Adaptive bandwidth measurements of importance functions for speech intelligibility prediction

Nathaniel A. Whitmal, III^{a)} and Kristina DeRoy^{b)}

Department of Communication Disorders, University of Massachusetts, Amherst, Massachusetts, 01003

(Received 6 October 2010; revised 23 August 2011; accepted 30 August 2011)

The Articulation Index (AI) and Speech Intelligibility Index (SII) predict intelligibility scores from measurements of speech and hearing parameters. One component in the prediction is the “importance function,” a weighting function that characterizes contributions of particular spectral regions of speech to speech intelligibility. Previous work with SII predictions for hearing-impaired subjects suggests that prediction accuracy might improve if importance functions for individual subjects were available. Unfortunately, previous importance function measurements have required extensive intelligibility testing with groups of subjects, using speech processed by various fixed-bandwidth low-pass and high-pass filters. A more efficient approach appropriate to individual subjects is desired. The purpose of this study was to evaluate the feasibility of measuring importance functions for individual subjects with adaptive-bandwidth filters. In two experiments, ten subjects with normal-hearing listened to vowel-consonant-vowel (VCV) nonsense words processed by low-pass and high-pass filters whose bandwidths were varied adaptively to produce specified performance levels in accordance with the transformed up-down rules of Levitt [(1971). *J. Acoust. Soc. Am.* **49**, 467–477]. Local linear psychometric functions were fit to resulting data and used to generate an importance function for VCV words. Results indicate that the adaptive method is reliable and efficient, and produces importance function data consistent with that of the corresponding AI/SII importance function. © 2011 Acoustical Society of America. [DOI: 10.1121/1.3641453]

PACS number(s): 43.71.An, 43.71.Es, 43.71.Gv [TD]

Pages: 4032–4043

I. INTRODUCTION

Theoretical models of speech perception can provide insight for researchers and clinicians concerned with the consequences of hearing impairment on speech intelligibility. Two widely used models are the Articulation Index, or “AI” (ANSI, 1969) and its successor, the Speech Intelligibility Index, or “SII” (ANSI, 1997). The underlying principle of the AI is that disjoint frequency regions make independent and additive contributions to intelligibility, as expressed in the equation

$$AI = \int_{f=0}^{\infty} I(f)W(f)df. \quad (1)$$

Here, $W(f)$ is a measure of the audible speech peak energy at frequency f , and $I(f)$ is a weighting function denoting the “importance” of frequency f . Values of $W(f)$ and $I(f)$ range between 0 and 1 (no contribution and maximum contribution to intelligibility, respectively). Measuring $I(f)$ has typically required time-intensive intelligibility testing with groups of subjects. The present work focuses on efficient measurement of importance functions (IFs) for modeling intelligibility in individuals with either normal or impaired hearing.

Early AI applications (Kryter, 1962; ANSI, 1969) used the Bell Telephone Laboratories IFs, which were derived from testing with nonsense syllable stimuli that were either high-

pass or low-pass filtered with a range of cutoff frequencies. Covariations in intelligibility and filter cutoff-frequency were converted to IFs, using either graphical curve bisection methods (French and Steinberg, 1947) or differential calculus (Fletcher and Galt, 1950). In both cases, the IF was deemed applicable to all types of test materials (e.g., word or sentence lists) and all subjects with presumably normal hearing. The AI values computed with these IFs were then input to empirically-derived “transfer functions” (TFs) that converted them into intelligibility score predictions for other types of test materials.

Later work (Pavlovic, 1984; Kamm *et al.* 1985; Pavlovic *et al.*, 1986) showed that AI predictions for hearing-impaired subjects were inaccurate. The SII (ANSI, 1997) attempted to correct these inaccuracies through use of test-specific IFs that were multiplied by “speech desensitization factors” ranging in value from 0 to 1 to account for individual hearing losses (Pavlovic *et al.*, 1986; Studebaker and Sherbecoe, 1993; Pavlovic, 1994). Ching *et al.* (1998) showed that remaining inaccuracies in the SII’s predictions for hearing-impaired subjects could be improved by augmenting the individual terms of Eq. (1) with subject-specific correction factors. Algebraically, these approaches are similar to using subject-specific IFs, even though all computations used the ANSI IFs and no intention to use subject-specific IFs was reported.

One likely reason that Ching *et al.* (1998) did not try to measure subject-specific IFs is that IF measurement can be time-consuming and tedious, as has been shown by investigators measuring IFs and TFs for non-standard materials. The measurements of Henry *et al.* (1998) for CVC word stimuli required 12 subjects to listen to 50-word lists for

^{a)}Author to whom correspondence should be addressed. Electronic mail: nwhitmal@comdis.umass.edu.

^{b)}Current address: Department of Speech, Language, and Hearing Sciences, Purdue University, West Lafayette, IN, 47907.

each combination of sixteen filter settings and four signal-to-noise ratios (SNRs). Eisenberg *et al.* (1998) required data from 20 subjects listening to 24-sentence lists in seven filter conditions simply to confirm that existing IFs could be used with HINT sentences (Nilsson *et al.*, 1994). Derivation of a HINT sentence transfer function required data from ten extra subjects listening to 25-sentence lists at six SNRs. Wong *et al.* (2007) required two stages of testing to obtain an IF and TF for Cantonese HINT sentences (Wong and Soli, 2005): a pilot study with six subjects who listened to 13 filter/SNR conditions, and a follow-up study with 78 subjects who listened to random selections of 115 different filter/SNR conditions. Further extension of the SII to new test materials will require similar measurements. A more efficient approach would encourage investigators to extend the SII to new materials and, if used with individual subjects as done indirectly by Ching *et al.* (1998), could improve the accuracy of the SII. Such improvements would also support short-time adaptations of the SII algorithm (Rhebergen and Versfeld, 2005; Rhebergen *et al.*, 2006) which improve SII predictions for speech masked by fluctuating noises. Currently, these extensions (which use ANSI IFs) work well for normal hearing listeners (Rhebergen *et al.*, 2008) but make inaccurate predictions for individual hearing-impaired listeners (Rhebergen *et al.*, 2010). Other short-time methods using importance functions built from instantaneous signal values (Ma *et al.*, 2009) have also been shown to improve the accuracy of speech quality and intelligibility prediction measures.

