Multiband product rule and consonant identification

Feipeng Li[a] and Jont B. Allen
Department of Electrical and Computer Engineering, University of Illinois at Urbana–Champaign, Urbana, Illinois 61801

(Received 31 July 2008; revised 5 February 2009; accepted 6 May 2009)

The multiband product rule, also known as band-independence, is a basic assumption of articulation index and its extension, the speech intelligibility index. Previously Fletcher showed its validity for a balanced mix of 20% consonant-vowel (CV), 20% vowel-consonant (VC), and 60% consonant-vowel-consonant (CVC) sounds. This study repeats Miller and Nicely's version of the hi-/lo-pass experiment with minor changes to study band-independence for the 16 Miller–Nicely consonants. The cut-off frequencies are chosen such that the basilar membrane is evenly divided into 12 segments from 250 to 8000 Hz with the high-pass and low-pass filters sharing the same six cut-off frequencies in the middle. Results show that the multiband product rule is statistically valid for consonants on average. It also applies to subgroups of consonants, such as stops and fricatives, which are characterized by a flat distribution of speech cues along the frequency. It fails for individual consonants. © *2009 Acoustical Society of America.* [DOI: 10.1121/1.3143785]

PACS number(s): 43.71.An, 43.71.Es, 43.71.Gv [MSS] Pages: 347–353

I. INTRODUCTION

A fundamental problem of human speech perception is how the human auditory system integrates speech cues across frequency. The most relevant study on this topic dates back to the 1920s, when Fletcher[1] at Bell Laboratories investigated speech articulation over voice communication systems. Low-pass and high-pass filtered "nonsense syllables" were used for the study of phone recognition. They found that the average phone error e of the full-band stimuli is equal to the product of the error of the low-pass filtered stimuli e_L and the error of the complimentary high-pass filtered stimuli e_H, that is,

$$e = e_L \times e_H. \tag{1}$$

In other words, the low-pass band and the high-pass band are consistent with the assumption that the low band and high band are independent. Equation (1) was then generalized, by assumption, into a multiple band form[1–3]

$$e = e_1 e_2 \cdots e_K. \tag{2}$$

The number of independent articulation bands is generally taken to be $K=20$, which makes each band correspond to about 1 mm along the basilar membrane.[2]

Let s denote the average phone articulation (i.e., the probability of the nonsense phones being correctly recognized), then the articulation error $e=1-s$, and the articulation band error $e_1=1-s_1$, etc. Given Eq. (2),

$$\log(1-s) = \sum_{k=1}^{K} \log(1-s_k). \tag{3}$$

Notice that $\log(1-s_k)$ is similar to the definition of entropy,[4] thus may be interpreted as the information carried by the kth band.[2,3] Equation (3) implies that the human speech recognition system consists of at least K parallel channels and that the total information is equal to the sum over the information in the K articulation bands. This relation may also be called the *additivity law of frequency integration*. It is the foundation of the two ANSI standards: articulation index (AI) (Ref. 5) and more recently speech intelligibility index (SII).[6]

Based on the assumption of independent articulation bands, French and Steinberg[7] developed a method for calculation of AI based on the intensity of the long-term average speech and noise. Following the verification by Beranek[8] and Kryter,[9] French and Steinberg's method[7] became an ANSI standard in 1969. Then in 1970–1980 Steeneken and Houtgast[10] extended the AI to the speech transmission index (STI) by introducing a modulation transfer function to account for reverberation and peak clipping. The original AI was developed for the use of normal hearing listeners. Later it was extended to estimate the speech intelligibility for the hearing-impaired listener,[11–14] resulting in a new ANSI standard named the SII. All the three models, AI, STI, and SII, are based on the same Fletcher–Galt assumption[3] that the total articulation is the sum of the contribution from multiple independent narrow bands.

Despite its importance to the widely used articulation models, the validity of the multiband product rule [Eq. (2)] has actually been a key open question.[15] For example, Kryter[16] showed that AI was a valid predictor of the intelligibility of speech under a wide variety of conditions of noise masking and speech distortion except for the cases of three non-contiguous pass bands at 0–600, 1200–2400, and 4800–9600 Hz. Grant and Braida[17] found that the predicted AI based on the sum of the AIs from individual bands was greater than the observed AI by approximately 18% for adjacent 1/3-octave bands, while the AI predicted for combinations of non-adjacent bands was less than the observed AI by approximately 41%. Lippmann[18] also found that the stop-band data did not agree with AI calculation. In 2001 Müsch and Buus[19,20] coined two new terms *synergistic* and *redun-*

[a]Author to whom correspondence should be addressed. Electronic mail: fli2@illinois.edu

dant interactions between neighboring bands to explain why the AI under, or over, estimates the wide-band error, compared to the product of the errors associated with the narrow bands. It has been conjectured that a revised model, which accounts for the mutual dependency between adjacent bands, might give a better prediction.[21] In a recent study, Ronan *et al.*[22] compared several frequency integration models for the prediction of individual consonant articulation score, for narrow-band cases. Results indicated that Fletcher's product rule[1] [Eq. (2)] made satisfactory predictions under various combinations of adjacent and non-adjacent narrow-band speech, except for the case of multiple high-frequency narrow bands, for which none of the evaluated methods are satisfactory. Investigation of SII (Ref. 23) also found that it greatly over-predicted performance at high sensation levels, and under-predicted performance at low sensation levels for many hearing-impaired listeners. The information contained in each frequency band is not strictly additive.

