INDIVIDUAL DISTANCE-DEPENDENT HRTFS MODELING THROUGH A FEW ANTHROPOMETRIC MEASUREMENTS

Mengfan Zhang, Xihong Wu, Tianshu Qu

Key Laboratory on Machine Perception (Ministry of Education), Speech and Hearing Research Center, Peking University, Beijing, China, qutianshu@pku.edu.cn

ABSTRACT

The lack of data is a major problem in individual HRTF modeling. There are many HRTF databases, but each database only has limited HRTFs with different characteristics, such as distance-dependent HRTFs or individual HRTFs. How to effectively model HRTFs through several different databases is an important task. In this paper, a method for predicting individual distance-dependent HRTFs using a few anthropometric parameters is proposed. By modeling the HRTFs in CIPIC database, which contains individual HRTFs in 1 meter, and the PKU&IOA database, which contains KEMAR HRTFs in eight distances, we predict the individual HRTFs in arbitrary directions and distances. The objective experiments show that the proposed model has less spectral distortions than distance variation function model. The subjective experiments show that the proposed model can predict the individual HRTFs in arbitrary directions and distances.

Index Terms— HRTF, SPCA, DNN, anthropometric parameters

1. INTRODUCTION

In recent years, spatial auditory display has gained attention in both academic research and practical applications. To realize the fidelity and immersive experience in binaural audio reproduction, Head Related Transfer Functions (HRTF) are often used as filters describing the sound transmission from a sound source to the listeners' eardrum. It is difficult to measure the high spatial resolution HRTFs for each potential user, so non-individual HRTFs are often used to achieve spatial audio system at present. However, this may lead to some perception errors such as in-head localization, front-back confusion, and a breakdown of elevation discrimination ability. Therefore, it is important to obtain individual HRTFs with high spatial resolutions.

In recent years, more and more researchers have concentrated on modeling individual HRTFs. Numerical calculation methods including the boundary element method (BEM) [1], the finite element method (FEM) [2], and the finite difference method (FDM) [3] can be used to model individual HRTFs. However, these methods are computationally expensive and depend on the availability of precise 3D geometry. The individual HRTFs can also be obtained based on the listeners' feedback. Fink et al. [4] let subjects tune the PCA weights from average HRTFs to obtain individual HRTFs. Nevertheless, the tuning procedure is very time-consuming for each potential user. Xie [5] used spatial principal component analysis (SPCA) to recover individual HRTFs from measurements at a few spatial directions; however, measuring individual HRTFs in a few directions is still a non-trivial task.

A more feasible way is using a few anthropometric parameters to model individual HRTFs. Zotkin et al. [6] selected the HRTF data of the subject whose anthropometric parameters are closest to the new subject. Hu et al. [7] applied the principal component analysis (PCA) to HRTF amplitude spectrum and used back-propagation artificial neural networks to map the PCA weights of HRTFs to the selected anthropometric parameters. Chun et al. [8] used the deep neural network (DNN) to map the anthropometric parameters to the head-related impulse response (HRIR). Zhang et al. [9] used DNN models based on spatial principal component analysis (SPCA) to predict HRTFs in arbitrary spatial directions. Those methods are based on one database, specifically CIPIC database [10], and can only predict HRTFs in 1 meter.

Since HRTF varies dramatically in near-field, there have been measurements to obtain distance-dependent HRTFs [11, 12], and algorithms and methods are proposed to model HRTFs in near-field. Duda et al. [13] applied a rigid sphere model to simulate sound propagation towards listener's head. Kan et al. [14] calculated a distance variation function (DVF), using a model of the acoustic scattering for a point-source on a rigid sphere, to apply to the HRTFs. Chen et al. [15] implemented a more specific model with head, neck and torso to model near-field HRTFs. Zhang et al. [16], used DNN models based on SPCA to predict HRTFs in arbitrary spatial distances.

