Interspeech 2018
2-6 September 2018, Hyderabad

Measuring the Band Importance Function for Mandarin Chinese with an Bayesian Adaptive Procedure

Yufan Du[1], *Yi Shen*[2], *Hongying Yang*[1], *Xihong Wu*[1], *Jing Chen*[1]

[1]Department of Machine Intelligence, Speech and Hearing Research Center, and Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing, China.
[2]Department of Speech and Hearing Sciences, Indiana University Bloomington, 200 S Jordan Ave., Bloomington, IN 47405

duyufan@pku.edu.cn, shen2@indiana.edu, 1501214507@pku.edu.cn, wxh@cis.pku.edu.cn, janechenjing@pku.edu.com

Abstract

A speech intelligibility index (SII) based band importance function (BIF) for Mandarin monosyllabic words spoken by a female speaker was derived with an adaptive procedure in this work. The adaptive procedure, namely the quick-band-importance-function (qBIF) procedure, optimized the stimulus on each trial according listeners' performance on proceeding trials in an iterative fashion. This method greatly improved the efficiency of data collection. Test-retest experiments were conducted and confirmed the reliability of this adaptive procedure at a group level. The BIF derived in this work showed generally consistence with the BIF derived with the traditional paradigm with noticeable differences at certain frequencies.

Index Terms: speech intelligibility index, band importance function, Mandarin Chinese, the qBIF procedure

1. Introduction

Speech intelligibility index (SII) [1] is a widely used model for estimating speech intelligibility under various listening conditions, such as additive noise and bandwidth reductions. In principle, the SII represents the proportion of the speech spectrum that is audible, with each frequency band weighted according to the typical contribution of that band (i.e. its importance) to intelligibility. Therefore, the band importance function (BIF) is a key component of the SII framework.

Previous studies have shown that the BIF depends on the speech material. For English speech, six BIFs are standardized for six types of speech materials, including nonsense syllables, phoneme-balanced words, and short messages [1]. For Chinese speech, Wong et al. [2] derived the BIF for sentences in Cantonese, and Chen et al. [3] estimated the BIF for monosyllabic words in Mandarin. Their results showed some difference between Chinese and English speech due to their acoustic characteristics. Overall, the characteristics of the BIF for Chinese speech have not been thoroughly studied, and this is partially due to the time-consuming process of estimating the BIF.

The traditional method to estimate the BIF used in the studies mentioned above involves the following steps [4, 5]. First, speech recognition is measured among normal-hearing listeners for the low-pass and high-pass filtered speech with systematically varied cut-off frequencies and at various signal to noise ratios (SNRs). At a certain intermediate cutoff frequency, equal performance is achieved for the low- and high-pass conditions, which can be considered as the performance associated with half of the speech information, i.e. an SII value of 0.5. A function relating the SII to speech recognition scores, called a transfer function, is achieved by repeating this procedure. Second, using the obtained transfer function, the relationship between performance and cutoff frequency is converted into a function between cut-off frequency and SII value. Finally, the BIF is derived by subtracting the SII value of successive cut-off frequencies and averaging high- and low-pass conditions. Such experimental procedure is very time-consuming. For example, each estimate of the BIF by Chen et al. [3] took about 14400 trials (2 speaker genders, 36 filtering conditions, and 4 SNRs) and more than 20 hours for each listener.

Due to this methodological limitation, it is often necessary to combine data collected from multiple listeners or using repeated speech tokens for each listener. Therefore, a more time efficient method for estimating the BIF is desirable. Shen and Kern [6] reduced the spectral resolution of the BIF to six octave bands and adopted a Bayesian adaptive testing technique in an effort to reduce the testing time and the amount of unique speech tokens required for BIF estimation. The resulting quick-band-importance-function (qBIF) procedure allowed the estimation of the BIF for monosyllabic English words from individual listeners using 300-400 experimental trials. Besides the improved test efficiency the qBIF procedure also included stimulus-generation steps that prevented issues associated with the traditional high- and low-pass filtering paradigm. The traditional paradigm is based on the assumption that various frequency bands contribute to speech intelligibility independently. However, several recent studies reported the redundancy and synergetic effects among frequency bands [7, 8, 9]. Some recently proposed procedures circumvented this issue by separating the speech material into sub-bands and randomly selecting a subset of the sub-bands for stimulus presentation [10, 11]. The qBIF procedure followed this more recent "compound" paradigm to improve the validity of the estimated BIFs.

