SYNTHESIS OF DEVICE-INDEPENDENT NOISE CORPORA FOR SPEECH QUALITY ASSESSMENT

Hannes Gamper, Lyle Corbin, David Johnston, Ivan J. Tashev

Microsoft Corporation, One Microsoft Way, Redmond, WA 98052, USA

ABSTRACT

The perceived quality of speech captured in the presence of background noise is an important performance metric for communication devices, including portable computers and mobile phones. For a realistic evaluation of speech quality, a device under test (DUT) needs to be exposed to a variety of noise conditions either in real noise environments or via noise recordings, typically delivered over a loudspeaker system. However, the test data obtained this way is specific to the DUT and needs to be re-recorded every time the DUT hardware changes. Here we propose an approach that uses device-independent spatial noise recordings to generate device-specific synthetic test data that simulate in-situ recordings. Noise captured using a spherical microphone array is combined with the directivity patterns of the DUT, referred to here as device-related transfer functions (DRTFs), in the spherical harmonics domain. The performance of the proposed method is evaluated in terms of the predicted signal-to-noise ratio (SNR) and the predicted mean opinion score (PMOS) of the DUT under various noise conditions. The root-mean-squared errors (RMSEs) of the predicted SNR and PMOS are on average below 4 dB and 0.28, respectively, across the range of tested SNRs, target source directions, noise types, and spherical harmonics decomposition methods. These experimental results indicate that the proposed method may be suitable for generating device-specific synthetic corpora from device-independent in-situ recordings.

***Index Terms*—** Speech quality, PMOS, PESQ, DRTF, spherical harmonics, microphone array, noise corpus

1. INTRODUCTION

Mobile and portable communication devices are being used in a large variety of acoustic environments. An important evaluation criterion for speech devices or processing algorithms is their performance in the presence of background noise. To evaluate various noise conditions, a device under test (DUT) can either be placed in a real noise environment for an in-situ recording, or subjected to synthetic noise environments delivered over a set of loudspeakers. While in-situ recordings may offer the most realistic test conditions, they can be cumbersome to obtain and typically cannot be controlled or repeated. Playing back noise signals over a loudspeaker array allows creating synthetic scenarios with specific noise conditions, including the signal-to-noise ratio (SNR) and the spatial distribution of noise and target sources. However, modelling complex real environments containing potentially hundreds of spatially distributed sources can be challenging.

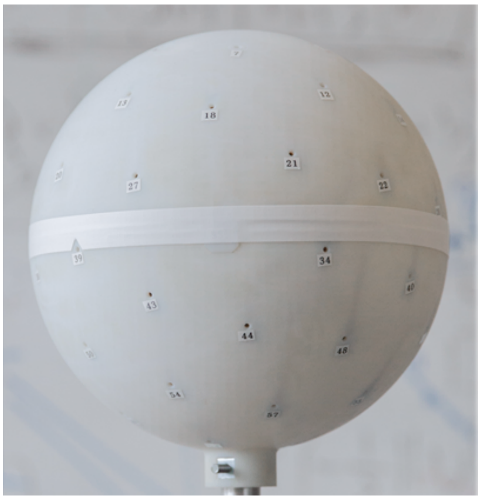

Fig. 1. 64-channel spherical microphone array.

To recreate actual noise environments as accurately as possible, the European Telecommunications Standards Institute (ETSI) specifies test methodologies that employ multichannel microphone and loudspeaker arrays to capture and reproduce real noise environments [1, 2]. Song et al. propose using a spherical microphone array to record a noise environment and deliver it to a DUT over a set of loudspeakers [3].

In previous work, the generation of a device independent noise corpus using a spherical microphone array (see Figure 1) for evaluating the performance of automatic speech recognition (ASR) on a DUT was introduced [4]. The approach aims at combining the realism of in-situ recordings with the convenience and controllability of a synthetic noise corpus. Here, the approach is extended for the evaluation of perceived speech quality. Experiments are conducted to assess the predicted mean opinion score (PMOS), estimated using the ITU-T P.862 Perceptual Evaluation of Speech Quality (PESQ) [5], of a DUT recording and its simulation.

2. PROPOSED METHOD

The proposed approach aims at simulating the perceived quality of speech recorded by a DUT in a noisy environment.

2.1. Sound field capture and decomposition

A convenient way to capture a sound field spatially is through a spherical microphone array [6]. Figure 1 shows the array used here, consisting of 64 digital MEMS microphones mounted on the surface of a rigid sphere of 100 mm radius.

