Speech-material and talker effects in speech band importance

Sarah E. Yoho, Eric W. Healy, Carla L. Youngdahl, Tyson S. Barrett, and Frédéric Apoux

Citation: The Journal of the Acoustical Society of America **143**, 1417 (2018); doi: 10.1121/1.5026787

View online: https://doi.org/10.1121/1.5026787

View Table of Contents: http://asa.scitation.org/toc/jas/143/3

Published by the Acoustical Society of America

Speech-material and talker effects in speech band importance

Sarah E. Yoho,[a] Eric W. Healy, and Carla L. Youngdahl[b]
Department of Speech and Hearing Science, The Ohio State University, Columbus, Ohio 43210, USA

Tyson S. Barrett
Department of Psychology, Utah State University, Logan, Utah 84322, USA

Frédéric Apoux[c]
Department of Speech and Hearing Science, The Ohio State University, Columbus, Ohio 43210, USA

(Received 17 April 2017; revised 12 February 2018; accepted 21 February 2018; published online 14 March 2018)

Band-importance functions created using the compound method [Apoux and Healy (2012). J. Acoust. Soc. Am. **132**, 1078–1087] provide more detail than those generated using the ANSI technique, necessitating and allowing a re-examination of the influences of speech material and talker on the shape of the band-importance function. More specifically, the detailed functions may reflect, to a larger extent, acoustic idiosyncrasies of the individual talker's voice. Twenty-one band functions were created using standard speech materials and recordings by different talkers. The band-importance functions representing the same speech-material type produced by different talkers were found to be more similar to one another than functions representing the same talker producing different speech-material types. Thus, the primary finding was the relative strength of a speech-material effect and weakness of a talker effect. This speech-material effect extended to other materials in the same broad class (different sentence corpora) despite considerable differences in the specific materials. Characteristics of individual talkers' voices were not readily apparent in the functions, and the talker effect was restricted to more global aspects of talker (i.e., gender). Finally, the use of multiple talkers diminished any residual effect of the talker.

https://doi.org/10.1121/1.5026787

[DB] Pages: 1417–1426

I. INTRODUCTION

The speech intelligibility index (SII; ANSI, 1997) represents more than a method for predicting intelligibility based on acoustic measurement without the need for human subjects. It also reflects in many ways our current understanding of human speech perception. One of its central components, the band-importance function, reflects our current understanding of how speech information is distributed across frequency. In the current study, two potential influences on the shape of the band-importance function are examined. These influences include the role of the speech material and the role of the individual talker.

Healy *et al.* (2013) termed these "speech-material" and "talker" effects. In the first possibility, importance function shape and/or structure is affected by the type of speech material employed, including the particular phonetic and semantic composition of the sentences or words. In the latter, importance function shape and/or structure is affected by the particular talker employed to produce the materials, including the specific idiosyncratic acoustic aspects of his or her voice.

Much work has gone into developing functions for different speech materials using the ANSI method (ANSI, 1969, 1997). These include the CID W-22 word lists (Studebaker and Sherbecoe, 1991), high- and low-context sentences (Bell *et al.*, 1992), and continuous discourse (Studebaker *et al.*, 1987). These different functions were created with the assumption that the primary determinant of their shape is the particular linguistic content of the speech, which could conceivably be obtained simply by seeing the printed text (i.e., the speech-material effect). In accord with this assumption that the speech-material effect dominates, these functions have typically been derived using a standard single-talker recording (usually male voice), and are typically assumed or explicitly stated to be accurate for any recording of that material.

There is good reason to believe that different speech materials will produce differently shaped band-importance functions, if those speech materials are restricted in their phonetic diversity. Take the limiting case of isolated phonemes. Whereas vowels and other voiced sounds have strong cues to their identity in the lower formant-frequency regions, fricative and sibilant sounds, as well as consonant-release bursts, have identifying features at much higher frequencies. These different phonetic classes surely have different frequency distributions of importance, and indeed band-importance functions are differently shaped for vowels versus consonants (Apoux and Healy, 2012). But it also seems reasonable to assume that differences in importance-function shape attributable to speech material should be considerably

[a] Current address: Utah State University, Logan, UT 84322, USA. Electronic mail: sarah.leopold@usu.edu

[b] Current address: Saint Mary's College at Notre Dame, Notre Dame, IL 46556, USA.

[c] Current address: Department of Otolaryngology—Head & Neck Surgery, The Ohio State University, Columbus, OH 43212, USA.

diminished once the array of different phonemes included in the corpus becomes sufficiently large. It could be reasonably argued that the standard sentence and word-list corpora on which band-importance functions are typically based should have sufficient phonetic diversity to overcome this aspect of the speech-material effect.

With regard to a potential talker effect, it is well known that the acoustic manifestation of speech depends largely on vocal-fold and tract size and therefore varies across individuals (Peterson and Barney, 1952). It is not unreasonable to expect the acoustic characteristics of an individual's voice to play a substantial role in the particular frequencies that are most important for understanding his or her speech. However, the development of SII band-importance functions from single-talker productions and the common extrapolation of these functions to other recordings of the same speech materials assumes that the talker effect is largely absent.

There has been some acknowledgement of these speech-material and talker effects in prior work employing the ANSI technique (ANSI, 1969, 1997). The potential for a strong talker effect is clearly articulated in early writings, "To obtain a desirable precision in the measurement of articulation, it is advisable to use at least five different voices…," (Fletcher and Steinberg, 1929; Fletcher and Galt, 1950; both in Fletcher, 1995, pp. 278–279). However, Studebaker *et al.* (1987) concluded that the difference between band-importance functions for different individual male and female talkers was smaller than the difference between functions for different speech materials. It was suggested that the use of masking noise matched to the talker should remove the influence of the talker. Studebaker and Sherbecoe (1991) acknowledged that all of their W-22 word-list recordings were made by the same talker and suggested that further data were needed to confirm the conclusion of Studebaker *et al.* (1987) that the talker has little, if any, influence. Bell *et al.* (1992) examined speech material while controlling for the talker, by having the same talker produce both high- and low-context sentences. Although a small difference in the crossover frequency was observed, the overall shapes of the importance functions were similar. This result could potentially be interpreted to contrast that of Studebaker *et al.* (1987) by suggesting that the similarity in function shape was driven by a similarity in talker characteristics.