In this work, we adapted the qBIF procedure to evaluate the BIF for monosyllabic words for Chinese Mandarin spoken by a female speaker. Although the qBIF procedure only provides estimates of the BIF at a reduced resolution (i.e. in six octave bands) compared to earlier studies [3], it provides a unique opportunity to investigate 1) the test-retest variabilities in BIF estimates and 2) the effect of different stimulus paradigms (i.e. the traditional high- and low-pass filtering

10.21437/Interspeech.2018-1825

paradigm versus the compound paradigm). The implementation details of the qBIF will be briefly described in Section 2; the experimental methods and results will be presented in Section 3 and 4, respectively; the implications of the results will be discussed in Section 5.

2. The quick band importance function (qBIF) method

The SII is a value between 0 and 1, which can be calculated by:

$$\mathrm{SII} = \sum_{i=1}^{n} w_i A_i \tag{1}$$

where n is the number of bands, w_i and A_i are the band importance function and the audibility function for the ith band, respectively. The spectral weights w_i in the BIF sum to unity. The audibility function represents the audibility for each band, which is derived from SNR and listener's hearing thresholds [4]. In the qBIF procedure, a logistic function is used to describe the relationship between the correct proportion of speech recognition and SII value:

$$p = \left[1 + \mathrm{e}^{-\beta(\mathrm{SII}-b)}\right]^{-1} \tag{2}$$

According to (1), (2) can be re-written as:

$$p = \left[1 + \mathrm{e}^{-\beta(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{A}-b)}\right]^{-1} \tag{3}$$

where β reflects the slope of the logistic function, and b is the SII value at 50% correct recognition, which is associated with the speech recognition threshold (SRT). Column vector $\boldsymbol{w}$ and $\boldsymbol{A}$ represent the BIF and audibility function, respectively. In the qBIF procedure, the speech signal is divided into six octave bands with center frequencies at 250, 500, 1000, 2000, 4000, 8000 Hz. It is assumed that the speech stimuli are presented sufficiently above the listener's hearing threshold, so that $\boldsymbol{A}$ in (3) is dominated by the SNR in each band:

$$A_i = \frac{SNR_i+15}{30}, i = 1, \cdots, 6 \tag{4}$$

where SNR_i represent the signal to noise ratio of the ith band and its value is assumed to be within the range of -15 to 15 dB. By this normalization, the value of A_i is always between 0 and 1. Additionally, the value of A_i is assumed to be 0 if the ith band is absent from the stimulus. The subjects' responses (correct or incorrect) under various SNR conditions (described by parameter $\boldsymbol{A}$) are collected, from which the band importance function (parameter $\boldsymbol{w}$) can be estimated by logistic regression.

In the current implementation of the qBIF procedure, the SNR could take one of five possible values (-5, 0, 5, 10, and 15dB). The combinations of various bands were limited so that the number of bands presented at the same time ranged between 2 and 5. This led to a total of 280 stimulus conditions and 280 different values of $\boldsymbol{A}$ $[5 \times (\mathrm{C}_6^2 + \mathrm{C}_6^3 + \mathrm{C}_6^4 + \mathrm{C}_6^5)]$.

An entropy-based criterion [12, 13, 14] is used for stimulus selection to maximize the information gain from each trial. After the kth trial of the qBIF procedure, the band importance function $\boldsymbol{w}_k$ is derived via logistic regression, and the performance for the next trial is predicted for all stimulus conditions according to (3). For each candidate stimulus, two band importance functions can be estimated base on two possible response (correct and incorrect) and their entropies are noted as $\boldsymbol{H}_{\mathrm{k+1}_{correct}}$ and $\boldsymbol{H}_{\mathrm{k+1}_{incorrect}}$. The overall expected entropy can be estimated:

$$\boldsymbol{E}(\boldsymbol{H}_{k+1}) = p_{correct} \times \boldsymbol{H}_{\mathrm{k+1}_{correct}} + (1 - p_{correct}) \times \boldsymbol{H}_{\mathrm{k+1}_{incorrect}} \tag{5}$$

A smaller overall expected entropy corresponds to greater expected information gain. Therefore, the stimulus condition that minimizes the overall expected entropy is selected for the next (k+1th) trial:

$$\boldsymbol{A}_{k+1} = \arg\min_{\boldsymbol{A}} \boldsymbol{E}(\boldsymbol{H}_{k+1}) \tag{6}$$

Additional implementation details about the qBIF procedure can be found in [6].