In 1955, Miller and Nicely[24] (MN55) repeated Fletcher and Galt's high-pass and low-pass filtering experiment[3] for the analysis of perceptual confusion. The speech stimuli includes 16 consonant sounds, /p,t,k,f,θ,s,ʃ,b,d,g,v,ð,z,ʒ,m,n/ spoken initially before the vowel /a/. Using the data from experiment MN55, we checked the validity of Fletcher's product rule [Eq. (1)].[1] Results[24] show that the model applies to the consonants on average, despite that it over-predicts the full-band error by 10%. We then plotted the product of e_L and e_H against the full-band error e for each of the 16 consonant sounds [see Fig. 2(b)]. To our surprise, more than half of the consonant sounds, specifically, /p,k,f,ʃ,b,d,g,ʒ,m,n/, show only small discrepancy.

Designed for the purpose of confusion analysis, the MN55 data are unsuitable for the study of the multiband product rule, for several reasons. First, the frequency samples are limited. Only six low-pass and five high-pass condition are included, in contrast Fletcher[1] and French and Steinberg[7] suggested $K=20$ frequency points. Second, the cut-off frequencies are not evenly distributed along the effective range of speech communication. Four out of six low-pass samples are below 1.5 kHz, with only one high-pass sample within the same frequency range. Interpolation between data points introduces significant error.

In the present study we investigate the validity of the multiband product rule for consonant sounds. The product rule is evaluated on three levels: (1) 16 consonants on average, (2) subgroups such as stops and fricatives, and (3) individual consonants. A computer-based high-pass and low-pass experiment, named HL07, is designed for this purpose (see Fig. 1). The new experiment utilizes the same 16 consonant sounds as experiment MN55. To address the problems listed above, the cut-off frequencies were chosen such that the basilar membrane is evenly divided into 12 bands over the frequency range from 0.25–8 kHz, with the low-pass and high-pass filters sharing the same six cut-off frequencies in the mid-frequency range.

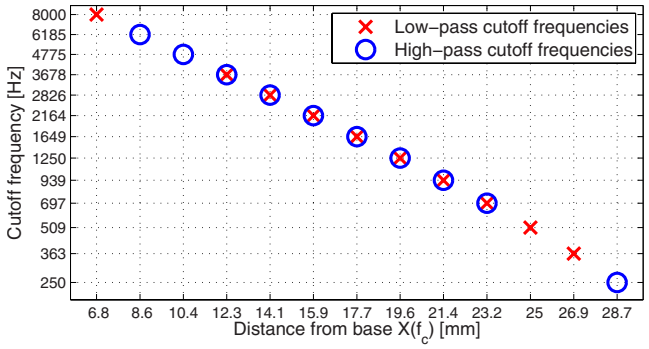

FIG. 1. (Color online) High-pass and low-pass cut-off frequencies of experiment HL07.

II. METHODS

A. Subjects

19 normal hearing subjects were enrolled in the experiment, of which 6 male and 12 female listeners completed. Except for one subject in her 40s, all the subjects were college students in their 20s. The subjects were born in the United States with English being their first language. All subjects were paid for their participation. IRB approval was obtained for the experiment. In order to make sure that all the data are of high quality, the performance of the listeners was assessed by their average recognition score. Those who had abnormally low scores will be excluded for further analysis. In experiment HL07, no subject has been removed for that reason.

B. Speech stimuli

The same 16 nonsense consonant-vowels (CVs) used by Miller and Nicely[24] were chosen. A subset of wide-band syllables sampled at 16 kHz were taken from the LDC-2005S22 corpus. Each CV was spoken by 20 talkers, among which only 6 utterances, half male and half female, were finally chosen for the test, to reduce the total duration of the experiment. The six utterances were selected such that they were representative of the speech material in terms of confusion patterns and articulation score based on the results of similar speech perception experiment.[25] The speech sounds were presented to both ears of the subjects at the listener's most comfortable level, but always less than 80 dB SPL.

C. Conditions

The subjects were tested under 19 filtering conditions, including 1 full-band 0.25–8 kHz, 9 high-pass, and 9 low-pass conditions. The cut-off frequencies were calculated from Greenwood's inverse cochlear map function[26] such that the full-band frequency range (0.25–8 kHz) was divided into 12 bands, corresponding to equal length along the basilar membrane. Figure 1 illustrates the frequency samples and the correspondent distances from the base on the human basilar membrane. The cut-off frequencies of the high-pass filtering were 6185, 4775, 3678, 2826, 2164, 1649, 1250, 939, and 697 Hz, with the upper-limit at 8000 Hz. The cut-off frequencies of the low-pass filter were 3678, 2826, 2164, 1649, 1250, 939, 697, 509, and 363 Hz, with a lower-limit at 250

Hz. The high-pass and low-pass filtering shared the same cut-off frequencies over the middle frequency range that contains most of the speech information. The filters were sixth order elliptical filter with 0.02 dB of peak-to-peak ripple and a stop-band attenuation of −60 dB. To make the filtered speech sound more natural and to mask the stop bands, white noise was used to mask the stimuli at the signal-to-noise ratio (SNR) of 12 dB, based on the average speech spectra of the 96 nonsense syllables.