In the current study, the relative strengths are examined of speech-material and talker effects on the shape of speech band-importance functions. This examination is prompted by the recent use of the compound method to determine band importance (Apoux and Healy, 2012; Healy *et al.*, 2013; Bosen and Chatterjee, 2016). It has been suggested that this method offers a more detailed view of the importance of different speech bands than do traditional methods, including the ANSI technique. This suggestion is based on the observation of reliable variations within a band-importance function ("microstructure"), in which adjacent bands can display considerably different importance (Healy *et al.*, 2013). The ability to produce functions that display more detail than those produced by the standard method both allows and necessitates the current re-examination of the potential influences of speech material and talker on the band-importance function. One possibility is that the compound-method functions more strongly reflect acoustic characteristics of the talker than do functions generated using the ANSI technique.

If the speech-material effect is dominant, as has been traditionally suggested, then it is essential that band-importance functions be established for each type of speech corpora. However, if the talker effect is of substantial influence, then the band-importance function created using one talker may not generalize to recordings from other talkers, even of the same speech corpus. A solution to a strong talker effect may involve the creation of band-importance functions based on materials spoken by numerous talkers, as the early founders intended, but as not commonly practiced today.

II. EXPERIMENT 1. SINGLE VS TEN-TALKER SENTENCES

In this experiment, the goal was to determine the influence of using multiple talkers to create the speech band-importance function. Sentences from the IEEE database (IEEE, 1969) were selected to represent the standard multi-talker sentence database, and the function derived from a single talker was compared to that derived for identical materials produced by multiple talkers. Of particular interest was the relative smoothness of the functions. If the function representing the single talker was substantially less smooth than that for the multiple talkers, then evidence for a talker effect on the speech band-importance function would be provided, because the variations in the single-talker function may potentially reflect acoustic idiosyncrasies of that voice.

A. Method

1. Subjects

Sixty normal-hearing listeners between the ages of 19 and 37 (mean = 21.8) years participated in this experiment. Fifty-five were female.[1] The listeners were recruited from courses at The Ohio State University and received course credit or a monetary incentive. All had pure-tone audiometric thresholds at or below 20 dB hearing level at octave frequencies from 250 to 8000 Hz (ANSI, 2004, 2010), all were native English speakers, and none had previous exposure to the speech materials used.

2. Stimuli and procedure

The IEEE-sentence corpus contains 720 sentences, and each sentence contains five scoring key words. The original 22 kHz, 16-bit recordings spoken by ten different talkers judged to have a general American dialect (5 male, 5 female) were used. For a multi-talker condition, all 10 talkers were employed. For a single-talker condition, one of the male talkers was chosen randomly.

The stimuli were filtered into the 21 critical-band divisions specified in the SII (see Table I). The filtering was essentially identical to that of Healy *et al.* (2013), which provided minimal band overlap and a high degree of acoustic band independence. The technique involved finite impulse

TABLE I. Band divisions employed for all functions.

Band	Center Frequency (Hz)	Band Limits (Hz)	Effective Filter Order
1	150	100–200	20 000
2	250	200–300	20 000
3	350	300–400	20 000
4	450	400–510	20 000
5	570	510–630	20 000
6	700	630–770	18 000
7	840	770–920	16 000
8	1000	920–1080	14 000
9	1170	1080–1270	12 000
10	1370	1270–1480	11 000
11	1600	1480–1720	9500
12	1850	1720–2000	8500
13	2150	2000–2320	7500
14	2500	2320–2700	6500
15	2900	2700–3150	6000
16	3400	3150–3700	5000
17	4000	3700–4400	4500
18	4800	4400–5300	3500
19	5800	5300–6400	3000
20	7000	6400–7700	2000
21	8500	7700–9500	2000

response filters with effective orders ranging from 2000 (for the highest frequency bands) to 20 000 (for the lowest frequency bands). This filtering produced equal slopes across the spectrum of approximately 8000 dB/octave, when measured from cutoff to noise floor. Due to limitations associated with filtering in the low spectral region, slope values decreased somewhat below 500 Hz. However, values remained over several thousand dB/octave at 300 Hz and were approximately 1000 dB/octave at 100 Hz. Transition bandwidths below 500 Hz remained in the 3–5 Hz range. The only difference between the filtering employed currently and that of Healy *et al.* (2013) is the current use of filtering in the forward and backward direction to eliminate group delays and ensure exact temporal synchrony across bands. Thus, the correction for group delays performed in the prior study was unnecessary. The effective filter order is double the input order during bi-directional filtering, which was accounted for currently by halving the input orders. This processing was performed in MATLAB.

Subjects were randomly divided into three groups. The first group was assigned target bands 1–7, the second group was assigned target bands 8–14, and the third group was assigned target bands 15–21 (after Healy *et al.*, 2013). In accord with the compound method, the target band was always presented with four "other" bands. The number of other bands was determined during pilot testing to ensure intelligibility scores that avoided floor and ceiling effects. The frequency locations of the other bands were selected randomly for each trial. Trials were paired such that in one trial, the target band was present along with four other random bands and in the other trial the same four other bands were presented without the target band. This pairing procedure allowed the importance of the target band to be assessed in a controlled manner and in the presence of many combinations of other spectral bands. It also allowed for a simplification of the weight calculation. [See Apoux and Healy (2012) and Healy *et al.* (2013) for a more detailed description of the compound method.]