Assume the microphone signals $P(\theta, \phi, \omega)$, where θ and ϕ are the microphone colatitude and azimuth angles and ω is the angular frequency, captured by M microphones uniformly distributed on the surface of a sphere [7]. Their plane wave decomposition can be represented using spherical harmonics [8, 6] as:

$$S_{nm}(\omega) = \frac{1}{b_n(kr_0)} \frac{4\pi}{M} \sum_{i=1}^{M} P(\theta_i, \phi_i, \omega) Y_n^{-m}(\theta_i, \phi_i), \quad (1)$$

where r_0 is the sphere radius, c is the speed of sound, and $k = \omega/c$. The spherical mode strength, $b_n(kr_0)$, is defined for an incident plane wave as:

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

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 complex spherical harmonic of order n and degree m is given as

$$Y_n^m(\theta, \phi) = (-1)^m \sqrt{\frac{2n+1}{4\pi} \frac{(n-|m|)!}{(n+|m|)!}} P_n^{|m|}(\cos\theta) e^{im\phi}, \quad (3)$$

where the associated Legendre function P_n^m represents standing waves in θ and $e^{im\phi}$ represents travelling waves in ϕ.

2.2. Characterising the DUT and spherical array

To simulate the response of the device under test (DUT) to a noise environment with the proposed method, its acoustic properties need to be measured. Assuming linearity, time invariance, and far field conditions, the directivity of the DUT microphones can be determined via impulse response measurements from loudspeakers positioned at a fixed distance and discrete azimuth and elevation angles, in an anechoic environment. Due to the similarity to the concept of head-related transfer functions (HRTFs) describing the directivity characteristics of a human head [9], we use the term *device-related transfer functions (DRTFs)* to describe the frequency-dependent DUT directivity patterns.

Similarly, the acoustic properties of the microphone array can be determined and used for calibration purposes or to derive spherical harmonics decomposition filters, as described in the next section.

2.3. Deriving spherical harmonics decomposition filters

Given the order-N plane wave decomposition of a sound field, $S(\omega)$, the acoustic pressure at the i-th array microphone, $\hat{P}(\theta_i, \phi_i, \omega)$, can be reconstructed via [10]:

$$\begin{aligned} \hat{P}(\theta_i, \phi_i, \omega) &= \sum_{n=0}^{\mathrm{N}} \sum_{m=-n}^{n} S_{nm}(\omega) b_n(kr_0) Y_n^m(\theta_i, \phi_i) && (4) \\ &= \mathbf{t}_{\mathrm{N},i}^{\mathrm{T}} \mathbf{S}_{\mathrm{N}} && (5) \end{aligned}$$

where

$$\begin{aligned} \mathbf{S}_{\mathrm{N}} &= \left[S_{0,0}(\omega), S_{1,-1}(\omega), S_{1,0}(\omega), \cdots, S_{\mathrm{N},\mathrm{N}}(\omega)\right]^{\mathrm{T}}, && (6) \\ \mathbf{t}_{\mathrm{N},i} &= \left[t_{0,0,i}, t_{1,-1,i}, t_{1,0,i}, \cdots, t_{\mathrm{N},\mathrm{N},i}\right]^{\mathrm{T}}, && (7) \\ t_{n,m,i} &= b_n(kr_0) Y_n^m(\theta_i, \phi_i). && (8) \end{aligned}$$

Note that from here on the dependence on ω is dropped for convenience of notation. For all microphones, this can be formulated as

$$\mathbf{P} = \mathbf{T}_{\mathrm{N}} \mathbf{S}_{\mathrm{N}}, \quad (9)$$

where

$$\mathbf{T}_{\mathrm{N}} = \left[\mathbf{t}_{\mathrm{N},1}, \mathbf{t}_{\mathrm{N},2}, \cdots, \mathbf{t}_{\mathrm{N},\mathrm{M}}\right]^{\mathrm{T}}. \quad (10)$$

The matrix $\mathbf{T}_{\mathrm{N}}$ relates the pressure recorded at the array microphones to the spherical harmonics, $\mathbf{S}_{\mathrm{N}}$. Spherical harmonics encoding filters, $\mathbf{E}$, are found by inverting $\mathbf{T}_{\mathrm{N}}$, e.g., via Tikhonov regularisation [10]:

$$\mathbf{E}_{\mathrm{L}} = \mathbf{T}_{\mathrm{L}}^{\mathrm{H}} \left(\mathbf{T}_{\mathrm{N}} \mathbf{T}_{\mathrm{N}}^{\mathrm{H}} + \beta^2 \mathbf{I}_{\mathrm{M}}\right)^{-1}, \quad (11)$$

where $\mathrm{L} \leq \mathrm{N}$ is the desired spherical decomposition order, typically dictated by the array geometry [10]. Note that lowering the desired order L toward higher frequencies ($kr_0 > 4$) may be considered to reduce spatial aliasing [11].