D. Procedure

The speech perception experiment was conducted in a sound-proof booth. A MATLAB code was developed for the collection of the data. Speech stimuli were presented to the listeners through Sennheisser HD 280-pro headphones. Subjects responded by clicking on the button labeled with the CV that they heard. In case the speech was completely masked by the noise, or the processed token did not sound like any of the 16 consonants, the subjects were instructed to click on a "noise only" button. A total of 2208 tokens were randomized and divided into 16 sessions, each of which lasted for about 15 min. A mandatory practice session of 60 tokens was given at the beginning of the experiment. To prevent fatigue the subjects were instructed to take frequent breaks. The subjects were allowed to play each token for up to three times. At the end of each session, the subject's test score, together with the average score of all listeners, was shown to the listener to provide feedback on their relative progress, as motivation.

E. Difference between HL07 and MN55

Although experiment HL07 can be regarded as a repeat of the MN55 study, the two experiments are distinguished in several important aspects. First, the subjects differ in gender and proficiency. In MN55 five extensively-trained female subjects served as both talkers and listening crew. This introduced a "coupling" effect between the talkers and the listeners, as well as an awareness of the relative difficulty of the sounds. In HL07 we use recorded speech prepared by ten male and eight female talkers from the LDC database. All the 18 subjects (6 male and 12 female) are naive listeners without any experience in speech perception tests. Second, the noise levels are different. Both experiments use white noise at 12 dB SNR. However, in experiment MN55, the speech level was controlled by a volume unit (VU) meter,[27] which measures the speech peaks, while in experiment HL07 the noisy speech were created by setting the rms level of the speech and noise. Thus 12 dB SNR in MN55 is about the same as 14 dB SNR in HL07.[27] As a consequence, the full-band error of MN55 is about 12% lower than that of HL07. Third, the filtering conditions are different. In MN55 the full-band speech was created by a wide-band filter of 0.2–6.5 kHz, and then the distorted speech were created by filtering the full-band speech with low-pass cut-off frequencies of 0.3, 0.4, 0.6, 1.2, 2.5, and 5 kHz and high-pass cut-off frequencies of 0.2, 1.0, 2.0, 2.5, 3.0, and 4.5 kHz. In contrast, the full-band speech in HL07 goes to 8 kHz. The loss of information from 6.5 to 8 kHz accounts well for the over-prediction of MN55 in the high frequency. Fourth, the test platforms are different. Data collection in MN55 was paper-based. The listeners were told to choose a response from the 16 nonsense CVs and write it down on the answer sheet within seconds following the presentation. The HL07 experiment is computer-based. No limit is applied for the responding time. Subjects were allowed to play each sound up to three times. In case the subjects could not tell which sound is presented, a noise only button was added.

F. Data analysis

The validity of Fletcher's product rule[1] [Eq. (1)] is investigated for average speech and individual consonants. The probability of error of a token (an utterance filtered at a frequency) is defined as the number of mis-labeled responses divided by the total number of presentations. The mean error of a consonant is the average over the six tokens pronounced by different talkers. Similarly, the total error of average speech can be calculated by averaging the errors of the 16 consonants. For both average speech and individual consonants, the fitness of the model to the data is evaluated in terms of average bias $B(f_c)$ and $\chi^2(f_c)$ computed from the error of all listeners. The average bias is given by

$$B(f_c) = e - e_L \times e_H, \tag{4}$$

where $e_L \times e_H$ and e are the model error and observed error at a cut-off frequency f_c. The chi-square statistic is

$$\chi^2(f_c) = N\frac{[(1-e_L\times e_H)-(1-e)]^2}{1-e_L\times e_H} + N\frac{[e-e_L\times e_H]^2}{e_L\times e_H}, \tag{5}$$

where N is the total number of presentations for the particular condition. The quantities $(1-e_L\times e_H)$ and $(1-e)$ are the predicted and observed scores. A significance level (the probability of this result not being due to chance) of 0.05 is chosen as the threshold of the chi-square test. A value of χ^2 greater than the threshold indicates that the measurements do not satisfy Eq. (1) at that condition, whereas when χ^2 is less than the threshold of significance, Fletcher's product rule[1] can be regarded as true.

The above analysis is carried out by treating the 18 listeners as average normal listeners. In order to determine if the same conclusion applies to any individual listeners, a one-way analysis of variance (ANOVA) test is applied to the $e-e_L\times e_H$ of different listeners following each χ^2 test. Due to the small number of responses, the 16 sessions are combined into 4 repeats, 4 sessions each. Let B_i denote the bias of $e_L\times e_H$ against e for subject i, and B_{ij} denote the bias of repeat j from subject i. Assuming that B_i has a Gaussian distribution $N(b_i,\sigma)$, where b_i is the mean of B_i, we can compare the mean of the various listeners by testing the hypothesis that they all have the same bias, against the general alternative that they are not all the same. If no two listeners are significantly different, we may conclude that the conclusion based on the average normal listeners is applicable to any individual listeners.