There were 14 conditions heard by each listener (7 target bands × 2 number-of-talkers). Subjects heard 20 sentences in each of these conditions for a total of 280 sentences (IEEE sentences 1–200 and 501–580). Half of these 20 sentences in each condition were target-band present and half were target-band absent. To obtain these 10-sentence sets in the multi-talker conditions, one sentence was presented from each of the ten talkers in random order. The individual talker was randomly selected for each listener and trial, and the frequency positions of the other bands were randomly selected for each listener and pair of trials. Half of the subjects heard the single-talker conditions followed by the multi-talker conditions, and the other half heard the reverse order. Target-band conditions were blocked so that all sentences in one target-band condition were completed before moving on to the next. Trials were paired and randomized such that a band-present trial and band-absent trial with the same other bands were contiguous. The order in which target-band conditions appeared and the condition-to-sentence correspondence was randomized for each subject.

Broadband sentences were set to play back at 70 dBA at each earphone using a flat plate coupler (Larson Davis AEC 101, Depew, NY) and ANSI Class 1 sound level meter (Larson Davis 824). The relative spectrum level of each band was maintained. The speech stimuli were converted to analog form using a PC and Echo Gina 3G D/A converters (Santa Barbara, CA), and presented diotically via Sennheiser HD 280 circumaural headphones (Wedemark, Germany).

Testing was performed in a double-walled IAC sound booth. A brief familiarization was conducted in which 20 sentences not used for formal testing (IEEE 701–720) were presented. The first five sentences were presented broadband, followed by five sentences consisting of 11 bands randomly selected for each trial, and finally ten sentences consisting of four bands randomly selected for each trial. Subjects responded after each trial by repeating as much of the sentence as possible to the experimenter, and were given correct/incorrect feedback during familiarization but not during formal testing. Subsequent to familiarization, subjects heard the 14 blocks of 20 sentences. The experimenter recorded the key words correctly repeated for each trial. Presentation of stimuli and collection of responses were performed using custom MATLAB scripts running on a PC. The total duration of testing was approximately 1 h and subjects were required to take a break half way through the experiment.

B. Results and discussion

The average intelligibility score for the single-talker conditions for band present was 62.7% (st. dev. = 4.5) and for band absent was 46.8% (st. dev. = 3.1). The average intelligibility score for the multiple-talker conditions for band present was 53.8% (st. dev. = 4.2) and for band absent was 37.8% (st. dev. = 3.6). The importance of each target band was calculated for the single- and ten-talker conditions according to the method of Apoux and Healy (2012). First, the band-present and band-absent scores were averaged across subjects for each target band. Then, the mean band-absent score was subtracted from the mean band-present score to create a mean difference score for each band. The difference scores were then normalized by dividing each by the sum of the 21 band-difference scores.

Figure 1 shows band importance for the single-talker and ten-talker IEEE stimuli. In general, the function for the ten-talker stimuli appears smoother than the function for the single-talker stimuli, especially in the upper half of the spectrum. In addition, there is a slight up-shift in the frequencies of greatest importance in the ten-talker function, likely reflecting the inclusion of female talkers having higher format frequencies.

To quantify the relative smoothness of the one- versus ten-talker functions, a Gaussian stochastic process was employed (Santner *et al.*, 2003). This model fits a curve across the frequency bands and computes the point-to-point correlation across subsequent bands for each function. The scale parameter theta is then used to indicate the overall smoothness of each function. A smaller scale parameter indicates a weaker correlation, which in turn indicates a smoother function. The Gaussian correlation function is given below in Eq. (1). The estimated scale parameter for the single-talker condition ($\theta = 0.659$) was larger than that for the multi-talker condition ($\theta = 0.493$), indicating that the multi-talker function is indeed smoother than the single-talker function:

$$R(x_i - x_j) = \exp\left[-\sum_k \theta_k (x_{ik} - x_{jk})^2\right]. \qquad (1)$$

In support of a supplementary examination of talker effects, data comprising the ten-talker condition were split into subgroups of five male and five female talkers. Band-importance functions were calculated for each of these gender subgroups and are presented in Fig. 2. Apparent is the considerable transposition in frequency across the two functions, likely corresponding to the different average frequency compositions of male versus female voices. Note that these functions cannot be used in an analysis of smoothness because each is composed of half the data involved in the main single- and ten-talker conditions.

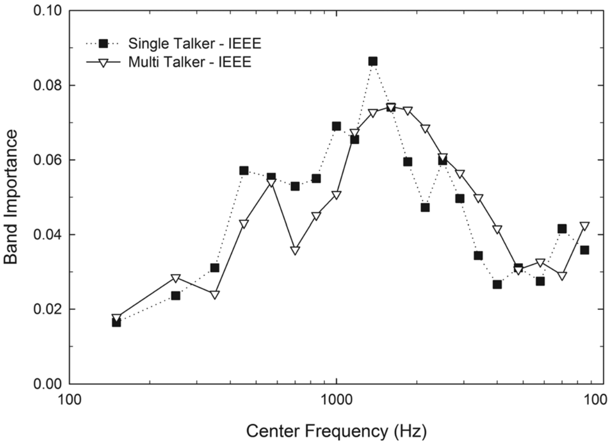

FIG. 1. Experiment 1: Effect of multiple talkers. Band-importance functions representing IEEE sentences. The closed symbols show functions for materials spoken by a single male talker, and the open symbols show functions for materials produced by multiple talkers (five male, five female).

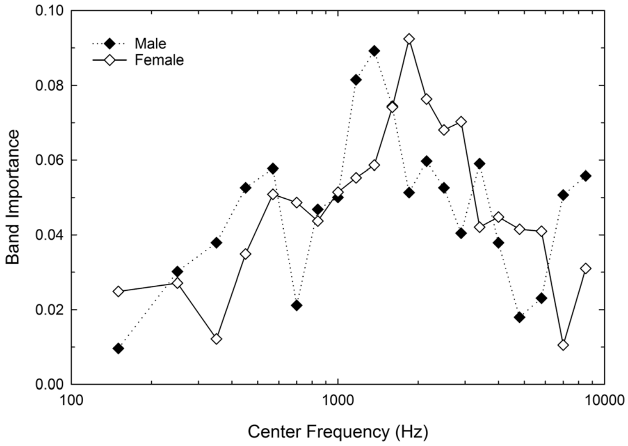

FIG. 2. Experiment 1: Effect of talker gender. Band-importance functions for the five male and five female talkers involved in the multiple-talker condition displayed in Fig. 1.