Given a matrix of measured array responses, $\mathbf{G}$, (9) becomes:

$$\mathbf{G} = \hat{\mathbf{T}}_{\mathrm{N}} \hat{\mathbf{S}}_{\mathrm{N}}, \quad (12)$$

where $\hat{\mathbf{S}}$ is composed of the expected spherical harmonic decompositions of unit amplitude plane waves incoming from the loudspeaker directions, θ_u and ϕ_u at radius r_u [10]:

$$\hat{S}_{nm} = e^{-ikr_u} Y_n^m(\theta_u, \phi_u). \quad (13)$$

Then, $\mathbf{T}_{\mathrm{N}}$ is derived as:

$$\hat{\mathbf{T}}_{\mathrm{N}} = \mathbf{G} \hat{\mathbf{S}}_{\mathrm{N}}^{\mathrm{H}} \left(\hat{\mathbf{S}}_{\mathrm{N}} \hat{\mathbf{S}}_{\mathrm{N}}^{\mathrm{H}} + \beta^2 \mathbf{I}_{(\mathrm{N}+1)^2}\right), \quad (14)$$

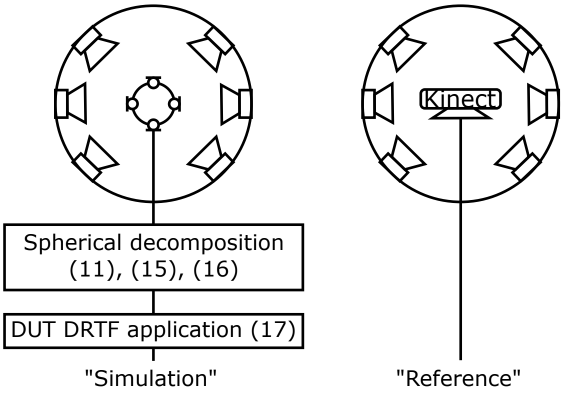
Spherical decomposition (11), (15), (16)
DUT DRTF application (17)
"Simulation"
Kinect
"Reference"

Fig. 2. Experimental setup.

and inverted using Tikhonov regularisation:

$$\hat{\mathbf{E}}_{\mathrm{L}} = \hat{\mathbf{T}}_{\mathrm{L}}^{\mathrm{H}} \left(\hat{\mathbf{T}}_{\mathrm{N}} \hat{\mathbf{T}}_{\mathrm{N}}^{\mathrm{H}} + \hat{\beta}^2 \mathbf{I}_{\mathrm{M}} \right)^{-1}. \tag{15}$$

In this work, $\beta = \hat{\beta} = 1$.

Alternatively, the decomposition filters can be derived from the measured array directivity using [12]

$$\hat{\mathbf{E}}_{\mathrm{L}} = \hat{\mathbf{S}}_{\mathrm{L}}^{\mathrm{T}} \mathrm{diag}(\mathbf{w}) \mathbf{G}^{\mathrm{H}} (\mathbf{G} \mathrm{diag}(\mathbf{w}) \mathbf{G}^{\mathrm{H}} + \lambda \mathbf{I})^{-1}, \tag{16}$$

where $\mathrm{diag}(\mathbf{w})$ is a diagonal matrix of weights accounting for the non-uniform distribution of the loudspeaker locations, $\mathbf{w} = [w_0, w_1, ..., w_{\mathrm{U}}]$ and $\sum_i w_i = 1$. Here, the weights are calculated from the areas of Voronoi cells associated with each location [13].

2.4. Simulating the DUT response

The response of the DUT to a sound field can be simulated by applying the DRTFs of the DUT to the sound field recording in the spherical harmonics domain. Note that this process is similar to binaural rendering in the spherical harmonics domain using head-related transfer functions [14].

Given a sound field recording from a spherical microphone array in the time domain, the estimated free-field decomposition, S_{nm}, is obtained via fast convolution in the frequency domain with the decomposition filters described in Section 2.3. The DUT response is simulated by applying the DUT directivity via the DRTF, $\breve{\mathcal{D}}_{n,-m}$, and integrating over the sphere [4]:

$$\hat{P} = \sum_{n=-\infty}^{\infty} \sum_{m=-n}^{n} S_{nm} \breve{\mathcal{D}}_{n,-m}. \tag{17}$$

3. EXPERIMENTAL EVALUATION

Experiments were conducted using the spherical microphone array shown in Figure 1 and a Kinect device [15] as the DUT.