The inefficiency of IF measurement is common to psychophysical procedures that use a constant-stimulus paradigm. One possible alternative is an adaptive psychophysical procedure that uses information from previous responses to steer subject responses towards desired performance levels, rather than levels far above or far below that contain superfluous information (Treutwein, 1995; Leek *et al.*, 2001). In principle, an adaptive procedure that varied the bandwidth of speech to target specific intelligibility performance levels could efficiently measure psychometric functions relating intelligibility to filter bandwidth. These psychometric functions could in turn be used to produce the IF using the methods of French and Steinberg (1947) or Fletcher and Galt (1950).

Noordhoek *et al.* (1999) demonstrated the utility of adaptive bandwidth measurements for constructing psychometric functions. Their test subjects heard Dutch sentence recordings (Plomp and Mimpen, 1979) processed by a 1 kHz bandpass filter with bandwidth varied adaptively in 0.45-octave steps to produce 50%-correct sentence recognition. The resulting bandwidth was called the speech reception bandwidth threshold, or “SRBT.” The percent-correct data for each subject’s sentences were then plotted versus the sentences’ “relative bandwidths” (the bandwidth the sentence was presented at divided by the subject SRBT) to produce a psychometric curve relating bandwidth to intelligibility. This result, while encouraging, was not explored further by those authors. Subsequent applications of SRBT measurements focused on suprathreshold deficits in hearing-impaired listeners (Noordhoek *et al.*, 2000, 2001; van Schijndel *et al.*, 2001a, 2001b) or perceptual integration of speech bands (Hall *et al.*, 2008), rather than psychometric functions.

Moreover, the authors only targeted the 50%-correct performance level with their stimuli; as a result, their results contain little reliable data for accurately determining importance at the ends of their usable frequency range.

The purpose of the present study is to see whether an adaptive-bandwidth approach would be capable of measuring IFs in individual subjects. Two experiments were conducted. Experiment 1 measured recognition of vowel-consonant-vowel (VCV) words processed by low-pass filters (or, “LPFs”) whose cutoff frequencies were adapted using transformed up-down rules (Levitt, 1971) to produce five levels of performance. Measured bandwidths for each level were consistent across subjects and showed high test-retest reliability. The VCV recognition tests were repeated with non-adaptive filters using the average cutoff frequencies of the adaptive measurements. Statistical analyses comparing values and psychometric functions obtained in adaptive and fixed measurement modes show that adaptive-mode results are not significantly different from (and less variable than) fixed-mode results. Experiment 2 repeated the tasks of Experiment 1 with high-pass filters (or, “HPFs”). Those data revealed only minor differences between adaptive and fixed modes, with higher variability inherent to HPF speech perception. Comparisons of derived IF data with that of French and Steinberg (1947) support the viability of the proposed method.

II. EXPERIMENT 1 – IMPORTANCE MEASUREMENT WITH LOW-PASS FILTERS

A. Methods

1. Subjects

Seven subjects between the ages of 18 and 22 years (mean age: 20.55 years) participated in this experiment. All of the subjects were native speakers of English and passed a hearing screening for thresholds at or below 20 dB HL at 500, 1000, 2000, 4000, and 6000 Hz. Partial course credit and/or monetary compensation were provided to subjects in exchange for their participation.

2. Materials

Stimuli consisted of the 23 consonants /b d g p t k f θ v ð h s ʃ z ʒ tʃ ʤ m n w l j r/, recorded in /a/C/a/ format for a study by Whitmal *et al.* (2007). The consonants were spoken by a female talker with an American English dialect and digitally recorded in a sound-treated booth (IAC 1604, Bronx, NY) with 16-bit resolution at a 22050 Hz sampling rate.

3. Processing

Stimuli were processed by 2047-th order digital FIR LPFs produced by the *fir1* command (version 1.15.4.4) implemented in MATLAB software (The Mathworks, Natick, MA, version 7.4.0.287). Filter cutoff frequencies ranged between 125 Hz and 8000 Hz and were either fixed at the beginning of a trial or varied adaptively as described below. Signal attenuation exceeded 80 dB for frequencies more than 200 Hz from the filter’s specified cutoff frequency. As in Noordhoek *et al.* (1999), the RMS level of the waveform was adjusted to match the RMS level of the unfiltered waveform. (For large reductions in

bandwidth, this operation results in a substantial gain increase.) The filtered syllables were output from a computer’s sound card (SigmaTel Digital Audio, Austin, TX) to a headphone amplifier (Behringer ProXL HA4700, Bothell, WA) driving a pair of Sennheiser HD580 circumaural headphones at 65 dB SPL (flat weighting) loaded by a free-standing flat-plate coupler (Bruel and Kjaer DB0843, Norcross, GA) secured with a coupling force of 2 N. The presentation level was calibrated daily using repeated loops of level-matched speech-spectrum noise, developed by creating a 110250-sample white-noise signal (i.e., five s at a 22050 Hz sampling rate), passing it through a 50th-order all-pole filter matched (via Levinson’s recursion) to the average autocorrelation function for the 23 VCV words, and scaling the filter’s output to the average RMS level of the VCV words. The noise was played by the COOL EDIT PRO software package (Syntrillium Software, Phoenix, AZ) through the signal chain and measured with a Class 1 sound level meter (Quest SoundPro SE/DL, Oconomowoc, WI) prior to the first testing session of each test day.

4. Procedures

Subjects were tested in a double-walled sound-treated booth (IAC 1604, Bronx, NY) in two separate test sessions. Each session lasted two to two-and-a-half h, with breaks provided as needed. The filtered syllables were presented to subjects by custom MATLAB software running on a laptop computer inside the booth. After hearing each syllable, subjects selected the perceived VCV from a list of 23 candidates provided by the software’s visual interface (described in detail by [Whitmal et al. 2007](#)). Before experimental data was recorded, a practice set consisting of 23 unfiltered tokens (one presentation per individual token) was presented to each subject to familiarize the listener with the use of the interface. These data were not included in the results.