III. EXPERIMENT 2. DIFFERENT TALKERS

In this experiment, the goal was to determine the influence of using different talkers to create band-importance functions for the same speech materials. A novel talker was used to create recordings of the Speech Perception in Noise (SPIN) test (Kalikow *et al.*, 1977) sentences, selected to represent standard sentence materials on which prior band-importance studies have been based. The function derived from this novel talker was compared to that based on identical speech materials, but from a different talker (the standard SPIN recording). Further, this comparison was extended to a function derived from a third talker, and for materials (IEEE sentences) that are in the same broad class ("sentences") but that differ in several ways (keyword scoring versus final-word scoring, semantic predictability versus mixed predictability, etc.). The same compound method was employed for all functions. Relative similarity across these functions would provide support for a speech-material effect on the shape of the speech band-importance function. Relative dissimilarity would support a talker effect on the shape of the speech band-importance function.

A. Method

1. Subjects

Sixty normal-hearing listeners between the ages of 19 and 31 (mean = 20.9) years participated in this experiment. Fifty-eight were female, and none had participated in experiment 1. The recruitment procedures, incentives, hearing criteria, native language, and previous-exposure characteristics were all the same as in experiment 1.

2. Stimuli and procedure

The speech materials used in this experiment were from the revised version of the SPIN. The test consists of 200 key words, each positioned as the final word in both a high- and low-predictability context sentence. For the purposes of the current study, a new recording of these materials was created using a male speaker having a general American dialect. The recordings were made in a double-walled IAC sound booth using a large-diaphragm condenser microphone having a flat frequency response (AKG C2000B) fitted with a commercial windscreen. The microphone was preamplified (Mackie 1202-VLZ, Woodinville, WA) and digitally recorded (Echo Gina 3G) at 44.1 kHz with 16-bit resolution. The talker sat 12 in. from the microphone and read the list of sentences twice. Recordings were monitored to ensure adequate gain and that no peak clipping occurred. A single production was selected for each sentence based on clarity, and the root-mean-square average level of each sentence was equated within 1 dB.

The processing of stimuli and experimental procedures of this experiment were essentially identical to those employed to create a band-importance function for the standard recordings of the SPIN sentences by Healy *et al.* (2013). The new recordings were first subjected to the same band divisions and filtering, and band-importance method as in the current experiment 1. Each subject heard 56 sentences in each of the seven target-band conditions for a total of 392 sentences. Half of the 56 sentences were high predictability and half were low predictability, and for each paired band-present/band-absent trial, the predictability was the same. The use of both high- and low-predictability SPIN sentences was based on the observation that both predictability subsets produce similarly shaped compound-method band-importance functions (Healy *et al.*, 2013), and the prior assertion that a single band-importance function can be used to represent both predictability subsets (Bell *et al.*, 1992).

Broadband sentences were set to play back diotically at 70 dBA at each earphone using the apparatus and procedures from experiment 1. A brief familiarization consisted of eight sentences not used for formal testing, presented first broadband and then repeated as five spectral bands, randomly selected from trial-to-trial. Subsequent to familiarization, subjects heard the seven blocks of test sentences. Subjects responded after each trial by typing the final word of the sentence on a custom MATLAB computer interface, and correct-incorrect feedback was given for familiarization only. The decision was made to have subjects type their own responses during this experiment, because only the final word was reported and scored, in accord with the established SPIN-test format. Testing was performed in a double-walled IAC sound booth, with a total duration of approximately 2 h and breaks were offered after every block.

B. Results and discussion

The average intelligibility score for band present was 65.5% (st. dev. = 4.5) and for band absent was 54.6% (st. dev. = 2.6). The importance of each band was calculated according to the method used in experiment 1. Figure 3 shows the band-importance function created for the novel SPIN talker employed in the current experiment. Also plotted is the function representing the standard male-talker recording of the SPIN sentences from Healy *et al.* (2013), and the IEEE single-talker function from experiment 1.

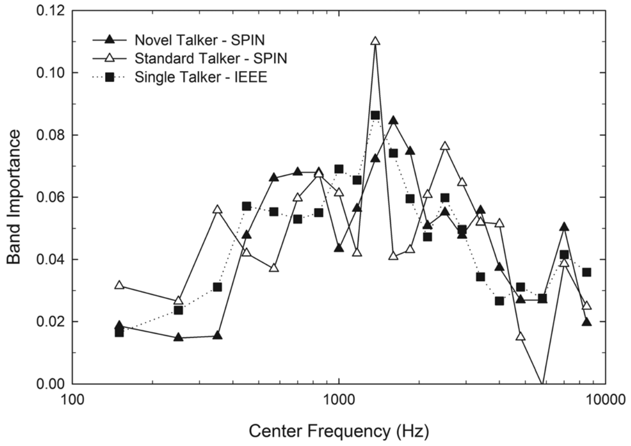

FIG. 3. Experiment 2: Effect of different talkers. Band-importance functions all representing sentence materials but created using different individual male talkers.

Of interest is the relative similarity between the three functions, despite the use of three different talkers. Most notably, all share a peak in importance at approximately 1500 Hz (and by definition have reduced importance on either side), another peak at approximately 2500 Hz, a broad drop in importance across 2500–6000 Hz, and a final peak at approximately 7000 Hz.

IV. EXPERIMENT 3: DIFFERENT MATERIALS

In this experiment, the goal was to determine the influence of using the same talker to create band-importance functions for speech materials across different broad classes (sentences versus word lists). The same novel talker employed in experiment 2 was used to create a function for the CID W-22 words (Hirsh *et al.*, 1952). This function was compared to that derived from the same talker for the SPIN sentences, using the same compound method. Relative similarity across these functions would support a talker effect on the shape of the speech band-importance function. Relative dissimilarity in these functions would support a speech-material effect on the shape of the function. The degree of relative similarity observed across functions in experiment 2 versus experiment 3 provides an indication of the relative strength of the speech-material versus talker effect on the shape of the band-importance function.