In this paper, we are aiming to combine HRTFs in different databases to model individual distance-dependent HRTFs. We first apply the SPCA to HRTFs in CIPIC database, and

This work was supported by the National Natural Science Foundation of China (No. 11590773, No.61175043), the State Key Laboratory of Robotics (2018-O09), and the High-performance Computing Platform of Peking University.

the HRTFs can be represented by a weighted combination of spatial principal components (SPCs) [5]. Then the HRTFs in PKU&IOA database are preprocessed to align with the statistics in CIPIC database. Through combining the individual HRTF model [9] and the distance-dependent HRTF model [16], we finally obtained the individual HRTFs in arbitrary spatial directions and distances by measuring a few anthropometric parameters.

The rest of the paper is organized as follows. In Section 2, the preprocessing of PKU&IOA database is presented. In Section 3, the SPCA is introduced. In Section 4, individual distance-dependent HRTF modeling is described. Section 5 gives the objective and subjective evaluations of the proposed model. In section 6, the conclusion is presented.

2. PREPROCESSING OF PKU&IOA DATABASE

In order to combine the individual HRTF model [9] and the distance-dependent HRTF model [16], we preprocess the raw HRIRs in PKU&IOA database.

a. Re-sample HRIRs. Since the sampling rate of PKU&IOA database is 65536Hz, and the sampling rate of CIPIC database is 44100Hz, we re-sample the HRIRs in PKU&IOA database to 44100Hz.

b. Change the length of HRIRs. The length of each HRIR in PKU&IOA database is 1024 samples, and we change it to 200 samples as those in CIPIC database. Due to different start time of HRIRs of different distances, we cut HRIRs from the eight points prior to the maximum absolute value.

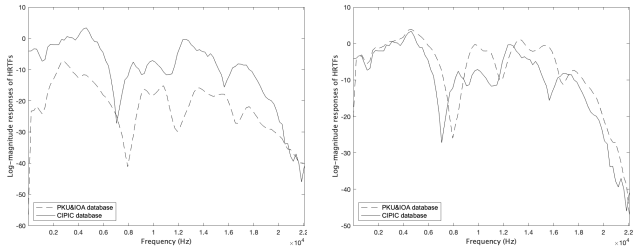
c. Transform HRIRs into the frequency domain. Fourier transformation is applied to HRIRs to obtain HRTFs.

d. Transform HRTFs into a logarithmic scale. The amplitude scale of the HRTFs are linear, and a logarithmic scale is much closer to our auditory perception [17]. We then compute the base 10 log-magnitude responses of HRTFs:

$$HRTF_{log}(\theta,\varphi,r,f,s) = 20log_{10}(|HRTF(\theta,\varphi,r,f,s)|). \quad (1)$$

To compare HRTFs in two databases, we randomly select a HRTF of KEMAR with small ear in both databases. Fig. 1.(a) shows the comparison of HRTFs in elevation of 0 degrees, azimuth of 30 degrees, and distance of 1 meter for two databases. Even though we re-sampled and changed the length of the HRIRs in PKU&IOA database, the HRIRs in two databases are still different. The reason for that is researchers used different measuring methods and equipments to measure HRTFs, so the mean and standard deviation of the two databases are different. Therefore, we first calculate the mean μ_{pku} and standard deviation σ_{pku} across the spatial directions and subjects and normalize the HRTFs of PKU&IOA database to obtain zero mean and unit variance.

$$H_{log}(\theta,\varphi,r,f,s) = \frac{HRTF_{log}(\theta,\varphi,r,f,s) - \mu_{\text{pku}}(r,f)}{\sigma_{\text{pku}}(r,f)}, \quad (2)$$



(a). before (b). after

Fig. 1. Comparison of HRTFs before and after aligning the mean and standard deviation of the two databases.

After that, we calculate the mean μ_{cipic} and standard deviation σ_{cipic} of CIPIC database. Finally, we make the normalized HRTF data in PKU&IOA database have the same mean and standard deviation as those in CIPIC database:

$$HRTF_{\text{pku}}(\theta,\varphi,r,f,s) = H_{log}(\theta,\varphi,r,f,s) \times \sigma_{\text{cipic}}(r,f) + \mu_{\text{cipic}}(r,f). \quad (3)$$