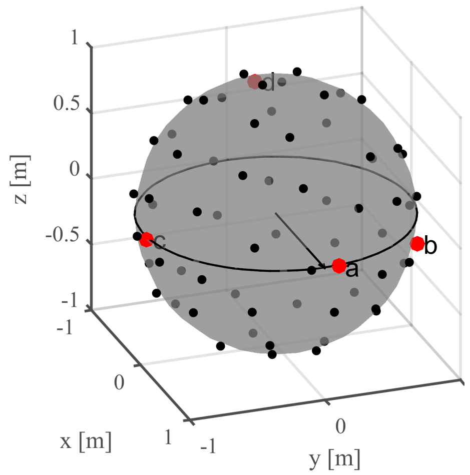

Fig. 3. Geometric layout of noise sources (black dots) and speech sources (red dots) at 5.6 degrees azimuth and 0 degrees elevation (a), 63.7 degrees azimuth and -10.4 degrees elevation (b), -84.4 degrees azimuth and 0 degrees elevation (c), and 172.1 degrees azimuth and 44.7 degrees elevation (d).

The experimental setup is depicted in Figure 2. Impulse response measurements were carried out for both the array and the DUT in an anechoic environment [16]. Two measurement runs, one with the DUT and array mounted upside down, were combined for a total of 512 measurement positions covering the sphere. The test data consisted of 50 short utterances from one male and one female speaker. Two noise types were used, random Gaussian noise with a 6 dB per octave roll-off (*brown noise*), and a sound field recording of a noisy outdoor market obtained with the spherical microphone array shown in Figure 1. Noise was rendered at 64 of the impulse response measurement directions approximating a uniform spatial distribution [7], either directly using 64 brown noise samples or by evaluating a spherical harmonics decomposition of the market noise recording at the 64 noise directions, shown in Figure 3.

Synthetic recordings were obtained by convolving the measured array and DUT impulse responses corresponding to the desired source and noise directions with the speech and noise samples. To simulate the DUT response, the DUT DRTF was applied to a 4th-order spherical decomposition of the synthetic array recordings via (17). From the simulated DUT response the SNR was estimated as the ratio between speech and noise energy in the range 100 to 2000 Hz. Given the estimated SNR, gains were derived for the synthetic speech and noise recordings to combine them at a target SNR, yielding the simulated DUT response (*simulation*). Those same gains were then used to combine the synthetic DUT noise and speech recordings (*reference*), yielding the reference SNR. The difference between the reference SNR and the simulation SNR provides a measure of the error predicting the DUT SNR via the simulated DUT response. The

	RMSE of SNR [dB]								RMSE of PESQ score							
	Brown noise				Market noise				Brown noise				Market noise			
	a	b	c	d	a	b	c	d	a	b	c	d	a	b	c	d
Eq. (11)	1.46	1.82	1.37	2.22	1.68	1.43	3.25	3.04	0.09	0.12	0.08	0.14	0.19	0.19	0.25	0.24
Eq. (15)	1.49	1.61	1.30	1.61	2.23	2.13	3.61	3.93	0.11	0.11	0.04	0.16	0.22	0.21	0.25	0.18
Eq. (16)	1.52	1.77	1.21	1.74	1.81	1.63	2.92	3.72	0.08	0.11	0.07	0.09	0.21	0.22	0.25	0.28

Table 1. Root-mean-squared errors of SNR and PMOS estimations, for the source direction a–d (see Figure 3).

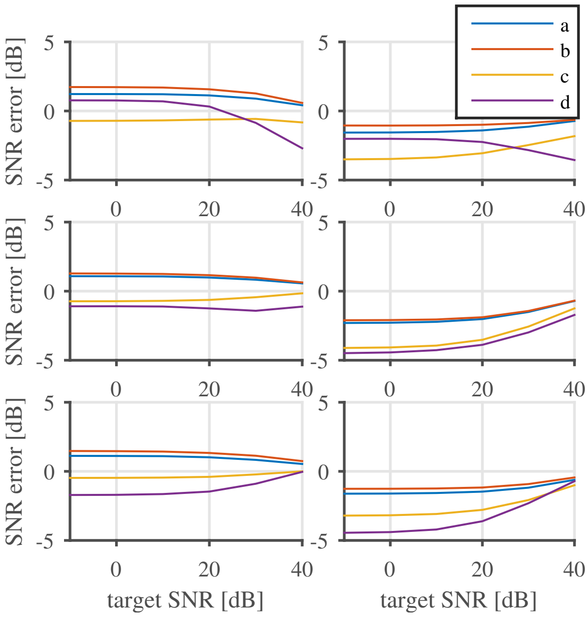

Fig. 4. SNR errors for brown noise (left) and market noise (right), for the three spherical decomposition methods: top: (11); middle: (15); bottom: (16). Labels a–d indicate the speech source locations labelled a–d in Figure 3.