A. Methods

1. Subjects

Sixty normal-hearing listeners between the ages of 19 and 31 (mean = 20.8) years participated in this experiment. Fifty-seven were female. One previously participated in experiment 1 and 28 previously participated in experiment 2. Subjects completing more than one experiment did so on different days, separated by two or more weeks. The recruitment procedures, incentives, hearing criteria, native language, and previous-exposure characteristics were all the same as in experiments 1 and 2.

2. Stimuli and procedure

The speech materials were the phonetically balanced W-22 words. The original corpus contains 200 words produced by a male speaker having a general American dialect and set in the carrier phrase, "Say the word ___." For the purposes of the current experiment, a new recording was made using the same talker and procedure as for the SPIN sentences in experiment 2.

A band-importance function was created for these materials using the band divisions, filtering, and procedures of experiment 2. Each subject heard 26 words in each of the seven target-band conditions. Half (13) of the trials were target-band present and the other half were target-band absent. Familiarization prior to testing included 15 words not heard during the test, heard first broadband and then as five bands randomly distributed in frequency for each trial. Subjects typed responses into a custom MATLAB interface and received feedback on response accuracy during familiarization only. Due to the open-set nature of monosyllable word identification, homophones of the target word were accepted for this experiment. The calibration and presentation apparatus, and the procedures, were all the same as in experiments 1 and 2. Testing lasted approximately 1 h.

B. Results and discussion

The average intelligibility score for band present was 57.6% (st. dev. = 4.9) and for band absent was 44.8% (st. dev. = 3.6).[2] The importance of each band was calculated according to the method used in experiments 1 and 2. Figure 4 shows band importance for the CID W-22 words spoken by the novel talker employed currently, as well as the function representing the SPIN sentences produced by the same talker from experiment 2. Of note are the relative differences between the functions. Specifically, the peak in importance present at 700 Hz for the words is absent for the sentences; a smaller peak at 1000 Hz exists for words, whereas sentences display a sharp valley; and a peak at 1600 Hz for sentences is absent for words. Substantial differences in importance exist across materials in the 2000- to 4000-Hz region, and opposite peaks and valleys exist across the two speech materials from approximately 5000 to 8500 Hz. The band of greatest importance for the CID W-22 words had a center frequency of 700 Hz, whereas the band of greatest importance for the SPIN sentences had a center frequency of 1600 Hz.

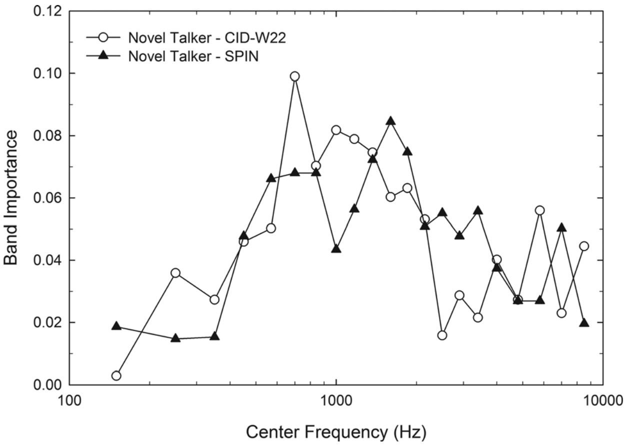

FIG. 4. Experiment 3: Effect of different speech materials. Band-importance functions representing different classes of speech materials (CID W-22 words and SPIN sentences) but both spoken by the same talker.

To quantify whether the region of maximum importance was different for the comparisons involving different talkers/same materials (experiment 2) and same talker/different materials (experiment 3), linear regression was used. Two regression models assessed the differences between the three functions for sentences with different talkers (model 1) and between sentences and words with the same talker (model 2). To prepare the data for the linear regression models, each individual was randomly assigned to a group for each region, talker, and material combination. For each group and condition (talker and material), the region with the highest importance was found, providing maximum importance measures for each group, talker, and material. If there was a tie (i.e., two bands had the same importance), both were kept for the analysis. These steps were undertaken to (a) control for the clustering of individuals (each individual had an equal chance of being in any of the groups for each region/talker/material combination) and (b) to enable the direct assessment of the region of maximum importance within linear regression. To further confirm that the groups were approximately independent, the intra-group correlation was assessed as well. $N = 127$ data points remained for the statistical analysis.

Table II presents the F statistic and p-value of the overall models, the estimated difference in the region of maximum importance (in band-number units), and the 95% confidence interval of the difference associated with each comparison. Only one of the comparisons approached significance at the 0.05 level and was statistically significant at $p < 0.10$—the model 2 comparison between sentences and words with the same talker ($p = 0.078$). As expected, the comparison between sentences with different talkers was not statistically significant ($p = 0.524$). This indicates that the region of maximum importance is marginally significantly different for the functions in experiment 3 representing sentences and words spoken by the same talker, but not for the three functions in experiment 2 representing sentences spoken by different talkers.

Additionally, the intra-group correlations were all below $r = 0.01$ suggesting that the randomization of the groupings helped to alleviate any lack of independence, adding evidence to the appropriateness of linear regression. The other assumptions of linear regression (i.e., normality, homoscedasticity) were also assessed and showed no meaningful deviations.

V. GENERAL DISCUSSION

The characteristics of the compound method and the functions resulting from it facilitate the current re-evaluation of potential influences on speech band importance. The functions derived currently using a single talker display a considerable amount of microstructure, in accord with previous

TABLE II. Results for the two linear regression models. Model 1 is the comparison of sentences with different talkers, and model 2 is the comparison between sentences and words with the same talker. Values within parentheses are the 95% confidence intervals.