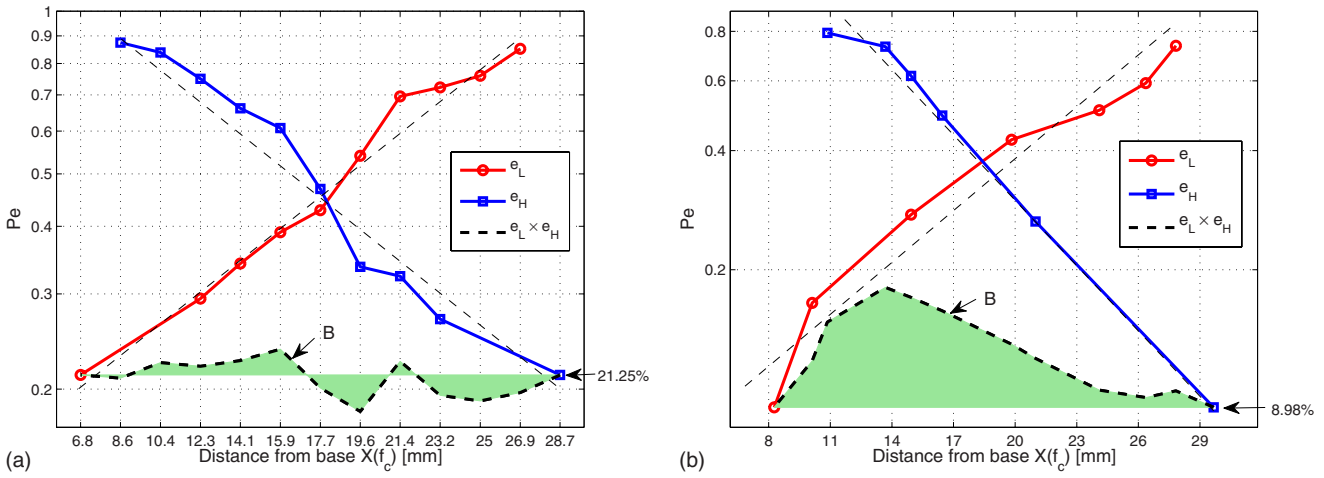

FIG. 2. (Color online) Grand probability of error and the average bias $B=e-e_L\times e_H$ for 16 consonants as a function of cut-off frequency. (a) shows the average low-pass error e_L (circles), the average high-pass error e_H (squares), and the product of the two, $e_L\times e_H$ (thick dashed) for experiment HL07. The full-band error e is defined as $e_L(f_c=8000\ \text{Hz})$ or $e_H(f_c=250\ \text{Hz})$. The average bias B is depicted by the shaded area. (b) shows the same data from experiment MN55, in which the full-band error e is defined as $e_L(f_c=6500\ \text{Hz})$ or $e_H(f_c=200\ \text{Hz})$. Note the log ordinate scale, which makes the figures easily read, actually magnifies the bias visually.

III. RESULTS

A. Multiband product rule for 16 consonants on average

Results indicate that the multiband product rule closely fits the recognition scores averaged over the 16 consonants. Figure 2(a) depicts the low-pass error e_L, the high-pass error e_H, and their product as a function of cut-off frequency. The full-band error e is equal to the low-pass error e_L at 8000 Hz and the high-pass error e_H at 250 Hz. The missing points of the low-pass error at 4775 and 6185 Hz, and the high-pass error at 363 and 509 Hz, are linearly interpolated from the nearest neighboring points. The average bias $B=e-e_L\times e_H$ is depicted by the shaded area. Suppose that the product rule is true, the shaded area would be zero. It is shown in Fig. 2(a) that the difference between $e_L\times e_H$ and the full-band error e is typically less than 3%, which is very close to zero.

Figure 2(b) depicts the results of experiment MN55.[24] Fletcher's product rule[1] over-predicts the full-band error over most frequencies for MN55, but still the measurements fit the model with reasonable accuracy. Since the low-pass and the high-pass conditions do not use the same set of cut-off frequencies, the low-pass error e_L and high-pass error e_H are linearly interpolated along the frequency to create the $e_L\times e_H$ curve, which introduces extra error in the prediction.

For both experiments, the intersection points of the low-pass and high-pass curves that divide the full band into two parts of equal information are about the same (1.5 kHz or 18 mm). The log low-pass error e_L and high-pass error e_H have been fitted by two straight lines that are symmetrical at the intersection point. This means the speech information is evenly distributed across frequency. A significant difference between the results of MN55 and HL07 lies in that the former has a maximum average bias B of 8.02%, which is considerably smaller than that of HL07 (21.25%). This might be due to the aforementioned coupling effect between the talkers and the listeners in experiment MN55, which makes the task relatively easier. Apart from that, the results of the two experiments are generally consistent. Due to the experimental design, experiment HL07 has a better precision (smaller bias) than experiment MN55, as we seen in Fig. 2. Therefore, in the remaining part of Sec. III, we will focus on analyzing the perceptual data of our experiment HL07.

Table I lists the average bias of the predicted score [the same data are depicted in Fig. 2(a) as the shaded area]. The results of the χ^2 tests indicate that e_L, e_H, and e are consistent with Fletcher's product rule[1] at all frequencies. An ANOVA test indicates that the difference between the 18 listeners is too small to be statistically significant at the level of 0.05. The discrepancy between the biases of any individual listeners and the overall average bias is generally less than 5%. Therefore the 18 listeners of normal hearing can be regarded as having the same bias $e-e_L\times e_H$ independent of cut-off frequencies. Thus Fletcher's product rule[1] may be applied to any individual normal hearing listener.