The first test session for each subject consisted of adaptive runs using transformed up-down responses ([Levitt, 1971](#)) to target performance levels of 15.9, 29.3, 50.0, 70.7, and 84.1 percent-correct syllable recognition. The filter cutoff frequency for each run was set initially at 1000 Hz and then varied (based on the subject’s responses) in accordance with one of the five response rules (see [Table I](#)). The step sizes for frequency changes were initially one octave for the first two reversals, 1/2-octave for the next two reversals, and 1/4-octave thereafter. It should be noted that the less-common “best-of-three” 50%-correct response rule used here was adopted following pilot trials in which the more common “1-up/1-down” rule produced its first four reversals before closely approaching the 50%-correct level.

Two runs were conducted for each performance level, resulting in ten runs per session (i.e., two runs for each of five target percent correct scores). All five target performance levels were presented once in random order, and then presented again in the same order. This was done so that subjects would not be presented with the same percent correct target for two consecutive runs, and to prevent any particular rule from having a significant advantage in learning effect. Previous studies ([Leek, 2001](#)) have addressed concerns about learning by interleaving adaptive tracks; nonetheless, results (shown below) indicate that the simple ordering used here was effective. 115 tokens (i.e., five repetitions of each VCV) were presented for each run, for a total of 1150 tokens presented per subject session.

In the second session, subjects performed the same task for VCVs filtered with constant LPF cutoff frequencies. The constant cutoff frequencies for each subject were calculated by averaging that subject’s cutoff frequencies for all reversals from the fifth reversal onward for each adaptive rule’s two runs. As in the adaptive runs, two runs were completed for each target level, for a total of 1150 tokens per session. Two of the low pass subjects were unable to return for the second session; hence, data from only five of the subjects are analyzed below.

B. Results

1. Comparison of adaptive and fixed-mode scores

Percent-correct scores were computed for each 115-trial fixed-mode run by dividing the number of correct guesses by the total number of trials. For the adaptive-mode trials, where all runs began with a 1000 Hz cutoff frequency, only the trials from the fifth reversal onward were used. The minimum number of trials used was 86; the maximum number used was 108. This restriction removed bias from initial parts of each run where the cutoff frequency was either too low or too high to approximate the target performance level well. Percent-correct scores for both adaptive-mode and fixed-mode trials are shown in [Fig. 1](#) as open and filled symbols (respectively); dashed lines are also shown depicting each target performance level.

Average percent-correct values and filter cutoff frequencies (computed as described above in the Methods section) are presented in [Table II](#). Inspection of [Fig. 1](#) and [Table II](#) show that the adaptive trials approximate the target levels very well; the largest deviation from any target is only 2.28 percentage points. It should be noted that the percentages for rules 1, 2, 4, and 5 are biased *away* from the target values and *toward* the 50%-level. This type of bias is consistent with

TABLE I. Adaptation strategies for performance targets (after [Levitt, 1971](#)).

Rule	Target level (%-correct)	Increase bandwidth after observing...	Decrease bandwidth after observing...
1	15.9	Four consecutive misses	Three or fewer misses followed by a correct guess
2	29.3	Two consecutive misses	One correct guess or one miss followed by a correct guess
3	50.0	Two misses in two or three consecutive trials	Two correct guesses in two or three consecutive trials
4	70.7	One miss or one correct guess followed by a miss	Two consecutive correct guesses
5	84.1	Three or fewer correct guesses followed by a miss	Four consecutive correct guesses

FIG. 1. (Color online) Recognition scores (in percent) for VCV syllables, low-pass filtered in either adaptive-bandwidth mode (open symbols) or fixed-bandwidth mode (filled symbols) with respect to filter cutoff frequency. Dashed lines indicate target performance levels specified in Table I.

theoretical predictions for reversal averages (Oron, 2007) and has been seen in simulations of adaptive trials using rules 4 and 5 (Kollmeier *et al.*, 1988; Schlauch and Rose, 1990; Saberi and Green, 1996; Garcia-Perez, 1998; Garcia-Perez, 2001). Saberi and Green attributed such biases to inherent imbalances in the transformed up-down rules; however, the bias predicted using their approach for rule 5 are (unlike our data) further away from the median level. To test the significance of these differences, subject percent-correct scores were converted to rationalized arcsine units, or “rau” (Studebaker, 1985) and compared with their corresponding rau-converted target percentage levels in a series of Wilcoxon signed-ranks tests. Test statistics and significance levels (shown in Table II) indicate that only the differences for rules 1 and 5 are significant. Similar comparisons (also shown in Table II) were made for the fixed trials, whose scores are (on average) 3.58 percentage points higher than their adaptive counterparts, more variable, and (in all but one case) further away from the target values. This difference has also been noted in previous adaptive-mode/fixed-mode comparisons (Kollmeier *et al.*, 1988). The largest adaptive/fixed difference of 8.86 percentage points was measured for the 70.7% target, where three of the subjects produced scores of 80%-correct or higher for the fixed-mode trials. These results notwithstanding, there is considerable overlap between scores for adaptive-mode and fixed-mode trials at each target level.

The rau-converted scores were also input to a repeated-measures mixed-model analysis of variance (ANOVA) of intelligibility scores, with subject identity assumed to be a

random factor. Within-subject main factors for the ANOVA included the filtering mode (adaptive or fixed), target percent-correct level, and order (i.e., first or second presentation of a mode/level combination). Among these main factors, only level ($F[4,16] = 478.04$, $p < 0.0001$) was significant at the 5% level. Filtering mode was not significant; the interaction between mode and level was, however ($F[4,60] = 5.03$, $p = 0.0015$), presumably reflecting the 8.86 percentage-point difference observed for the 70.7%-correct level. No other interactions were significant.

2. Test-retest reliability

The lack of significance of presentation order in the previous ANOVA suggests both high test-retest reliability and the absence of a learning effect for the two filtering conditions. These prospects were further explored through paired t-test comparisons and Pearson product-moment correlations between intelligibility scores of first and second runs. Results for adaptive-mode trials showed no significant difference ($t(24) = 0.637$, $p = 0.51$) and high test-retest reliability ($r(23) = 0.986$, $p < 0.0001$). Similar results were observed for fixed-mode trials ($t(24) = -0.753$, $p = 0.49$; $r(23) = 0.984$, $p < 0.0001$).

Comparisons of cutoff frequency values measured on first and second adaptive runs were also conducted. Results showed no significant differences between paired cutoff frequency values ($t(24) = 0.076$, $p = 0.94$) with high test-retest reliability ($r(23) = 0.936$, $p < 0.0001$).