	Model 1	Model 2
Variable	n = 72	n = 55
Material		
Sentences	—	[reference]
Words	—	2.621 (−0.305, 5.547)
Talker		
CID	[reference]	—
IEEE	1.369 (−1.057, 3.795)	—
SPIN	0.929 (−1.497, 3.355)	—
F-Statistic	0.653 (df = 2; 69)	3.229 (df = 1; 53)
P-Value	0.524	0.078

examinations (Apoux and Healy, 2012; Healy *et al.*, 2013). In contrast, the function for the ten-talker IEEE stimuli is somewhat less jagged, particularly in the 1000- to 5000-Hz region. The band of highest importance is also shifted up in frequency by one band in the ten-talker function. An obvious interpretation of this latter result involves the inclusion of female voices in the ten-talker stimulus set, whereas the single talker was male.

There are two possible interpretations for the difference in smoothness between functions in experiment 1. In the first, the more pronounced peaks and valleys of the single-talker function reflect particular acoustic characteristics of the individual voice, and the inclusion of multiple talkers averages out those idiosyncrasies across talkers. In the second interpretation, the variability in talker from trial-to-trial resulted in listeners monitoring frequencies more broadly, and placing less emphasis on any individual band. Data from Assgari and Stilp (2015) suggest that listeners are less sensitive to modest differences in spectral peaks when forced to continually recalibrate to new talkers. There are also differences in higher-level processing of single- versus multiple-talker sentence lists, as exemplified by improved intelligibility of novel utterances by familiar versus less familiar talkers (Nygaard *et al.*, 1994) and an increased demand on working memory in a multiple-talker context (Mullennix *et al.*, 1989).

A pair of supplementary analyses was undertaken to further examine potential effects of talker on the speech band-importance function. The first analysis involved an attempt to identify the individual talkers represented in Fig. 3 based on typical acoustic characterizations of the human voice. All functions in Fig. 3 represent sentence materials and male talkers, so this analysis represents a detailed examination of talker. The fundamental frequency ($F0$) and first three formant frequencies ($F1$, $F2$, and $F3$) were calculated for the first 25 sentences from each talker, using Praat software (Boersma and Weenink, 2011) and an upper formant-frequency limit of 5000 Hz. These data are presented in Table III.[3] In addition to these primary spectral characteristics, $F0$ variation, amplitude variation, and syllable rate are provided.

TABLE III. Acoustic characteristics of the talkers employed. Shown are values for the three talkers used to create the band-importance functions displayed in Fig. 3 and the mean values for the five male and five female talkers in Figs. 1 and 2. Shown are fundamental frequencies ($F0$) and the first three formant frequencies ($F1$, $F2$, and $F3$), all in Hz. Also shown is the standard deviation (SD) of the $F0$ variations, the SD of the amplitude envelope in dB, and the syllable rate.

Condition	$F0$	$F1$	$F2$	$F3$	$F0$ SD	Amp SD	Syllables/ sec
Novel Talker—SPIN	108	646	1621	2555	35.6	13.7	3.6
Standard Talker—SPIN	120	726	1765	2742	25.5	14.1	4.1
Single Talker—IEEE	118	724	1776	2653	31.7	15.2	3.6
Male IEEE Talker Mean	116	647	1710	2620	25.2	15.5	3.8
Female IEEE Talker Mean	217	723	1969	3018	52.5	14.9	3.5

A comparison of the primary spectral talker characteristics and their corresponding band-importance functions in Fig. 3 yields little correspondence. A primary difference between the functions involves the upward transposition of the peak around 1500–2000 Hz for the novel SPIN talker relative to the other two talkers. But in contrast to what is observed in the functions, the $F2$ value for the novel SPIN talker is not higher than the others and is in fact the lowest of the three. This lack of ability to readily identify detailed talker characteristics in the speech band-importance function based on traditional spectral voice measures provides little support for a talker effect.

This lack of correspondence might reflect the spectral resolution with which the band-importance function is measured. Indeed, the $F1$ value for all three talkers in Fig. 3 falls within a single band. For $F2$ and $F3$, the values for two of the talkers fall within a single band, and that for the remaining talker falls in the adjacent band. This is true despite the use of the maximum SII resolution (21 bands).

A second supplementary analysis involved a more global analysis of the talker effect. The functions displayed in Fig. 2 are split by gender into male and female voices. Apparent is that these functions are generally similar in shape, but transposed in frequency relative to one another, with the female-voice function being higher. Thus, there is evidence that more global effects of talker, namely talker gender, influences the speech band-importance function. This global effect of talker likely explains the smoother function observed in experiment 1 for the ten-talker materials relative to the single-talker materials.

Potentially interesting is the magnitude of the function transposition relative to the magnitude of the difference between genders on traditional acoustic voice measures. These measures are also displayed in Table III and were calculated in the same fashion as the other values in that table. The most prominent peak in the Fig. 2 functions are transposed by two bands (band 10 versus band 12), which corresponds to a center-frequency difference of 480 Hz. This transposition far exceeds the gender difference observed for the primary spectral acoustic voice measures, and exceeds the difference in $F2$ by roughly a factor of 2.

Together, these analyses provide limited support for a strong effect of talker on the speech band-importance function. They suggest that traditional spectral measures used to characterize voices do not strongly characterize frequency band-importance functions for individual male talkers. This conclusion is in accord with the observation by Healy *et al.* (2013) of a general alignment between formant frequencies and band-importance function peaks, but a lack of close correspondence. The current analyses do suggest that the more global characteristic of talker gender is reflected in the band-importance function as a frequency transposition. But the magnitude of the transposition is larger than might be predicted based on these traditional measures.

To more directly examine the relative strength of the speech-material and talker effects on band importance, functions were created having the same or different speech-material types, and same or different talkers. Figure 3 displays a relatively high degree of similarity between the three functions reflecting the same speech-material type (sentences) but three different talkers. The available statistical analysis indicated that the region of maximum importance for the three functions is not statistically different. This finding suggests that there is a particular characteristic shape to band-importance functions for sentences, regardless of the voice used to make the recording—a strong speech-material effect.