B. Multiband product rule for stops and fricatives

Analysis of the perceptual data indicates that the multiband product rule applies to the stops and fricatives as well. Figure 3(a) depicts the average low-pass error e_L, average high-pass error e_H, and the product of the two $e_L\times e_H$ for the six stop consonants (/pa, ka, ta, ba, ga, da/). The average bias $B=e-e_L\times e_H$, as depicted by the shaded area, is rather small. The high-pass error and the low-pass error cross each other at about 1.5 kHz, which is about the same position (18 mm)

TABLE I. The average bias of 16 consonants on average in experiment HL07 for various cut-off frequencies.

Frequency (Hz)	363	509	697	939	1250	1649	2164	2826	3678	4775	6185
$B=e-e_L\times e_H$	−1.9	−2.6	−1.8	1.3	−3.1	−1.2	2.5	1.3	0.8	1.7	0.3

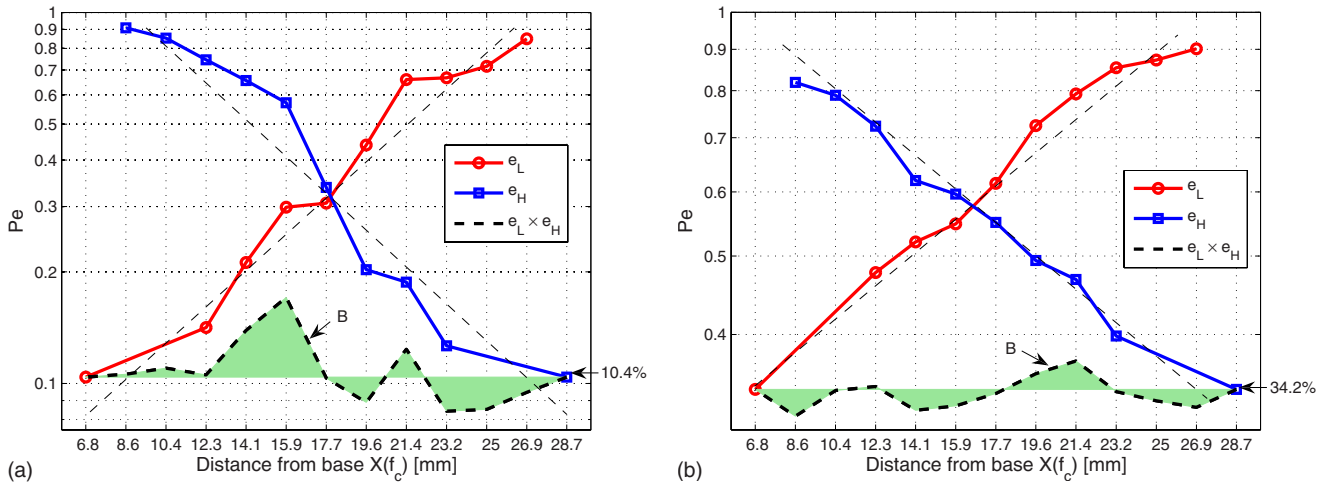

FIG. 3. (Color online) Average probability of error and the average bias $B=e-e_L\times e_H$ for stops (/pa, ka, ta, ba, ga, da/) and fricatives (/fa, θa, sa, ʃa, va, ða, za, ʒa/) as a function of cut-off frequency. (a) shows the average low-pass error e_L (circles), the grand high-pass error e_H (squares), and the product of the two, $e_L\times e_H$ (thick dashed), for stops. The average bias $B=e-e_L\times e_H$ is the shaded area. (b) shows the same results for the fricatives.

as the weight of the 16 consonants on average. The logarithms of e_L and e_H are well approximated by straight lines having complementary but identical slopes.

The results for the eight fricative consonants (/fa, θa, sa, ʃa, va, ða, za, ʒa/) are depicted in Fig. 3(b). The average bias B is almost flat with the maximum prediction error being less than 3%. Like the case of average consonants, $e_L(f_c)$ and $e_H(f_c)$ have near constant equal slopes of opposite signs when the two curves are plotted on log scales, suggesting that the fricative information is evenly distributed across the frequency range.

Table II lists the average bias B for the two sound groups at various cut-off frequencies. All values satisfy the χ^2 test at a significance level of 0.05. An ANOVA test shows no significant difference between the results of the 18 listeners.

C. Multiband product rule for individual consonants

Analysis of our HL07 data reveals that Fletcher's product rule[1] applies to the 16 consonants over limited frequencies for about 80% of the cases (CVs × frequencies). Figure 4 depicts the low-pass error e_L, high-pass error e_H, and the product of the two $e_L\times e_H$ for the 16 consonants. Based on the shape of $e_L\times e_H$, the 16 consonants can be roughly classified into flat and non-flat groups. The flat group includes /pa, ka/ and /ʃa, da, ma, na, za, ga, sa, fa, va/, for which the prediction error $e_L\times e_H-e$ is less than 5% over all frequencies, or less than 5% for most of the cut-off frequencies. The rest of the consonant sounds, /ta, ba, ʒa, θa, ða/, form the biased (non-flat) group.

Table II lists the average bias of the predicted score (the same data are depicted in Fig. 4 as the shaded area). A χ^2 test of significance level 0.05 was applied to each of the 16 consonants. A total of 136 out of 176 cases (16 CVs × 11 frequencies) statistically satisfy Fletcher's product rule[1] at a significance level of 0.05. Only two consonants /pa, ka/ passed the χ^2 test over all frequencies. Most of the unsatisfied cases come from the biased group, such as /ta, ba, ða, ʒa/, for which the fail rate is 50%.