Fig. 1.(b) shows the comparison of HRTFs after aligning the mean and standard deviation. The amplitude of HRTFs of two databases are much closer. To align the mean and standard deviation of all eight distances, we need a model to predict the mean and standard deviation of different distances for CIPIC database, which only contains 1 meter's HRTFs. We use DNN model in [16] to predict mean of different distances. For standard deviation, we use the same DNN architecture as the mean model's. Specifically, the input of DNN is $\sigma_{\text{pku}}(r_0, f)$ of distance $r_0 = 1m$ and target distance r_d, and the ground truth is $\sigma_{\text{pku}}(r_d, f)$. Based on the mean and standard deviation models learned by PKU&IOA database, we can use the same structures and parameters to estimate the mean and standard deviation for CIPIC database. After that, we can calculate $HRTF_{\text{pku}}$ of all eight distances by Eq. (3). Thus, the mean and standard deviation of HRTFs in PKU&IOA database are aligned with those in CIPIC database.

e. The means of $HRTF_{\text{pku}}$ are subtracted:

$$HRTF_{pku\Delta}(\theta,\varphi,r,f,s) = HRTF_{\text{pku}}(\theta,\varphi,r,f,s) - \mu_{\text{cipic}}(r,f). \quad (4)$$

f. The mean spatial function H_{av} is calculated. H_{av} is the mean of $HRTF_{pku\Delta}$ over frequencies and subjects.

$$H_{av}(\theta,\varphi,r) = \frac{1}{NS}\sum_{s=1}^{S}\sum_{f=1}^{N} HRTF_{pku\Delta}(\theta,\varphi,r,f,s). \quad (5)$$

3. SPATIAL PRINCIPAL COMPONENT ANALYSIS

The traditional PCA method is generally used in the time or the frequency domain of HRTFs [18, 19], while SPCA is applied to the spatial domain. The high spatial resolution

HRTFs are decomposed into the combination of SPCs and SPCA weights [5]. To model individual distance-dependent HRTFs, $c_q(r)$ is used to predict the relationship between the SPCA weights of different distances [16]:

$$HRTF_{pku\Delta}(\theta,\varphi,r,f,s) = \sum_q d_{q,r_0}(f,s)c_q(r)W_q(\theta,\varphi) + H_{av}(\theta,\varphi,r), \quad (6)$$

where W_q is SPCs, which depends only on the source direction. φ is elevation angle, and θ is azimuth angle. d_q, r is SPCA weights which vary as functions of frequency f, individual s and distance r. d_{q,r_0} is the SPCA weights of distance r_0. H_{av} is the function of source direction and distance.

4. DISTANCE-DEPENDENT INDIVIDUAL HRTF MODELING

Fig. 2 shows the framework of individual distance-dependent HRTF modeling. We first performed SPCA to HRTFs in CIPIC database, then SPCs, SPCA weights, and H_{av} are obtained. By measuring a few anthropometric parameters, head width, head depth, shoulder width, cavum concha height, cavum concha width, fossa height, pinna height and pinna width, we model SPCA weights and ITDs in 1 meter, and H_{av} is also predicted in arbitrary spatial directions of 1 meter [9]. Due to the spatial directions contained in two databases are different, $D_c = 1250$ directions in CIPIC database and $D_p = 793$ directions in PKU&IOA database, we use the SPCs modeling in [9] to predict the direction vector of SPCs (DV-SPCs) in all the D_p directions and then combine all the DV-SPCs into a $Q \times D_p$ matrix:

$$\boldsymbol{W} = \begin{bmatrix} W_1(0), & W_1(1) & \dots & W_1(D_p-1) \\ W_2(0), & W_2(1) & \dots & W_2(D_p-1) \\ \vdots & \vdots & \vdots & \vdots \\ W_Q(0), & W_Q(1) & \dots & W_Q(D_p-1) \end{bmatrix}, \quad (7)$$

where $\boldsymbol{W}$ is composed of the first $Q = 200$ SPCs [9], and DV-SPCs is a column of $\boldsymbol{W}$ and varies as a function of spatial directions. The original $\boldsymbol{W}$ obtained by applying SPCA to HRTFs in CIPIC database is a $Q \times D_c$ matrix. By predicting SPCs for PKU&IOA database using the data in CIPIC database, we align SPCs of the two HRTF databases.