Further, there is evidence that the speech-material effect generalizes across the broad class of speech-material type and is not specific to the particular speech corpus. This evidence comes from the similarity across functions representing different sentence corpora (IEEE and SPIN). Although these corpora belong to the same broad class "sentences," they differ in several ways, including scoring technique (multiple component key words for IEEE versus single final-word scoring for SPIN) and semantic predictability (predictable for IEEE versus mixed predictability for SPIN). But despite these considerable differences in the specific characteristics of the corpora, their common assignment to the broad class of sentences appears sufficient to drive the speech-material effect and produce similarity in the shape of the band-importance functions.

Additional evidence supporting the strength of the speech-material effect and the relative weakness of the talker effect is a relative dissimilarity in the functions representing the same talker but different materials. As Fig. 4 shows, the two functions for the very same individual producing speech materials in different broad classes (sentences versus word lists) vary considerably in the location of excursions. Indeed, for multiple regions of the spectrum, the locations of peaks and valleys seem to be in direct opposition. In addition, the location of greatest importance differs considerably, with that for the CID W-22 words 900 Hz lower than that for the SPIN sentences, and this difference is supported by statistical analysis.[4] These results together indicate that the details present in the band-importance functions resulting from the compound method do not simply reflect acoustic characteristics of the particular talker's voice, but rather depend more primarily on the type of speech material under evaluation.

One possible explanation for the observed differences in functions from the same talker producing different speech materials involves the contribution of top-down processing. Early articulation-testing work (Miller *et al.*, 1951) showed large differences in the articulation functions for sentences, nonsense syllables, and digits. The authors attributed these differences to the number of alternatives to the target in the testing set. In other words, stimuli such as digits and sentences have a restricted number of possible responses, digits due to their limited number of alternatives and sentences due to their grammatical and contextual constraints. Nonsense syllables on the other hand have a larger number of alternatives, therefore requiring heavier reliance on bottom-up acoustic information to be correctly identified. What is less clear is why differences in the reliance on bottom-up acoustic properties to identify a speech target would produce differences in frequency regions of importance.

Top-down factors can influence speech perception in a variety of ways. Whereas data exist to clarify the influences of these factors on overall intelligibility, the extent to which these factors can serve to differentially influence the weighting of different speech frequencies is far more poorly understood. It should be emphasized that these relationships are complex, and that future studies designed to examine them directly are required to provide clarity. But it is possible to speculate based on what is known.

Coarticulation exists in both typical sentences and word lists. It restricts the set of possible lexical alternatives based on articulatory trajectory and corresponding phonetic content. But the constant carrier phrase in typical word lists causes the coarticulation preceding the target word to be constant, therefore reducing coarticulatory variability. Stickney and Assmann (2001) found a slightly lower overall intelligibility for speech that had more constant preceding coarticulation, relative to the same speech items having more diverse preceding coarticulation. Because it has very direct and strong acoustic ramifications, coarticulation could differentially influence a listener's dependence on various speech frequencies. This possibility may serve to contribute to the differences in the function shape observed for sentences versus words.

With regard to sentences, grammatical context typically restricts word class (e.g., noun versus verb). Semantic context typically restricts more narrowly to a particular set of semantically plausible lexical entries. The extent to which, and manner in which, these specific factors serve to affect the shape of the band-importance function is not well known. But data from Healy *et al.* (2013) showing a general similarity in the shape of functions for low- versus high- predictability SPIN sentences suggest that the influence of semantic context is not strong. In contrast to sentences, word lists have neither grammatical nor semantic context. But the SPIN-sentence data just mentioned may suggest that the considerable differences in shape observed across the functions for words versus sentences is not driven largely by differences in semantic context.

It may be considered somewhat surprising that a speech-material effect dominates the shape of the band-importance function, and that the talker effect is weaker. Although this is a common modern assumption for the less-detailed ANSI functions, recall that the earliest formulators of the band-importance concept stressed the need for multiple talkers. Further, both the sentences and word lists employed currently possess considerable phonetic diversity. Thus, the seemingly reasonable assumption described in Sec. I that speech band-importance functions will become similar once a threshold amount of phonetic diversity is achieved does not appear to hold. Instead, it is the assembly of those phonetic entities into words or sentences that drives with more strength the importance of information contained at various frequencies.

Finally, it is noted that other techniques exist to assess speech band importance. These include but are not limited to the successive low-pass and high-pass filtering technique reflected in the SII (ANSI, 1997), the redundancy correction of Steeneken and Houtgast (1991), the correlational technique of Doherty and Turner (1996), and the joint optimization procedure of Kates (2013). All of these techniques have employed speech in background noise, and many rely on noise either indirectly (to control overall intelligibility level) or directly (to correlate the signal-to-noise ratio with intelligibility). One advantage of the currently used compound method is the ability to assess speech band importance either in quiet or in noise [also see discussions by Apoux and Healy (2012) and Healy *et al*. (2013)]. But the importance of speech in noise is also a topic of considerable interest and differences across functions created in quiet versus noise may be observed (see Yoho *et al*., 2017).

VI. CONCLUSIONS

Taken together, these findings support a strong influence of speech material and weaker influence of the talker on the shape of the highly detailed speech band-importance functions created using the compound method. The speech-material effect appears to generalize across corpora in the same broad class of "sentences" despite considerable differences in the particular aspects of those corpora. The talker effect does not appear strong enough to reflect acoustic aspects of individual talkers, and instead appears restricted to more global aspects of a talker, including gender. The use of multiple talkers, although not critical, appears to largely diminish any residual effect of a talker and smooth the functions slightly. These data suggest the need to generate different functions for different broad classes of speech materials, but perhaps not for every individual corpus. Further, the ability to generalize from one talker to others using the compound method appears to be relatively strong.

ACKNOWLEDGMENTS

This work formed portions of a dissertation submitted by S.E.Y., under the direction of E.W.H., in partial fulfillment of the degree requirements for the Ph.D. in Speech and Hearing Science at The Ohio State University. The work was supported in part by the National Institute on Deafness and other Communication Disorders (Grant Nos. R01 DC008594 and R01 DC015521 to E.W.H.). The authors are grateful for the data-analysis and manuscript-preparation assistance of Jordan Vasko and Brittney Carter.