In the above-described one-step-ahead search algorithm, the estimated band importance function $\boldsymbol{w}_k$ relies on the accuracy of $\boldsymbol{w}_{k-1}$. However, at the beginning of the qBIF procedure, the estimated $\boldsymbol{w}_k$ is not sufficiently accurate due to the lack of data. In a related study using the qBIF procedure in English speech, the initial SNR was set as 5 dB and response of 30 trials were collected before the activation of the one-step-ahead search algorithm [6]. However, in our pilot experiments, we found the task was too difficult to active the algorithm correctly with the previous setting. Hence, the initial SNR was set as 15 dB, and 50 trials were tested before the entropy-based criterion method activated.

3. Experiment

3.1. Subjects

Eleven listeners (18-23 years old, mean=21.2 years, 8 females) took part in this study. They all had pure-tone thresholds ≤ 25 dB HL at octave frequencies from 250 to 8000 Hz. All subjects were native speakers of Mandarin Chinese and were paid for their participation.

3.2. Stimuli

The monosyllabic words were from the national standard of China "Acoustics-Speech articulation testing method" [15]. There were 10 word lists and each list contained 75 monosyllabic words. These words were phonetically balanced and spoken by 5 female and 5 male speakers. Lists 1 through 8 spoken by Talker F4 (a female talker) were used in this study. The speech signal was digitalized at 16 kHz and 16-bit resolution. The root-mean-square amplitudes of all words were equalized. A steady speech-spectrum noise (SSN) was produced according to the long-term average spectrum of Talker F4 [3], which was used as the masker signal. A single talker was used in the current study, so that the observed variabilities in the results could not be attributed to talker variations. To construct the stimulus for each trial, the speech signal and the SSN were mixed at a certain SNR. The mixture signal was then separated into the six octave bands using 12th-order butterworth filters. A subset of the six sub-bands was recombined to generate the test stimulus. The SNR and the choice for the sub-bands varied from trial to trial and were controlled by the stimulus optimization algorithm described in (6).

3.3. Procedure

The current experiment was conducted in an anechoic chamber, which was 5.6 m in length, 4 m in width and 1.93 m in height. Speech signals were presented via a soundcard (Fireface UC) and a loudspeaker (Dynaudio Acoustics, BM64). During the testing, the listeners were seated at the centre of

anechoic chamber and in front of loudspeaker. The distance between loudspeaker and the centre of listeners' head was 2 m. The height of loudspeaker and subjects' ear was approximately the same. The speech level was 65 dBA, measured at the subjects' head position.

The experiment consisted of a test and a retest phase, which were designed to evaluate the reliability of the BIF estimates. Four word lists (300 trials), randomly drawn for each listener, were used in the test phase, and the other four word lists (300 trials) were used in the retest phase. For each test phase, the word lists were tested in random order. For each list, the words were also tested in random order. During each trial, the listener was instructed to verbally repeat each word that he or she heard. The experimenter scored the listener's response in term of correctness. The experiment was completed in about 1 hour for each listener, and a break was allowed between the test and retest phases.

4. Results

4.1. Test-retest reliability of the BIF estimates

Figure 1 shows the mean BIF across eleven subjects derived from the test and retest phases of the current experiment, as well as their average. The BIFs obtained from the test and retest phases had a similar shape. The highest weight was found for the 2000-Hz band, which was consistent with the previous study [3] for Mandarin Chinese.