The HRTFs are measured in interaural-polar coordinate system in CIPIC database but measured in spherical coordinate system in PKU&IOA database. Therefore, we transform the azimuth angle and the elevation angle in PKU&IOA database to interaural-polar coordinate system. The transformation formulas are as follows:

$$\begin{aligned} \sin(\theta') &= \sin(\theta)\sin(\varphi), \\ \tan(\varphi') &= \cot(\varphi)/\cos(\theta), \end{aligned} \quad (8)$$

where θ and φ refer to the azimuth angle and the elevation angle in spherical coordinate system respectively, and θ' and φ' are the azimuth angle and the elevation angle in interaural-polar coordinate system respectively.

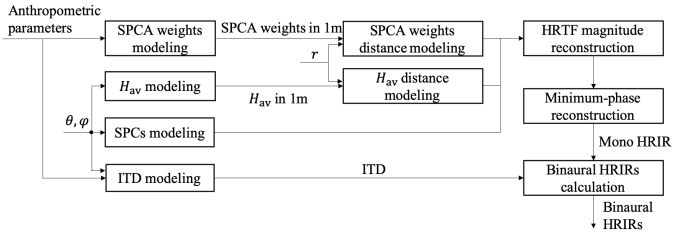


Fig. 2. The framework of individual distance-dependent HRTF modeling.

After transforming the angles, we predict DV-SPCs in all the D_p directions and combine them into the $Q \times D_p$ matrix. Then the HRTFs in PKU&IOA database are projected to the combined SPCs. SPCA weights and H_{av} for PKU&IOA database are then obtained. Through SPCA weights distance modeling and H_{av} distance modeling [16], we obtained the structures and parameters of DNN models for predicting SPCA weights as well as H_{av} in arbitrary spatial distances. Those structures and parameters we obtained through training DNN models in PKU&IOA database can be used directly to the SPCA weights and the H_{av} in CIPIC database, since we aligned the mean and standard deviation of two databases and also projected the PKU&IOA HRTFs to the SPCs obtained by CIPIC database. At this time, we can predict SPCA weights and H_{av} in arbitrary spatial distances for CIPIC database.

To sum up, by measuring a few anthropometric parameters of an individual, we first predict its SPCA weights and H_{av} in 1 meter using CIPIC database. By employing the structures and parameters learned from PKU&IOA database, we then predict SPCA weights and H_{av} in arbitrary spatial distances. Thus, the HRTF magnitude of arbitrary spatial directions and distances can be reconstructed by solving Eq. 6. The minimum phase reconstruction method is applied to HRTF magnitudes to generate mono HRIRs [18]. Since ITD only varies slightly when sound source moves from far-field to near-field [20,21], we consider ITDs in arbitrary spatial distances are equal to the ITDs in distance of 1 meter. Finally, binaural HRIRs in arbitrary spatial directions and distances can be reconstructed.

5. EVALUATION EXPERIMENTS

5.1. Objective experiments

Our proposed model and DVF model [14] are evaluated by spectral distortion (SD):

$$SD = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(20lg\frac{|H(f_k)|}{|\hat{H}(f_k)|})^2}, \quad (9)$$

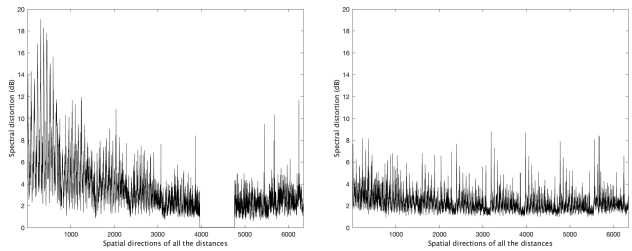


Fig. 3. Comparison of the SD between DVF model (left) and the proposed model (right).