An ANOVA test was used to investigate the listener's dependence. Since the number of tokens per CV × frequency for each listener is only 6, a number too small for a useful statistical test, the 18 listeners are ranked according to their speech recognition scores and artificially divided into three groups. The top six are attributed to the H group. The middle six are attributed to the M group. The lower six are classified as the L group. For 173 out of 176 combinations (16 CVs × 11 frequencies) ANOVA tests produce the same result that the H, M, and L groups are not significantly different in terms of the average bias per CV × frequency. In other words, the three groups of listeners are close to each other in terms of the fitness to the multiband product rule (see Table III).

The perceptual data provide important information on the perceptual cues for the initial consonants. Usually the primary cue of a consonant is located around the intersection point of e_L and e_H, which divides the full band into two parts having equal information (e.g., score). When the primary speech cue is removed, the error climbs dramatically.[28]

TABLE II. The average bias of stops and fricatives in experiment HL07 for various cut-off frequencies.

	Frequency (Hz)										
Subgroup	363	509	697	939	1250	1649	2164	2826	3678	4775	6185
Stop	−1.1	−2.0	−2.0	2.0	−1.5	−0.1	6.6	3.5	0.1	0.9	0.5
Fricatives	−2.1	−1.5	−0.2	2.9	1.5	−0.4	−1.6	−2.0	0.3	0.8	−1.5

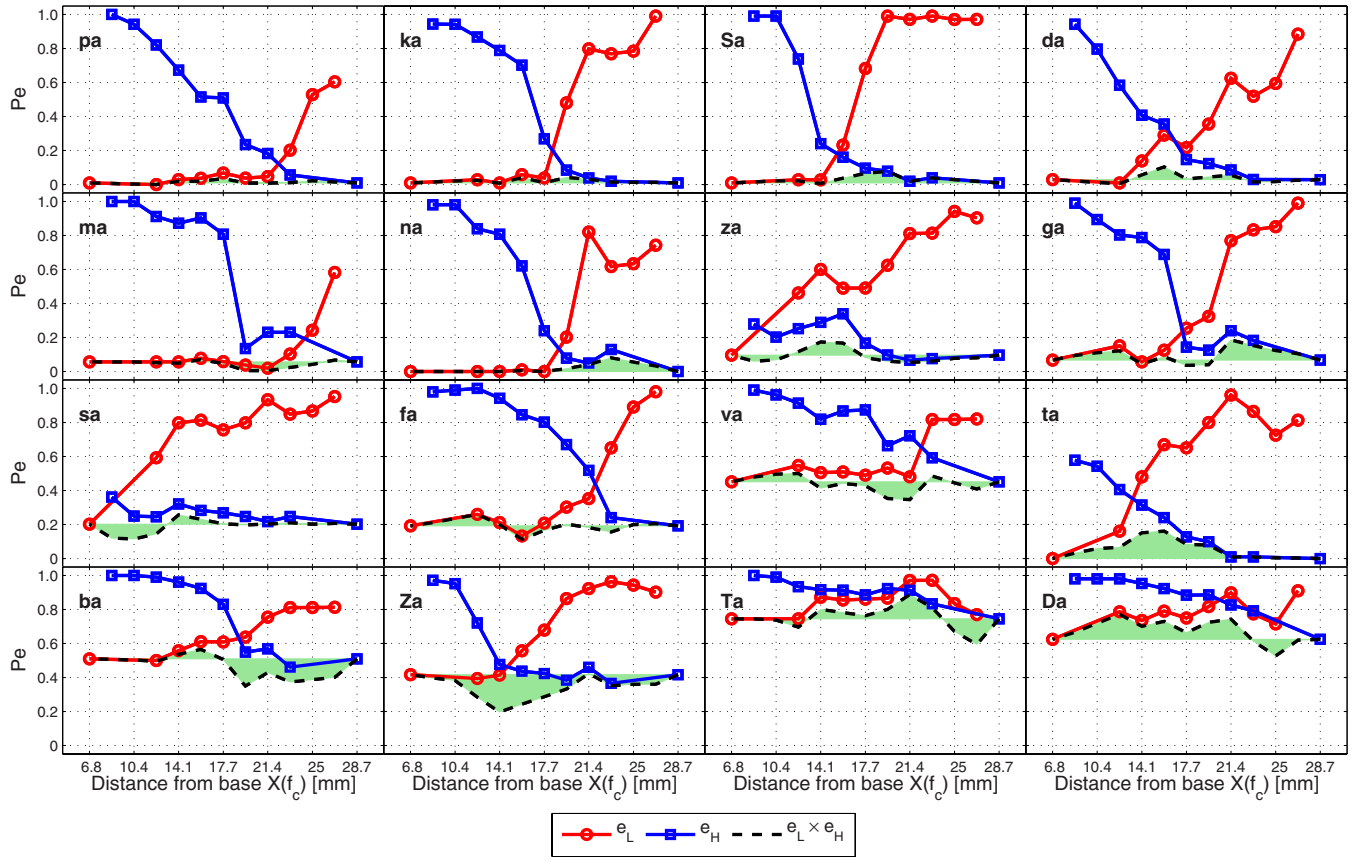

FIG. 4. (Color online) Probability of error for 16 consonants as a function of cut-off frequency. The low-pass error $e_L(f_c)$ and the high-pass error $e_H(f_c)$ are marked by circles and squares, respectively. The dashed curve depicts the product of the two $e_L \times e_H$. The full-band error e is equal to $e_L(f_c=8000$ Hz) or $e_H(f_c=250$ Hz). The bias $B(f_c)=e-e_L \times e_H$ is illustrated by the shaded area. The International Phonetic Alphabet (IPA) symbols for Ta, Sa, Da, and Za are /θa, ʃa, ðа, ʒа/, respectively.