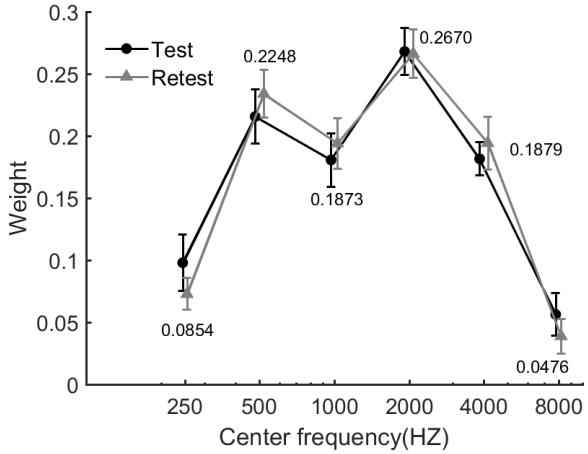

Figure 1: *The BIFs estimated using the qBIF procedure for Mandarin Chinese monosyllabic words. The black line (circles) represents the average results across all subjects from the test phase of the current experiment, while the gray line (triangles) represents the average results from the retest phase. Error bars represent ± one standard errors. The text labels provide the average spectral weights across the test and retest results.*

The average BIFs obtained from the test and retest phases were highly correlated (r = .99, p < .001), with a root-mean-square deviation (RMSD) of 0.016, indicating satisfactory test-retest reliability at a group level (for a group of 11 listeners). To investigate whether adequate test-retest reliability can be obtained with a reduced number of listeners, a bootstrap simulation was conducted for group sizes of 1 to 10. For each group size, 1000 simulation runs were conducted. Within each simulation run, a subset of listeners were randomly drawn according to the group size and the average BIFs across the drawn listeners were calculated for the test and retest phases. The correlation coefficient and the RMSD were calculated based on the average BIFs. Independent draws were conducted across simulation runs.

Figure 2 shows the average correlation coefficient and RMSD across 1000 simulation runs as a function of group size. For individual listeners (i.e. a group size of 1), the BIFs from the test and retest phases were not expected to correlate to each other and the RMSD was expected to be above 0.08. Therefore, intra-subject variability in the estimated BIF was quite substantial. Since an identical speech material (monosyllabic words produced by a single female talker) was used in both test and retest phases, the observed intra-subject variability could not be explained by variations in stimuli. As the group size increased, the average correlation coefficient increased and the RMSD decreased. A significant test-retest correlation was expected for group sizes above 3 (with a critical r value of 0.811 for df = 4 and α = 0.05, two tailed).

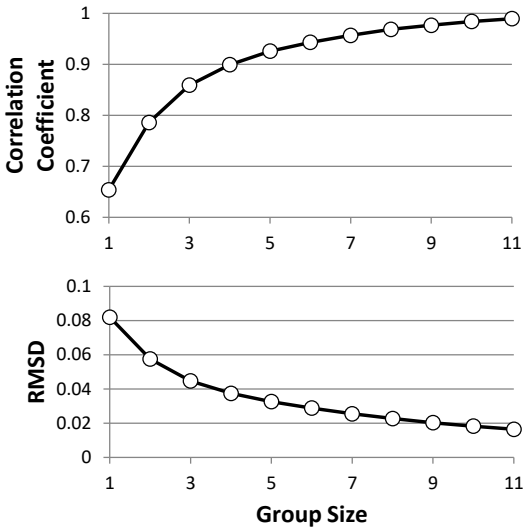

Figure 2: *The correlation coefficient (top) and the root-mean-square deviation (RMSD) between the average BIFs from the test and retest phases as a function of group size. Results for each group size less than 11 were based on 1000 bootstrap simulations with resampling of listeners.*

4.2. Comparison between stimulus paradigms

Chen et al. [3] derived the 1/3 octave-band BIF using a similar speech corpus as the current study. However, instead of the compound stimulus paradigm, these authors used the traditional, high- and low-pass filtering approach. Figure 3 shows the BIF obtained from the two studies for female talkers. To enable direct visual comparisons, the weights estimated by Chen et al. were summed within each octave bands.

Results from both studies showed a prominent peak in the BIF located at the 2000-Hz frequency band, indicating the specific importance of this frequency range for recognizing monosyllabic words in Mandarin [3]. In the work by Chen et al. using a similar speech material [3], it was found that the frequency bands centred at 1600 and 2000 Hz were especially important for mandarin Chinese because the F2 formant frequencies mainly located in this range and the F2

information was critical for identifying Chinese vowels [3, 16, 17].

