R. Thomas, and I. Tashev, "HRTF magnitude modeling using a non-regularized least-squares fit of spherical harmonics coefficients on incomplete data," in *Proc. APSIPA Annual Summit and Conference*, Hollywood, CA, USA, 2012.

[11] L. S. Davis, R. Duraiswami, E. Grassi, N. A. Gumerov, Z. Li, and D. N. Zotkin, "High order spatial audio capture and its binaural head-tracked playback over headphones with HRTF cues," in *Proc. Audio Engineering Society Convention*, New York, NY, USA, 2005.

[12] P. Bilinski, J. Ahrens, M. R. P. Thomas, I. J. Tashev, and J. C. Platt, "HRTF magnitude synthesis via sparse representation of anthropometric features," Florence, Italy, 2014, pp. 4501–4505.

[13] S. Moreau, J. Daniel, and S. Bertet, "3D sound field recording with higher order ambisonics - objective measurements and validation of spherical microphone," in *Proc. Audio Engineering Society Convention 120*, Paris, France, 2006.

[14] C. T. Jin, N. Epain, and A. Parthy, "Design, optimization and evaluation of a dual-radius spherical microphone array," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 193–204, 2014.

[15] H. Gamper, L. Corbin, D. Johnston, and I. J. Tashev, "Synthesis of device-independent noise corpora for speech quality assessment," in *Proc. Int. Workshop on Acoustic Signal Enhancement (IWAENC)*, Xi'an, China, 2016.

[16] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Proc. Interspeech*, Florence, Italy, 2011, pp. 437–440.