3. Comparative feature analysis

Feature analyses were conducted by determining the percentage of correct voicing, manner, and place-of-articulation consonant classifications produced in the subjects' identification tasks. (Categories are shown in Tables III–V of Whitmal *et al.*, 2007.) Results are shown in Fig. 2. Among the three features, voicing was least sensitive to cutoff-frequency changes; 90%-correct classification was achieved for all cutoff frequencies above 758.7 Hz (i.e., the cutoff frequency for the 29.7%-correct level). Manner decreased gradually from 94.5%-correct to 80%-correct (on average) as cutoff frequency decreased from 2868.4 Hz to 758.7 Hz, and dropped sharply as cutoff frequency decreased to 416.4 Hz. The most common manner error was misidentification of non-strident fricatives and affricates as stops. This error mode decreased as bandwidth was increased and high-frequency spectral cues were restored. Changes in place (the most vulnerable feature) resembled changes in intelligibility. In nearly all conditions,

TABLE II. Statistical measures for adaptive-mode and fixed-mode trials with low-pass filtered consonants in Experiment 1.

Target %	Bandwidth (kHz)		Mean %-correct, adaptive mode				Mean %-correct, fixed mode			
	Mean	SD	Mean	SD	T	p	Mean	SD	T	p
15.9	0.416	0.088	18.18	1.9	25.5	0.006	19.39	4.3	18.5	0.065
29.3	0.759	0.121	30.12	2.1	8.5	0.416	34.78	2.7	27.5	0.002
50.0	1.462	0.227	49.05	3.9	-8.5	0.436	51.57	5.9	8.5	0.416
70.7	2.259	0.404	69.05	1.6	-18.5	0.061	77.91	8.3	21.5	0.027
84.1	2.868	0.468	82.39	1.5	-25.5	0.006	83.04	4.8	-3.5	0.754

T: Wilcoxon signed-ranks statistic comparing measurement to target. p: false-alarm probability for T.

FIG. 2. (Color online) Reception scores for phonetic features of VCV consonants filtered in either adaptive-bandwidth mode (dashed line and open symbols) or fixed-bandwidth mode (solid line and filled symbols) vs filter cutoff frequency. Error bars denote $\pm$ SE.

feature reception accuracy for the fixed-mode trials was slightly higher (2.5 percentage points, on average) than reception accuracy for adaptive-mode trials; the largest differences of (approximately) eight to nine percentage points were observed for manner at 758.7 Hz, and for place at 2259.4 Hz.

Percent-correct scores for all features were converted to *rau* and analyzed in a mixed-model repeated-measures ANOVA; within-subject main factors included filtering mode and target percent-correct level. Results showed no significant differences between filtering modes for voicing ($F[1,4] = 0.01$, $p = 0.92$), manner ($F[1,4] = 5.26$, $p = 0.08$), or place ($F[1,4] = 4.16$, $p = 0.11$). First-order interactions between mode and target level were significant for both manner and place; these presumably reflect the two inter-mode differences described above.

4. Fitting of psychometric functions

The recognition and cutoff-frequency data of Fig. 1 were used to derive psychometric curve estimates for each subject. Initial approaches with conventional logistic and Weibull functions (Wichmann and Hill, 2001a) provided poor fits to the data. Subsequently, the nonparametric local linear approach of Zychaluk and Foster (2009) was adopted. The functions produced by this approach are weighted summations of smooth kernel functions, the arguments of which are locally-fitted linear functions of the stimulus variable. The stimulus range for fitting is determined by a smoothing parameter whose value is selected by a “cross-validation” method to minimize deviance for psychometric curve estimates derived from a subset of the data. Note that deviance, defined as the difference between the log-likelihood values for a given curve estimate and a theoretical curve estimate with no prediction error, is given as

$$D = 2 \sum_{i=1}^M \left[r_i \log \frac{r_i}{n_i P(f_i)} + (n_i - r_i) \log \frac{(n_i - r_i)}{n_i - n_i P(f_i)} \right], \quad (2)$$

where M is the number of performance levels investigated, $P(f_i)$ is the given curve estimate for performance level i at cutoff-frequency f_i , n_i is the number of trials at that performance level, and r_i is the number of correctly identified syllables.

Deviance is asymptotically distributed as $\chi^2(M)$ and is useful as a measure of goodness-of-fit. The reader is referred to Wichmann and Hill (2001a) for a discussion of its application to psychometric functions.

Custom software developed for local linear fits of psychometric functions (Zychaluk and Foster, 2009) was used to fit curves to each subject’s data for both adaptive-mode and fixed-mode runs. To facilitate convergence for all five subjects, the base-2 logarithm of the cutoff-frequency was used as the stimulus variable, and the cross-validation smoothing parameter was set to a minimum value of 0.6. Measured data and fitted curves for each subject are shown in the five panels of Fig. 3 marked with initials; a sixth panel shows the local linear curve fit to the log of the average cutoff-frequency and percent-correct values, with all sets of adaptive-mode data and fitted curves included for comparison. This average curve is also plotted in the other panels for reference. Inspection of Fig. 3 suggests good agreement in all cases between the curves and the measured data. Agreement between fixed-mode and adaptive-mode data is also good. Small but noticeable differences are visible between the fixed- and adaptive-mode curves of subjects JB and NH; the other subjects’ fixed- and adaptive-mode curves overlap substantially.

Goodness-of-fit for each of the curve estimates of Fig. 3 was assessed using the simulation approach of Wichmann and Hill (2001a). These tests utilize simulated distributions of deviance that closely resemble $\chi^2(5)$ distributions. The subjects’ curve estimates were used to simulate the number of correct answers the subject would produce at the individual performance levels in each of 10000 experiments. Deviance values computed for each simulation were compiled into simulated deviance distributions against which the deviance values of the measured performance data could be compared. Deviance values for the subjects’ adaptive-mode runs ranged between 0.35 and 2.61, well within the 95%-confidence interval ($\chi^2(5) \leq 11.07$). Deviance values for the fixed-mode runs were higher, ranging between 0.86 and 3.74 for four subjects and reaching a statistically significant value of 15.41 ($p = 0.009$) for subject NH. This unusually high value is a result of the subject’s unusually high score for rule four trials, which is inconsistent with the other data points and the best-fit curve.