The main difference between the two studies occurred at 500 Hz, where a second peak in the BIF was observed in the current study but not in the study by Chen et al. This difference may be caused by the difference in stimulus paradigms. The speech stimuli were always presented in contiguous bands in the traditional high- and low-pass filtering approach, but they were presented in a randomly distributed manner in the qBIF procedure. The latter approach, i.e. the compound paradigm, was designed to limit the confounding factors such as the redundancy and synergetic effects among frequency bands. It has been reported that the compound paradigm could lead to significant changes to the BIFs below 2000 Hz compared to those estimated using the traditional procedure [11]. It is worth noting that the average BIFs shown in Figure 3 were measured using monosyllabic words produced by two different female talkers, however, it is unlikely that the observed differences between the two average BIFs were due to talker differences. This is because the talker effect within the same gender has been shown to have a fairly weak effect on BIF [18]. Moreover, Chen et al. reported BIF estimates in 1/3-octave-band resolution, and the weights shown in Figure 3 were the result of summing weights within each octave range. Therefore, they cannot be considered as equivalent to the octave-band BIFs [3]. The discrepancies due to methodology need to be further studied for Mandarin speech.

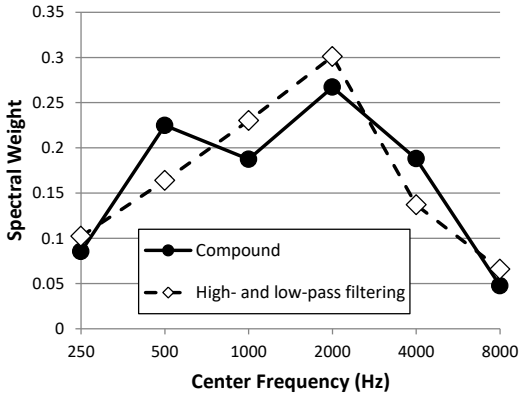

Figure 3: *The average band importance function estimated using the qBIF procedure for Mandarin Chinese monosyllabic words is plotted using filled circles. The band importance function for Mandarin Chinese monosyllabic words derived using traditional method [3] is plotted using diamonds.*

5. Discussions

5.1. Limitations of the qBIF procedure

The qBIF procedure is able to estimate BIF efficiently but it still has some limitations. First, the SII model was simplified to accelerate the test process at the cost of spectral resolution. The current implementation of the qBIF procedure derived octave-band BIFs instead of 1/3-octave-band BIFs as in many previous studies. If the qBIF procedure was implemented using 1/3-octave bands, the complexity of the SII model would increase significantly. Whether the qBIF procedure would lead to stable and reliable estimates in such a situation would need to be explored in the future.

Second, although excellent test-retest reliability was found for the average BIFs estimated using the qBIF at a group level, substantial deviations in the BIFs between the test an retest phases was observed for individual listeners. Therefore, at least for Mandarin monosyllabic words tested using the qBIF procedure of 300 trials, relative large intra-subject variability is expected.

5.2. Other issues

One goal of improving BIF estimation is to guide hearing-aid fitting, individually estimated BIF may provide insights into how hearing-impaired listeners utilize speech cues in various frequency regions. Toward this goal, a crucial step is to investigate the intra- and inter-subject variability while removing the effect caused by variations in stimulus features (e.g., material type, talker gender). In this work, the BIF was measured with the same type of speech material spoken by the same female in order to minimize the variations in the stimuli. However, it is known that the shape of the BIF depends on the speech materials, caused by the redundancy of the speech materials and the characteristics of the speaker [3-5]. Results from the current study may not be diversified enough to represent the band importance of Mandarin Chinese for daily speech communication. Future studies with systematically varied speech materials are warranted.

In addition, the prominent advantage of the qBIF procedure is the efficiency of data collection. To derive the BIF with the traditional method for Mandarin Chinese [3], a listener was required to finish 7200 test trials (36 filtering conditions ×4 SNRs × 50 words) for one speaker's speech, while only 300 trials were required for the qBIF procedure in this work. This makes it possible to estimate BIF with other Mandarin speech materials, e.g., disyllable words and sentences, efficiently.

6. Summary

The present work derived the octave-band BIF for Chinese monosyllabic words spoken by a female speaker, using an adaptive procedure, namely the qBIF procedure. With the improved efficiency provided by the qBIF procedure, the test-retest reliability of the estimated BIFs was studied. Excellent test-retest reliability was demonstrated at a group level, but not at an individual level. The main trend of the derived octave-band BIF was similar to the BIF derived with the traditional method, with minor deviations likely due to the different stimulus paradigms. Further work is necessary to extend the current results for a wider range of Mandarin speech materials (e.g., other speakers, disyllable words and sentences).