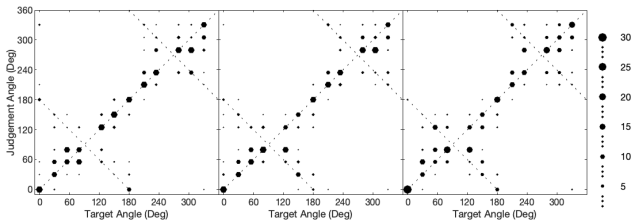


Fig. 4. Judged direction versus target direction of all subjects for the distance of 50 (left), 100 (middle) and 160 (right) centimeters. Each solid circle represents the amount of judgments for a target angle. Size of circle is increased with the judgments.

where $H(f_k)$ is the magnitude response of the HRTF from PKU&IOA database (after the first two steps of preprocessing), $\hat{H}(f_k)$ is the magnitude response of the HRTF estimated by DVF model or the proposed model.

The SD of the reconstructed HRTFs of KEMAR with small ear is shown as Fig. 3. The abscissa values are 793 directions of eight distances, and the distance becomes larger from left to right. Note that our proposed model projecting HRTFs in 200 SPCs results in an average SD of 1.62 dB, so this is the reason that the SD of all spatial locations in our proposed model is larger than approximately 1.5 dB. SD equaling zero in some spatial directions of DVF model is because we use HRTFs at 1 meter to estimate HRTFs at other spatial distances. The SD value is quite large for distances close to the head in the DVF model, because a rigid sphere model is used, and ignoring the details about human head leads to a bad prediction performance when sound source is closer to the head. The average SD of the proposed model in all the sampled directions and distances is 2.34 dB, and the average SD of DVF model is 3.79 dB. This demonstrates that the proposed model is superior to DVF model.

5.2. Subjective experiments

The stimuli in this experiment is a train of eight 250-ms bursts of Gaussian noise (20-ms cosine-squared onset-offset ramps), with 300 ms of silence between the bursts. The HRIRs of twelve azimuth angles (0, 30, 55, 80, 125, 150, 180, 210, 235, 280, 305, and 330 degrees) in three distances, 50, 100 and 160 centimeters, are generated by the proposed model. Then, the stimulus is filtered by the HRTFs to obtain the virtual sounds. A total of three azimuth localization experiments are performed. The three experiments correspond to three distances, 50, 100 and 160 centimeters, respectively. Before each experiment, the subject is trained using the sound of other eight azimuth angles (0, 45, 90, 135, 180, 225, 270, and 315 degrees). Through listening to these sounds, the subject can build up the spatial perception for the virtual sound. After that, thirty-six binaural sounds are randomly played to the subject by a Sennheiser HD 650 headphone through a Sound Blaster sound card. The thirty-six sounds contain twelve directions' sounds, and each direction appears three times. The subject gave the exact direction of each sound he/she perceived during the experiments through an interface on a computer. After each experiment, there are five minutes for a break. Twelve subjects (10 male 2 female, age from 21 to 29) with normal hearing took part in the experiments. All experiments were performed in a sound booth (Background noise level : 20.9 dBA), with no light during the experiments.

Fig. 4 shows the results of localization experiments of all twelve subjects in three distances respectively. The judgments are plotted as a function of the coordinates of the targets. There are 432 judgments shown in each panel, corresponding to the thirty-six judgments made for each of the twelve binaural sounds. Each solid circle represents the amount of judgments for a target angle. Size of circle is increased with the judgments. The average correction rates for the distance of 50, 100 and 160 centimeters are 59.7%, 59.5%, and 58.6% respectively. The average confusion rates for the distance of 50, 100 and 160 centimeters are 18.8%, 26.9%, and 29.2% respectively. The average angle of errors for the distance of 50, 100 and 160 centimeters are 11.12, 11.06, and 11.27 degrees respectively. Results show that our distance-dependent individual HRTF modeling method effectively predicts HRTFs in arbitrary spatial directions and distances.

6. CONCLUSION

The paper proposed an individual distance-dependent HRTF modeling method based on CIPIC and PKU&IOA databases. By combining the individual model and the distance-dependent model, we predict HRTFs in arbitrary spatial directions and distances. Objective experiments show that our proposed model is superior to the DVF model. Subjective experiments show that the HRTFs predicted by our proposed method are effective. Therefore, by measuring a few anthropometric parameters for an individual, we can predict its HRTFs in arbitrary spatial directions and distances.

7. REFERENCES

[1] W. Kreuzer, P. Majdak, and Z. Chen, "Fast multipole boundary element method to calculate head-related transfer functions for a wide frequency range," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1280–1290, 2009.