The reliability of the curve estimates of Fig. 3 was assessed using the *pfcmp* procedure of Wichmann and Hill (2001a, 2001b) to see whether the first- and second-run curves for each subject were significantly different. *pfcmp* uses Monte Carlo simulations to construct bivariate normal distributions of inter-curve threshold and slope differences; these are then used to test the null hypothesis that the threshold and slope parameters of the curves are identical. The original *pfcmp* procedure was modified to use the local linear fitting software of Zychaluk and Foster (2009). Hypothesis tests for each subject (using 2000 Monte Carlo trials) produced *p*-values ranging between 0.17 and 0.75, supporting the null hypothesis of identical curve parameters.

5. Convergence properties

Measures of cutoff frequency estimate (CFE) bias and variance were computed to illustrate the adaptive-mode

FIG. 3. (Color online) Local linear psychometric curves fit to percent-correct recognition vs cutoff frequency data for VCV syllables, low-pass filtered in adaptive-bandwidth mode runs for five subjects. Panels marked “Subject” show scores and curves for a subject’s two trials. Run 1 data are denoted by triangles and dash-dotted lines; run 2 data are denoted by squares and dashed lines. The lower right-hand panel plots the average data measured for each target level (filled symbols), along with their best-fit local linear psychometric curve (solid line); individual subject data (open symbols) and fitted curves (dashed lines) are also shown for reference. The best-fit line is also shown in each of the five subjects’ panels.

runs’ convergence properties. Bias for the CFE values are shown in the left half of Fig. 4. There, reversal data for each run were used to compute running CFE values in base-2 logarithm units as a function of trial number. (CFE values for trials prior to the fifth reversal were set to 9.97—the base-2 log of 1000—by default.) CFE values for subject runs at each target level were averaged to produce average CFE functions; the final CFE for that level was then subtracted from the running CFE data to produce the final result. Note that these “relative bias” functions do not account for the biases observed above in subsection 1. A typical bias function may be divided into three regions:

- (1) an early region where none of the runs have achieved four reversals, the CFE is still equal to 9.97, and bias magnitude is maximized;
- (2) a transition region where at least one run has achieved four reversals and bias decreases; and,
- (3) an asymptotic region where all runs have achieved at least four reversals and the remaining residual bias moves slowly towards zero.

The distinction between transition and asymptotic regions is most evident for the 15.9, 70.7, and 84.1%-correct runs where the initial bias magnitude was greater than one octave. The boundary between regions appears to occur at the first trial for which (a) bias is less than or equal in size to the 1/4-octave step and (b) more than one 1/4-octave-step reversal is available for averaging. These boundary values are denoted by vertical dash-dot lines in Fig. 4. The number of reversals completed at the boundary trials (averaged across all five levels) was 6.06.

The standard deviation (SD) for CFE functions for each level are plotted versus trial number in the right half of Fig. 4. Regions of behavior like those of the bias functions are visible, albeit with different boundaries and behavior. The early region (which has constant bias) shows no variation; the tran-

FIG. 4. (Color online) Convergence statistics for cutoff-frequency estimates for VCV syllables, low-pass filtered in adaptive-bandwidth mode trials and presented in order of target level. Left panels: relative cutoff-frequency estimate bias (in octaves) vs trial number. Dashed horizontal lines denote a range of ± 0.25 octaves (one adaptive step) from zero bias. Dashed vertical lines denote the trial boundary between transition and asymptotic regions: numbers to the right of the line denote the average number of reversals recorded at the trial boundary. Right panels: standard deviation (in octaves) of cutoff-frequency estimates vs trial number. Dashed horizontal lines denote deviations of 0.25 and 0.50 octaves (i.e., one or two adaptive steps) from zero. Numbers in the upper-right-hand corner of the panel denote the average number of reversals occurring in each 115-trial run.

sition region shows a transient increase in variation due to inter-track differences in reversal location. The asymptotic region is marked by the gradual approach of the SD to the asymptotic value of 1/4th octave (i.e., one frequency step). For the 15.9, 29.3, and 50%-correct runs, the boundary between the transition and asymptotic regions appears at the trial where the SD equals 1/2 octaves (i.e., two frequency steps). For all five runs, the boundary between transition and asymptotic behavior appears before the 40th trial. The SD panels also display the average number of reversals completed during each run. Reversal counts for the 15.9, 29.3, 70.7, and 84.1%-correct runs were consistent with those observed by García-Pérez (2001) in simulations of k-up/1 down adaptive runs with small step sizes. (Data for the 1-up/1-down 50%-correct rule used in

FIG. 5. (Color online) Recognition scores (in percent) for VCV syllables, high-pass filtered in either adaptive-bandwidth mode (open symbols) or fixed-bandwidth mode (filled symbols) with respect to filter cutoff frequency. Dashed lines indicate target performance levels specified in Table I; arrows denote outliers at the 15.9%-level.

TABLE III. Statistical measures for adaptive-mode and fixed-mode trials with high-pass filtered consonants in Experiment 2.

Target %	Bandwidth (kHz)		Mean %-correct, adaptive mode				Mean %-correct, fixed mode			
	Mean	SD	Mean	SD	T	p	Mean	SD	T	p
15.9	4.889	1.092	20.34	4.1	25.5	0.006	18.78	4.2	15.5	0.131
29.3	3.675	0.689	32.51	2.6	25.5	0.006	40.78	10.2	23.5	0.014
50.0	2.671	0.440	54.46	3.1	26.5	0.004	61.83	8.6	24.5	0.010
70.7	1.906	0.488	73.62	2.3	25.5	0.006	76.78	4.9	26.5	0.004
84.1	1.259	0.434	85.18	1.9	13.5	0.184	87.83	3.6	23.5	0.014

T: Wilcoxon signed-ranks statistic comparing measurement to target. p: false-alarm probability for T.

that paper could not be compared directly with data for the best-of-3 rule used here.)

III. EXPERIMENT 2 – IMPORTANCE MEASUREMENT WITH HIGH-PASS FILTERS

A. Methods

Methods for Experiment 2 were identical to those used in Experiment 1, with two exceptions: filtering, and subject participation. Stimuli for Experiment 2 were processed by 2047-th order digital FIR HPFs, computed using the same MATLAB procedure. Subjects for Experiment 2 included six adults between the ages of 18 and 21 years (mean age: 20.25 years). All subjects met the eligibility requirements of Experiment 1 and were similarly compensated. None of the subjects had participated in Experiment 1. One of the HPF subjects was unable to return for the second session; hence, data from only five of the subjects are analyzed.