7. Acknowledgements

The work was supported by the National Natural Science Foundation of China (Grant Nos. 61473008, 61771023, and 11590773), the Newton alumni funding by the Royal Society, UK, and NIH grant R21 DC013406.

8. References

[1] ANSI, "S3.5-1997, methods for the calculation of the speech intelligibility index," *New York: American National Standards Institute*, 1997.

[2] L. L. N. Wong, A.H.S. Ho, E. W. W. Chua, and S.D. Soli, "Development of the Cantonese speech intelligibility index,"

The Journal of the Acoustical Society of America, vol. 121, no. 4, pp. 2350–2361, 2007.
[3] J. Chen, Q. Huang, and X. Wu, "Frequency importance function of the speech intelligibility index for mandarin Chinese," *Speech Communication*, vol. 83, pp. 94–103, 2016.
[4] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *The Journal of the Acoustical society of America*, vol. 19, no. 1, pp. 90–119, 1947.
[5] G. A. Studebaker and R. L. Sherbecoe, "Frequency-importance and transfer functions for recorded CID W-22 word lists. *Journal of Speech, Language, and Hearing Research*, vol. 34, no. 2, pp. 427-438, 1991.
[6] Y. Shen and A. B. Kern, "An analysis of individual differences in recognizing monosyllabic words under the Speech Intelligibility Index framework," *Trends in Hearing*, vol. 22, 2331216518761773, 2018.
[7] R. M. Warren, K. R. Riener, J. A. Bashford, and B. S. Brubaker, "Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits," *Attention, Perception, & Psychophysics*, vol. 57, no. 2, pp. 175–182, 1995.
[8] E. W. Healy and R. M. Warren, "The role of contrasting temporal amplitude patterns in the perception of speech," *The Journal of the Acoustical Society of America*, vol. 113, no. 3, pp. 1676–1688, 2003.
[9] L. E. Humes and G. R. Kidd, "Speech recognition for multiple bands: Implications for the Speech Intelligibility Index," *The Journal of the Acoustical Society of America*, vol. 140, no. 3, pp. 2019-2026, 2016.
[10] F. Apoux and E. W. Healy, "Use of a compound approach to derive auditory-filter-wide frequency importance functions for vowels and consonants," *The Journal of the Acoustical Society of America*, vol. 132, no. 2, pp. 1078–1087, 2012.
[11] E. W. Healy, S. E. Yoho, and F. Apoux, "Band importance for sentences and words reexamined," *The Journal of the Acoustical Society of America*, vol. 133, no. 1, pp. 463–473, 2013.L. L.
[12] Kontsevich and C. W. Tyler, "Bayesian adaptive estimation of psychometric slope and threshold," *Vision research*, vol. 39, no. 16, pp. 2729–2737, 1999.
[13] Y. Shen and V. M. Richards, "Bayesian adaptive estimation of the auditory filter," *The Journal of the Acoustical Society of America*, vol. 134, no. 2, pp. 1134-1145, 2013.
[14] Y. Shen, R. Sivakumar, and V. M. Richards, "Rapid estimation of high-parameter auditory-filter shapes," *The Journal of the Acoustical Society of America*, vol. 136, no. 4, pp. 1857-1868, 2014.
[15] GB-T15508, "Acoustics-speech articulation testing method," *Standardization Administration of the People's Republic of China,* 1995.
[16] T Lin and L. J. Wang, *Yuyinxue Jiaocheng (in Chinese)*. Peking University Press, 1992.
[17] F. Chen, L. L. N. Wong, and E. Y. W. Wong, "Assessing the perceptual contributions of vowels and consonants to mandarin sentence intelligibility," *The Journal of the Acoustical Society of America*, vol. 134, no. 2, pp. EL178–EL184, 2013.
[18] S.E. Yoho, E. W. Healy, C. L. Youngdahl, T. S. Barrett, and F. Apoux, "Speech-material and talker effects in speech band importance," *The Journal of the Acoustical Society of America*, vol. 143, no. 3, pp. 1417-1426, 2018.