IV. GENERAL DISCUSSION

In Sec. III A, we demonstrated that Fletcher's product rule[1] [Eq. (1)] is true for the average consonants at all cut-off frequencies. This can be regarded as a significant verification of the multiband product rule of frequency integration [Eq. (2)]. Suppose that Eq. (2) is a consequence of the fact that the frequency bands b_k, associated with e_k, are independent in terms of speech perception. A strict proof would require a

TABLE III. The average biases of 16 consonant sounds in experiment HL07 for various cut-off frequencies. Cases for which the χ^2 test was statistically significant at the 0.05 level are marked with an asterisk.

	Frequency (Hz)										
CV	363	509	697	939	1250	1649	2164	2826	3678	4775	6185
pa	0.3	1.0	0.2	−0.1	−0.1	2.5	0.9	1.0	−1.0	−0.7	−0.4
ka	0.2	0.2	0.5	2.1	3.1	0.1	3.1	−0.2	1.5	1.2	0.7
ʃa	0.7	1.7	3.0	0.9	6.8*	5.6*	2.8	−0.3	1.2	1.4	0.8
da	−0.3	−1.1	−1.3	2.5	1.5	0.3	7.5*	2.8	−2.3	−1.7	−0.9
ma	0.2	−1.8	−3.3*	−5.2	−5.1	−1.0	1.4	−0.7	−0.5	0.0	0.0
na	2.4	4.8*	8.0*	4.0*	1.6	0.0	0.6	0.0	0.0	0.0	0.0
za	−1.4	−1.7	−3.5	−4.3*	−3.5	−1.4	7.0*	7.7*	2.0	−2.2	−2.6
ga	2.8	4.7	8.6*	11.8*	−2.6	−3.0	1.9	−2.2	5.5	4.9	3.4
sa	0.1	−0.4	0.8	0.1	−0.4	0.2	2.8	5.4	−5.7	−7.9*	−7.0*
fa	0.8	0.4	−3.6	−1.0	1.0	−2.6	−7.9*	0.7	6.7	4.8	2.4
va	−5.1	−1.5	3.3	−10.4*	−9.8*	−2.3	−0.9	−3.7	4.9	5.2	3.6
ta	0.2	0.4	0.8	0.9	7.8*	8.3*	16.1*	15.1*	6.5*	6.5*	3.9*
ba	−10.5*	−11.9*	−13.6*	−8.1	−16.0*	−0.4	5.5	2.6	−1.4	−0.7	−0.4
ʒа	−5.3	−5.2	−6.5	0.7	−8.4	−12.9*	−17.3*	−22.0*	−13.2*	−3.6	−2.1
θа	−15.5*	−8.0	6.5	14.2*	5.5	1.8	3.7	5.5	−4.9	−0.7	0.0
ðа	−1.7	−10.8*	−1.1	11.8*	9.9*	3.8	10.5*	7.6	14.7*	10.6*	5.5

speech perception test that actually measures the 20 narrowband recognition scores. This is totally impractical for $K=20$, as it would require $20!=2.5\times10^{18}$ tests.

If we look at the real perceptual data [Fig. 2(a)], it actually provides much more information. The logarithms of both e_L and e_H can be closely fitted by two lines symmetrical across the intersection point of the two curves. This clearly indicates that (1) the speech information is uniformly distributed across the basilar membrane, as independently measured by both low-pass and high-pass tests; and (2) the articulation bands are additive in log error in speech perception. Similar results are observed for the two groups of stops and fricatives [Figs. 3(a) and 3(b)].

Based on the observation, it is conjectured that the multiband product rule is a combined property of the peripheral auditory system that has multiple independent parallel channels, and that the input speech stimuli are characterized by a uniform distribution of speech cues along the basilar membrane. It does not apply to individual consonants because the distribution of individual consonant speech cues is not flat. Due to the priori dependence between the speech cues, sometimes the high-pass and low-pass errors do not fit the model. For example, when the primary cue of a sound covers more than one band, the product of the low-pass and high-pass error $e_L\times e_H$ may be lower or higher than full-band error e, due to the fact that the bands neighboring the cut-off frequency are not really independent. To fully understand the interactions between the speech cues and explain why the multiband product rule fails at certain points necessitates knowledge of the speech features.[29]

V. CONCLUSION

The multiband product rule of frequency integration is an empirical formula justified by the two properties about speech and hearing, specifically, (1) the speech information is evenly distributed across the frequency, and (2) the auditory critical bands are independent in terms of speech perception. Results of our experiment HL07 show that the multiband product rule is statistically valid for consonants on average. It may also apply to subgroups of consonant sounds, such as stops and fricatives, which are characterized by a flat distribution of speech cues along the frequency. It fails for individual consonants, as expected.[30,31]

ACKNOWLEDGMENTS

The authors are grateful for thoughtful discussions with Bryce E. Lobdell, Michael Kramer, and members of the HSR group at University of Illinois, Urbana.