B. Results

1. Comparison of adaptive and fixed-mode scores

Percent-correct scores were computed for both adaptive-mode and fixed-mode trials, using the approach employed in Experiment 1. (The number of usable trials for adaptive-mode scores ranged between 73 and 106.) Results are shown in Fig. 5; average percent-correct values and filter cutoff frequencies are presented in Table III. Figure 5 and Table III show that both approaches tend to overshoot the performance target levels. The adaptive-mode scores are (on average) 3.2 percentage points above their target levels, and the fixed-mode scores are (on average) 4.0 percentage points higher than their adaptive counterparts. Variability for both modes was higher than that observed for LPFs. What's more, the overlap seen between adaptive and fixed scores is substantially reduced; for the 29.3% and 50%-correct levels, the fixed scores are (on average) 11.6 percentage points above target, with trials for those levels approaching the next higher target level (see Fig. 6). Wilcoxon signed-ranks tests performed on rau-converted percent-correct scores (see Table III) indicate that the overshoot is statistically significant for four of the five rules.

As for Experiment 1, rau-converted percent-correct scores were input to a repeated-measures mixed-model ANOVA. Among the main factors, only level ($F[4,16] = 292.19$, $p < 0.0001$) was significant at the 5% level. Filtering mode approached (but did *not* achieve) significance ($F[1,4] = 6.61$, $p = 0.062$). The interaction between mode and level was, as in

Experiment 1, significant ($F[4,60] = 4.01$, $p = 0.006$), reflecting the intermode differences observed at the 29.3% and 50%-correct levels. No other interactions were significant.

2. Test-retest reliability

As for Experiment 1, test-retest reliability and learning effects were explored through paired t-test comparisons and Pearson product-moment correlations between intelligibility scores of first and second runs. Results for adaptive-mode trials showed no significant difference ($t(24) = 1.52$, $p = 0.14$) and high test-retest reliability ($r(23) = 0.986$, $p < 0.0001$). Differences approached (but did *not* achieve) significance for fixed-mode trials ($t(24) = -1.95$, $p = 0.063$; $r(23) = 0.988$, $p < 0.0001$). Comparisons of cutoff frequency values measured on first and second adaptive runs showed no significant differences between paired cutoff frequency values in the two runs ($t(24) = 0.563$, $p = 0.58$) and good test-retest reliability ($r(23) = 0.881$, $p < 0.0001$). The latter correlation is greatly affected by two first-run scores at the 15.9%-correct level; these are identified by the arrows in Fig. 5.

3. Comparative feature analysis

Feature analyses were conducted as described for Experiment 1, with results shown in Fig. 2. Each feature's reception decreases in similar fashion as the cutoff frequency increases, with (relatively) gradual decreases for cutoffs below 2670.7 Hz (the 50%-correct cutoff frequency) and steeper decreases for higher cutoffs, where stops and fricatives were increasingly misidentified as either /s/ or /z/ — two phonemes with most of their spectral energy concentrated at frequencies above 3500 Hz. Feature reception accuracy for the fixed-mode trials was again slightly higher (3.2 percentage points, on average) than reception accuracy for adaptive-mode trials. As in Experiment 1, feature scores were converted to rau and analyzed in a mixed-model repeated-measures ANOVA. Results of the ANOVA showed significant differences between filtering modes for voicing ($F[1,4] = 8.58$, $p = 0.043$) and manner ($F[1,4] = 8.58$, $p = 0.046$). This result reflects the higher feature transmission observed for the fixed-mode conditions. No significant differences were observed for place ($F[1,4] = 3.38$, $p = 0.140$). First-order interactions between mode and target level were not significant.

4. Fitting of psychometric functions

Psychometric curves were fit to the data of Experiment 2, using the approach employed for Experiment 1, with one

difference: the minimum cross-validation smoothing parameter was increased to 1.1 to facilitate convergence. (Lower values resulted in numerical errors.) Measured data and fitted curves for each subject are shown in the five initialed panels of Fig. 6; these show good agreement between the first and second curves of subjects DR and KM, small but noticeable differences for subjects DL and LT (who scored below the overall average), and substantial differences for subject EH (who scored above average). Fixed-mode data (shown in each panel) resemble data from both runs for subjects DR and KM, and are most similar to second-run adaptive-mode data for subjects DL, LT, and EH.

Goodness-of-fit and reliability tests were conducted using the methods of Experiment 1. Deviation values for the second-trial curves of subjects EH and KM were 11.10 and 10.05 (respectively), both of which approach (but do not exceed) the distribution's critical value for the 95%-confidence interval. These high deviation values reflect the large prediction errors visible for several of the curves' data points. Deviation values for the remaining adaptive-mode and fixed-mode curves were similar (range: 0.28–5.65). Reliability tests of threshold/slope differences in first-run and second-run curve estimates showed significant differences for subjects EH ($p < 0.001$) and DL ($p = 0.029$); the remaining subjects presented p -values ranging between 0.16 and 0.66.

5. Convergence properties

Relative bias and SD measures were computed for all Experiment 2 runs, using the approach employed for Experiment 1. Results are presented in Fig. 7. A comparison of Figs. 4 and 7 reveals two major differences between the data sets. First, the initial bias values for Experiment 2 runs are all negative and inversely correlated with performance level. In contrast, the data of Fig. 4 shows both positive and negative initial bias values that are never more than 1.5 octaves from 1000 Hz. This difference reflects the fact that all of the measured cutoff frequencies in Experiment 2 are greater than the initial 1000 Hz start frequency, whereas the Experiment 1 cutoff frequencies are situated on either side of 1000 Hz. Second, the asymptotic regions for Experiment 2 runs are characterized by later boundary locations (30.6 trials, on average), more reversals at boundary (6.86, on average), and higher asymptotic SD values for the 70.7% and 84.1%-correct levels. Stable SD values are observed by trial 40 for all runs except the 50%-correct run, which moves suddenly toward the asymptote at trial 43. Runs for the 29.3, 50.0, and 70.7%-correct levels also show 3.3 fewer total reversals (on average) than their Experiment 1 counterparts. These differences reflect the higher variability seen in Fig. 5.