[2] F. Ma, J. Wu, M. Huang, W. Zhang, W Hou, and C. Bai, "Finite element determination of the head-related transfer function," *Journal of Mechanics in Medicine and Biology*, vol. 15, no. 05, pp. 1550066, 2015.

[3] T. Xiao and H. L. Qing, "Finite difference computation of head-related transfer function for human hearing," *The Journal of the Acoustical Society of America*, vol. 113, no. 5, pp. 2434–2441, 2003.

[4] K. J. Fink and L. Ray, "Tuning principal component weights to individualize hrtfs," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, 2012, pp. 389–392.

[5] B. Xie, "Recovery of individual head-related transfer functions from a small set of measurements.," *Journal of the Acoustical Society of America*, vol. 132, no. 1, pp. 282–294, 2012.

[6] D. N. Zotkin, J. Hwang, R. Duraiswaini, and L. S. Davis, "Hrtf personalization using anthropometric measurements," in *2003 IEEE Workshop on Applications of Signal Processing To Audio and Acoustics*, New Paltz, NY, USA, 2003, pp. 157–160.

[7] H. Hu, L. Zhou, H. Ma, and Z. Wu, "Hrtf personalization based on artificial neural network in individual virtual auditory space," *Applied Acoustics*, vol. 69, no. 2, pp. 163–172, 2008.

[8] C. J. Chun, J. M. Moon, G. W. Lee, N. K. Kim, and H. K. Kim, "Deep neural network based hrtf personalization using anthropometric measurements," in *Audio Engineering Society Convention 143*. Audio Engineering Society, 2017.

[9] M. Zhang, Z. Ge, T. Liu, X. Wu, and T. Qu, "Modeling of individual hrtfs based on spatial principal component analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, no. 1, pp. 785–797, 2020.

[10] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The cipic hrtf database," in *2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, New Platz, NY, USA,, 2001, IEEE, pp. 99–102.

[11] T. Qu, Z. Xiao, M. Gong, Y. Huang, X. Li, and X. Wu, "Distance-dependent head-related transfer functions measured with high spatial resolution using a spark gap," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1124–1132, 2009.

[12] H. Wierstorf, M. Geier, and S. Spors, "A free database of head related impulse response measurements in the horizontal plane with multiple distances," in *Audio Engineering Society Convention 130*. Audio Engineering Society, 2011.

[13] R. O. Duda and W. L. Martens, "Range dependence of the response of a spherical head model," *The Journal of the Acoustical Society of America*, vol. 104, no. 5, pp. 3048–3058, 1998.

[14] A. Kan, C. Jin, and A. van Schaik, "Distance variation function for simulation of near-field virtual auditory space," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2006, vol. 5, pp. 325–328.

[15] Z. Chen, G. Yu, B. Xie, and S. Guan, "Calculation and analysis of near-field head-related transfer functions from a simplified head-neck-torso model," *Chinese Physics Letters*, vol. 29, no. 3, pp. 034302, 2012.

[16] M. Zhang, Y. Qiao, X. Wu, and T. Qu, "Distance-dependent modeling of head-related transfer functions," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2019, pp. 276–280.

[17] J. O. Smith, *Techniques for digital filtering design and system identification with the violin*, Ph.D. thesis, CCRMA, Stanford, 1983.

[18] D. J. Kistler and F. L. Wightman, "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction," *Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1637, 1992.

[19] M. Zhang, R. A. Kennedy, T. D. Abhayapala, and W. Zhang, "Statistical method to identify key anthropometric parameters in hrtf individualization," in *The Workshop on Hands-Free Speech Communication & Microphone Arrays*, Edinburgh, UK, 2011, pp. 213–218.

[20] D. S. Brungart and W. M. Rabinowitz, "Auditory localization of nearby sources. head-related transfer functions," *The Journal of the Acoustical Society of America*, vol. 106, no. 3, pp. 1465–1479, 1999.

[21] M. Otani, T. Hirahara, and S. Ise, "Numerical study on source-distance dependency of head-related transfer functions," *The Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3253–3261, 2009.