[1]H. Fletcher, *Speech and Hearing in Communication*, ASA edition (Acoustical Society of America, Woodbury, NY, 1995).
[2]J. B. Allen, "How do humans process and recognize speech?," IEEE Trans. Speech Audio Process. **2**, 567–577 (1994).
[3]H. Fletcher and R. Galt, "The perception of speech and its relation to telephony," J. Acoust. Soc. Am. **22**, 89–151 (1950).
[4]C. E. Shannon, "The mathematical theory of communication," Bell Syst. Tech. J. **27**, 379–423 (1948); "The mathematical theory of communication," **27**, 623–656 (1948).
[5]American National Standard methods for the calculation of the articulation index, A.S3.5-1969 (American National Standards Institute, New York, NY, 1969).
[6]Methods for calculation of the speech intelligibility index (SII-97), A.S3.5-1997 (American National Standards Institute, New York, NY, 1997).
[7]N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am. **19**, 90–119 (1947).
[8]L. L. Beranek, "The design of speech communication systems," Proc. IRE **35**, 880–890 (1947).
[9]K. D. Kryter, "Methods for the calculation and use of the articulation index," J. Acoust. Soc. Am. **34**, 1689–1697 (1962).
[10]H. Steeneken and T. Houtgast, "A physical method for measuring speech transmission quality," J. Acoust. Soc. Am. **67**, 318–326 (1980).
[11]V. Duggirala, G. A. Studebaker, C. V. Pavlovic, and R. L. Sherbecoe, "Frequency importance functions for a feature recognition test material," J. Acoust. Soc. Am. **83**, 2372–2382 (1988).
[12]C. V. Pavlovic, "Use of the articulation index for assessing residual auditory function in listeners with sensorineural hearing impairment," J. Acoust. Soc. Am. **75**, 1253–1258 (1984).
[13]C. V. Pavlovic, G. A. Studebaker, and R. L. Sherbecoe, "An articulation index based procedure for predicting the speech recognition performance of hearing-impaired individuals," J. Acoust. Soc. Am. **80**, 50–57 (1986).
[14]G. A. Studebaker, C. V. Pavlovic, and R. L. Sherbecoe, "A frequency importance function for continuous discourse," J. Acoust. Soc. Am. **81**, 1130–1138 (1987).
[15]J. B. Allen, *Articulation and Intelligibility* (Morgan and Claypool, Princeton, NJ, 2005).
[16]K. D. Kryter, "Validation of the articulation index," J. Acoust. Soc. Am. **34**, 1698–1702 (1962).
[17]K. W. Grant and L. D. Braida, "Evaluating the articulation index for auditory visual input," J. Acoust. Soc. Am. **89**, 2952–2960 (1991).
[18]R. P. Lippmann, "Accurate consonant perception without mid-frequency speech energy," IEEE Trans. Speech Audio Process. **4**, 66–69 (1996).
[19]H. Müsch and S. Buus, "Using statistical decision theory to predict speech intelligibility I. Model structure," J. Acoust. Soc. Am. **109**, 2896–2909 (2001).
[20]H. Müsch and S. Buus, "Using statistical decision theory to predict speech intelligibility II. Measurement and prediction of consonant-discrimination performance," J. Acoust. Soc. Am. **109**, 2910–2920 (2001).
[21]H. Steeneken and T. Houtgast, "Mutual dependence of octave-band weights in predicting speech intelligibility," Speech Commun. **28**, 109–123 (1999).
[22]D. Ronan, A. K. Dix, P. Shah, and L. D. Braida, "Integration across frequency bands for consonant identification," J. Acoust. Soc. Am. **116**, 1749–1762 (2004).
[23]T. Y. Ching, H.Dillon, and D. Byrne, "Speech recognition of hearing-impaired listeners: Predictions from audibility and the limited role of high-frequency," J. Acoust. Soc. Am. **103**, 1128–1140 (1998).
[24]G. A. Miller and P. E. Nicely, "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**, 338–352 (1955).
[25]S. Phatak and J. Allen, "Consonant and vowel confusions in speech-weighted noise," J. Acoust. Soc. Am. **121**, 2312–2326 (2007).
[26]D. D. Greenwood, "A cochlear frequency-position function for several species—29 years later," J. Acoust. Soc. Am. **87**, 2592–2605 (1990).
[27]B. Lobdell and J. Allen, "A model of the vu (volume-unit) meter, with speech applications," J. Acoust. Soc. Am. **121**, 279–285 (2007).
[28]M. S. Régnier and J. B. Allen, "A method to identify noise-robust perceptual features: Application for consonant /t/," J. Acoust. Soc. Am. **123**, 2801–2814 (2008).
[29]J. B. Allen, "Consonant recognition and the articulation index," J. Acoust. Soc. Am. **117**, 2212–2223 (2005).
[30]P. Heil, H. Neubauer, A. Tiefenau, and H. von Specht, "Comparison of absolute thresholds derived from an adaptive forced-choice procedure and from reaction probabilities and reaction times in a simple reaction time paradigm," J. Assoc. Res. Otolaryngol. **7**(3), 279–298 (2006).
[31]C. E. Shannon, "Communication in the presence of noise," Proc. IRE **37**, 10–21 (1949).