IV. DISCUSSION

A. Comparison of LPF and HPF data

The purpose of Experiments 1 and 2 was to determine whether the proposed adaptive-mode approach for measuring IFs would give valid, reliable results that were consistent with data obtained using a fixed-mode approach. Results of both experiments indicate that they do, albeit to different degrees. LPF scores (Experiment 1) more closely approached target

FIG. 6. (Color online) Local linear psychometric curves for VCV syllables, high-pass filtered in adaptive-bandwidth mode runs for five subjects. Data format is identical to that of Fig. 3.

levels than HPF scores (Experiment 2); the latter were consistently above target levels and showed higher variability. The higher variability resulted in psychometric function estimates with poorer fits to the data and poorer test-retest reliability. Moreover, filtering mode significantly reduced feature reception in Experiment 2, whereas it had no significant effect in Experiment 1. These findings suggest that LPF data better meet the goals of this work than HPF data.

The relationships between feature reception and bandwidth seen in the present study are similar to those of Miller and Nicely (1955), who compiled confusion matrices and feature reception data for 16 consonant-vowel words processed by LPFs and HPFs. Miller and Nicely noted that LPF feature patterns were similar to those produced by masking noise, and observed predictable consonant confusions which they attributed to loss of high-frequency place cues. In contrast, confusions for HPF speech were attributed to the loss of audibility of all features, and were observed to be less predictable and distributed randomly across their confusion matrices. In their words, “(w)hen an error occurs with high-pass filtering, there is little chance of predicting what the

FIG. 7. (Color online) Convergence statistics for cutoff-frequency estimates for VCV syllables, high-pass filtered in adaptive-bandwidth mode trials and presented in order of target level. Data format is identical to that of Fig. 4.

error will be (Miller and Nicely, 1955).” This randomness presumably explains the higher variability of HPF speech intelligibility, where consecutive errors would be more likely to occur despite changes in bandwidth intended to reduce them. It also suggests that the difficulties of measuring HPF speech are not restricted to the present method.

B. Plausibility of derived importance functions

The lower precision seen in the HPF data may raise concerns about the plausibility of importance functions derived from that data. This section explores these concerns by comparing both the data and a derivative IF with those of previous reports to see whether they exhibit properties expected in an importance function. The purpose of these comparisons is to show that the essential properties of the present data are consistent with those of previous IFs (and *not* to show that they are identical to those IFs, which represent different source materials).

The top panel of Fig. 8 compares local linear psychometric functions fitted to the ten average LPF and HPF CFEs measured in Experiments 1 and 2 in both adaptive-mode and fixed-mode measurements. These are compared with the corresponding psychometric function pair of French and Steinberg (1947), measured for consonant-vowel-consonant (CVC) nonsense syllables presented at approximately 75 dB SPL. While the function pairs differ in both asymptotic value and slope (presumably because of differences in stimulus type, level, and filter response), they exhibit very similar LPF/HPF intersection (or “crossover”) frequencies: the French/Steinberg curves intersect at 1930 Hz, whereas the adaptive- and fixed-mode curves intersect at 2080 Hz and 2089 Hz (respectively). The crossover frequency is the frequency above or below which only half of intelligible information is transmitted (i.e., where $AI = 0.5$) and is an intrinsic property of the source material (Hirsh *et al.*, 1954; Studebaker and Sherbecoe, 1993). The value measured here is consistent with crossover frequencies reported for other nonsense syllable tests ranging between 1660–2354 Hz (Beranek, 1947; Fletcher and Galt, 1950; Hirsh *et al.*, 1954; Wang *et al.*, 1978; Studebaker and Sherbecoe, 1993; Henry *et al.*, 1998; Ardoint and Lorenzi, 2010).

A second intrinsic property of interest is band additivity. French and Steinberg (1947) noted that the sum of AI values for complementary LPF/HPF bands should equal the AI value of the unfiltered speech. This property follows from the observation (Fletcher and Galt, 1950) that the error rate for unfiltered speech perception is the product of the error rates for filtered speech from individual bands:

$$e_{BB} = e_{LP}e_{HP}. \quad (3)$$

The AI therefore equals

$$\begin{aligned} AI &= -K \log_{10} e_{BB} = -K \log_{10}(e_{LP}e_{HP}) \\ &= -K \log_{10}(e_{LP}) - K \log_{10}(e_{HP}) = AI_{LP} + AI_{HP}, \end{aligned} \quad (4)$$

where K is a fitting constant determined by the subject test data. (Figure 8 shows that $e_{BB} = 0.329$ at the crossover fre-

quency where $AI = 0.5$; hence, $K = 1.04$.) Equation (3) implies that the product of individual band errors should remain approximately constant as the cutoff frequency is varied; this was recently verified by Li and Allen (2009) for both their own subject data and the data of Miller and Nicely (1955). The bottom panel of Figure 8 shows results of a similar evaluation, made by converting the functions in the top panel of Fig. 8 to error functions and multiplying the complementary error terms to produce estimates of broadband error. Error products for both the present data and the data of French and Steinberg range between 0.89 and 0.99, with the largest deviation produced in the crossover regions of each function pair. This deviation is consistent with the 10% overprediction reported by Li and Allen for the Miller and Nicely data, which they attributed to the use of small numbers of unmatched cutoff frequencies in the LPFs and HPFs, and suggests that band additivity holds to the extent that the measured cutoff frequencies can reveal.