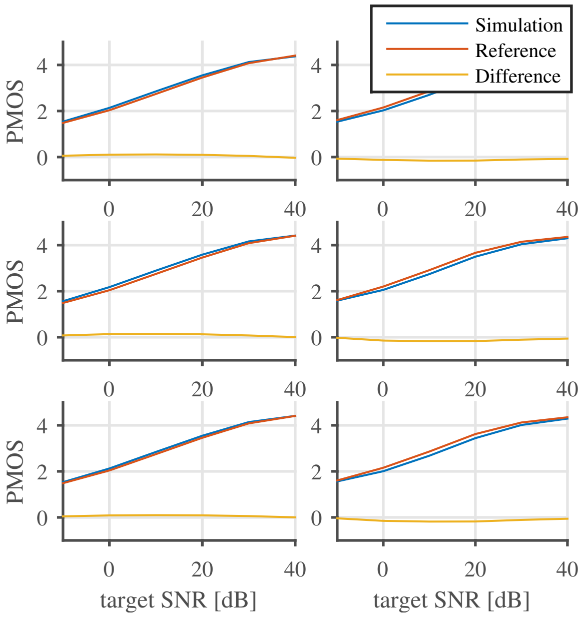

Fig. 5. PMOS estimates for brown noise (left) and market noise (right), for the source direction labelled a in Figure 3 and the three tested spherical decomposition methods: top: (11); middle: (15); bottom: (16).

SNR estimation errors across the range of tested target SNRs, for all noise types, target speech directions, and spherical decomposition methods are illustrated in Figure 4. As can be seen, the SNRs are estimated to within 5 dB across test conditions. The differences between the tested spherical decomposition methods indicate that there may be room for improvement by tuning the decomposition parameters.

The degradation of the simulation and reference samples in terms of perceived speech quality as a result of the additive background noise was evaluated via the Predicted Mean Opinion Score (PMOS), ranging from -0.5 to 4.5, implemented via the ITU-T P.862 Perceptual Evaluation of Speech Quality (PESQ) [5]. A comparison of PMOSs estimated for simulation and reference for one source direction is shown in Figure 5. The PMOS calculated for the simulation matches the PMOS of the reference quite well across test conditions. Table 1 summarises the root-mean-squared errors (RMSEs) of the SNR and PMOS estimations. The results indicate that the differences between the various spherical decomposition methods are marginal, despite the differences in the SNR estimates, and that the market noise condition proved more challenging, resulting in higher error rates.

4. CONCLUSION

The proposed method allows generating device-specific synthetic test corpora for speech quality assessment using device-independent spatial noise recordings. Experimental results indicate that the Predicted Mean Opinion Score (PMOS) of a device under test (DUT) in noisy conditions can be estimated reasonably well. An advantage of the experimental framework used here is that generation and evaluation of the synthetic test corpus can be done significantly faster than real time, as no actual recordings are performed on the DUT or the array. Future work is needed to evaluate the proposed method under echoic conditions and in real noise environments.

5. REFERENCES

[1] ETSI TS 103 224, "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," 2011.

[2] ETSI EG 202 396-1, "Speech and multimedia transmission quality (STQ); a sound field reproduction method for terminal testing including a background noise database," 2015.

[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, Sep 2015.

[4] H. Gamper, M. R. P. Thomas, L. Corbin, and I. J. Tashev, "Synthesis of device-independent noise corpora for realistic ASR evaluation," in *Proc. Interspeech*, San Francisco, CA, USA, Sep 2016.

[5] ITU-T P.862, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," Feb. 2001.

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

[7] 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.

[8] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*, Academic Press, London, first edition, 1999.

[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, Sep 1999.

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

[11] J. Meyer and G. W. Elko, "Handling spatial aliasing in spherical array applications," in *Proc. Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, Trento, Italy, May 2008, pp. 1–4.

[12] 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, May 2006.

[13] A. Politis, M. R. P. Thomas, H. Gamper, and I. J. Tashev, "Applications of 3D spherical transforms to personalization of head-related transfer functions," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Brisbane, Australia, Mar 2016, pp. 306–310.

[14] 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, Oct 2005.

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

[16] P. Bilinski, J. Ahrens, M. R. P. Thomas, I. J. Tashev, and J. C. Platt, "HRTF magnitude synthesis via sparse representation of anthropometric features," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, May 2014, pp. 4501–4505.