[1] The gender balance employed in the current study (strong majority female) matches that used by Healy *et al.* (2013) to create band-importance functions used for comparison. Although there is little reason to believe that listener gender differentially affects the weighting of speech information across the spectrum, the current results may be seen as restricted to female listeners.

[2] The subjects in experiment 3 who also participated in experiment 2 displayed an overall intelligibility within 2.8 percentage points of their naive counterparts in experiment 3, suggesting that prior experience in an experiment employing different speech materials did not substantially affect performance.

[3] Values for the novel talker across SPIN sentence versus W-22 word productions were within 2% for $F0$ and within 6% for all three formant frequencies.

[4] An additional comparison is possible between the current novel-talker W-22 function and that created using similar procedures but the standard W-22 recordings in Healy *et al.* (2013). The same regression analysis employed in experiment 3 indicated that the region of maximum importance was not significantly different across these two functions, supporting a stronger speech-material effect and a weaker talker effect. But the apparent similarity between these functions is decreased somewhat relative to that for sentences by different talkers. This suggests that the increase in bottom-up processing associated with the lower context materials might be accompanied by an increase in reliance on bottom-up acoustic talker characteristics.

ANSI (**1969**). ANSI S3.5, *American National Standard Methods for the Calculation of the Articulation Index* (American National Standards Institute, New York).

ANSI (**1997**). ANSI S3.5 (R2007), *American National Standard Methods for the Calculation of the Speech Intelligibility Index* (American National Standards Institute, New York).

ANSI (**2004**). ANSI S3.21 (R2009), *American National Standard Methods for Manual Pure-Tone Threshold Audiometry* (American National Standards Institute, New York).

ANSI (**2010**). ANSI S3.6-2010, *American National Standard Specification for Audiometers* (American National Standards Institute, New York).

Apoux, F., and Healy, E. W. (**2012**). "Use of a compound approach to derive auditory-filter-wide frequency-importance functions for vowels and consonants," J. Acoust. Soc. Am. **132**, 1078–1087.

Assgari, A. A., and Stilp, C. E. (**2015**). "Talker information influences spectral contrast effects in speech categorization," J. Acoust. Soc. Am. **138**, 3023–3032.

Bell, T. S., Dirks, D. D., and Trine, T. D. (**1992**). "Frequency-importance functions for words in high- and low-context sentences," J. Speech Hear. Res. **35**, 950–959.

Boersma, P., and Weenink, D. (**2011**). "Praat: Doing phonetics by computer (Version 4.3.22) [computer program]," http://www.praat.org (Last viewed February 9, 2018).

Bosen, A. K., and Chatterjee, M. (**2016**). "Band importance functions of listeners with cochlear implants using clinical maps," J. Acoust. Soc. Am. **140**, 3718–3727.

Doherty, K. A., and Turner, C. W. (**1996**). "Use of a correlational method to estimate a listener's weighting function for speech," J. Acoust. Soc. Am. **100**, 3769–3773.

Fletcher, H. (**1995**). *Speech and Hearing in Communication*, edited by J. B. Allen (Acoustical Society of America, Woodbury, NY), pp. 278–279.

Fletcher, H., and Galt, R. H. (**1950**). "The perception of speech and its relation to telephony," J. Acoust. Soc. Am. **22**, 89–151.

Fletcher, H., and Steinberg, J. C. (**1929**). "Articulation testing methods," Bell System Tech. J. **8**, 806–854.

Healy, E. W., Yoho, S. E., and Apoux, F. (**2013**). "Band importance for sentences and words reexamined," J. Acoust. Soc. Am. **133**, 463–473.

Hirsh, I. J., Davis, H., Silverman, S. R., Reynolds, E. G., Eldert, E., and Benson, R. W. (**1952**). "Development of materials for speech audiometry," J. Speech Hear. Disord. **17**, 321–337.

IEEE (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**, 225–246.

Kalikow, D. N., Stevens, K. N., and Elliott, L. L. (**1977**). "Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability," J. Acoust. Soc. Am. **61**, 1337–1351.

Kates, J. M. (**2013**). "Improved estimation of frequency importance functions," J. Acoust. Soc. Am. **134**, EL459–EL464.

Miller, G. A., Heise, G. A., and Lichten, W. (**1951**). "The intelligibility of speech as a function of the context of the test materials," J. Exp. Psychol. **41**, 329–335.

Mullennix, J. W., Pisoni, D. B., and Martin, C. S. (**1989**). "Some effects of talker variability on spoken word recognition," J. Acoust. Soc. Am. **85**, 365–378.

Nygaard, L. C., Sommers, M. S., and Pisoni, D. B. (**1994**). "Speech perception as a talker-contingent process," Psych. Sci. **5**, 42–46.

Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**, 175–184.

Santner, T. J., Williams, B. J., and Notz, W. I. (**2003**). *The Design and Analysis of Computer Experiments* (Springer, New York).

Steeneken, H. J. M., and Houtgast, T. (**1991**). "Mutual dependence of the octave-band weights in predicting speech intelligibility," Speech Commun. **28**, 109–123.

Stickney, G. S., and Assmann, P. F. (**2001**). "Acoustic and linguistic factors in the perception of band-pass-filtered speech," J. Acoust. Soc. Am. **109**, 1157–1165.

Studebaker, G. A., Pavlovic, C. V., and Sherbecoe, R. L. (**1987**). "A frequency importance function for continuous discourse," J. Acoust. Soc. Am. **81**, 1130–1138.

Studebaker, G. A., and Sherbecoe, R. L. (**1991**). "Frequency-importance and transfer functions for recorded CID W-22 word lists," J. Speech Hear. Res. **34**, 427–438.

Yoho, S. E., Healy, E. W., and Apoux, F. (**2017**). "How susceptibility to noise varies across speech frequencies," J. Acoust. Soc. Am. **141**, 3819.