The plausibility of IFs derived from the present data can also be assessed, using the method of Fletcher and Galt (1950). For speech at the optimum level input to an ideal LPF with cutoff frequency f_{LP} , Eqs. (1) and (4) state that:

$$AI_{LP} = -K \log_{10}(1 - s(f_{LP})) = \int_{f=0}^{f_{LP}} I(f) df. \quad (5)$$

where $s(f_{LP})$ is the proportion of correct answers at f_{LP} . The AI vs frequency relationships of Eqs. (4) and (5) were modeled for both adaptive and fixed-mode trials by computing AI values for the ten data points shown in the top panel of Fig. 8 and fitting a local linear psychometric curve to them (Zychaluk and Foster, 2009). $I(f)$ values were then computed from the derivative of the fitted curve and plotted in the upper panel of Fig. 9, using the AI-per-Hz units of Fletcher and Galt on the ordinate. Alternate versions of $I(f)$ are shown in the lower panel of Fig. 9, with the computation changed to produce units of one-third-octave band frequency on the abscissa and band-importance percentage on the ordinate. This change allows visual comparisons with the one-third-octave band version of the French-Steinberg IF reported by Studebaker and Sherbecoe (1993). The adaptive-mode and fixed-mode IFs (which are virtually identical) each show strong peaks near 2000 Hz, like the French-Steinberg IF and other IFs measured for nonsense syllables and monosyllabic words (Wang *et al.*, 1978; Studebaker and Sherbecoe, 1993; Henry *et al.*, 1998; Ardoint and Lorenzi, 2010). The concentration of importance in the new IFs is noticeably stronger near 2000 Hz and weaker at lower frequencies than the French-Steinberg IF. One possible explanation for this difference is the present use of VCV tokens. Wang *et al.* (1978) showed that crossover frequencies for place features were higher for vowel-consonant syllables than for consonant-vowel syllables that used the same 16 stops, fricatives and affricates. Since Figs. 2 and 6 indicate that intelligibility is more strongly correlated with place reception than manner or voicing reception, it seems likely that an upward shift in

FIG. 8. (Color online) Top panel: average recognition scores (in percent-correct) for low-pass and high-pass filtered VCV syllables obtained in adaptive-mode (solid lines) and fixed-mode (dashed lines) trials, compared with average recognition scores (in percent-correct) for the filtered CVC syllables of French and Steinberg, 1947 (dash-dotted lines). The “crossover” frequency for VCV syllables is denoted by the dashed vertical line. Bottom panel: Recognition scores from broadband error estimates vs cutoff frequency for VCV syllables, obtained in adaptive-mode (solid lines) and fixed-mode (dashed lines) trials, compared with estimates for the French and Steinberg data.

place crossover frequency would be associated with an upward shift in importance density.

C. Practical implications

The results of Experiments 1 and 2 suggest that reliable CFEs for each performance level with either LPF or HPF speech may be obtained within 40 trials. For software used in this experiment, average trial times are 5.92 s for LPF speech and 6.10 s for HPF speech. Running a block of 40 trials would therefore require about four min; running ten blocks to obtain the data of Experiments 1 and 2 would require about 40 min per subject. This estimate is admittedly very conservative, as these measurements were the first ones made for this application and the bandwidths and convergence rates were

FIG. 9. (Color online) Top panel: Importance function for VCV syllables in AI units/Hz as derived from Eq. (5) and data from adaptive-mode (solid lines) and fixed-mode (dashed lines) trials. Bottom panel: Importance functions for VCV syllables from top panel, recomputed in units of %-contribution per one-third-octave band and compared with the one-third-octave-band importance function of French and Steinberg 1947 (dash-dotted lines) as reported by Pavlovic (1994).

FIG. 10. (Color online) Convergence statistics for adaptive-mode cutoff-frequency estimates using starting bandwidths of either 1000 Hz (dashed line) or rule-specific CFE values from Experiments 1 and 2 (solid line). Low-pass filter results are shown on the left side; high-pass filter results are shown on the right. Computation of relative bias is similar to that of Figs. 4 and 8 with only the first two reversals discarded.

unknown. Estimating the time needed to measure complete importance functions in practical applications is more complex, and depends on three factors described below.

1. Selection of initial bandwidths

Figures 4 and 7 indicate that CFE values for the early regions of adaptive runs are often biased strongly toward their initial value (1000 Hz). The 1000 Hz initial bandwidth was chosen arbitrarily, as the bandwidths required to produce the performance levels of Table I were unknown. Initial biases were also observed in adaptive trial simulations (Garcia-Perez, 1998; Garcia-Perez, 2001) that resembled the present experiment in two aspects: step size was small compared to the psychometric function’s “spread” (i.e., stimulus range for non-asymptotic behavior), and the difference between the initial value and true threshold was equal to or greater than one-third of the spread. For these simulations, initial bias was reduced (and convergence was accelerated) by placing the starting value close to the true threshold. It’s possible that convergence could also be accelerated in the proposed method if the 1000 Hz bandwidth was replaced in each rule by the measured bandwidths shown in Tables II and III.

To test this hypothesis, adaptive trials were simulated using two sets of initial bandwidths: the 1000 Hz bandwidths of Experiments 1 and 2, and the average CFE values of Tables II and III. For each simulated trial, the average psychometric curve of either Fig. 3 or 6 was used to derive a recognition probability for use in a Bernoulli trial to simulate a subject response. This response was used (as in Experiments 1 and 2) to determine the CFE for the next simulated trial in accordance with the decision rules of Table I. One thousand simulations were run for each of the ten rule/filter combinations and used to compute relative bias measures for the two initial bandwidth sets. Computations were similar to those described by Figs. 4 and 7, with one exception: CFE values were computed after only two reversals (Garcia-Perez, 2001) to allow better inspection of early bias. Results (shown in Fig. 10) indicate that

starting at the average CFE values produces considerably lower initial bias than starting at 1000 Hz. Use of similar starting values in clinical settings could provide fast convergence to shorten testing of subjects with normal hearing, given appropriate stopping rules, using trial counts that would result in high-variance estimates using fixed-mode testing.

2. Sampling of psychometric curve

The IFs displayed in Fig. 9 were derived from curves fitted to ten data points, selected because they conformed to the transformed up-down rules of Levitt (1971) and also provided a fine sampling of relevant cutoff-frequency values. Fine sampling is crucial for establishing the viability of the method, since poor sampling schemes can impose bias on psychometric curve estimates (Wichmann and Hill, 2001a). At the same time, the large number of samples also increases the time required to measure the IF. This increase may be unnecessary, since the HPF and LPF data appear to characterize the same underlying curve and contain redundant information in the crossover region. Fletcher and Galt (1950) exploited this redundancy by using HPF data for the low-frequency portion of their IF and LPF data for the high-frequency portion. Developing a similar approach for the proposed method would reduce the number of trials by about half. Other savings may be obtained by using computer simulations to optimize the number and frequency values of samples (Lam *et al.*, 1999), with non-uniform step sizes (Kaernbach, 1991) used to converge on performance levels other than those specified by Levitt's rules.

3. Use of curve bisection method

The IF displayed in Fig. 9 was derived from the derivative of a single psychometric curve in accordance with Eq. (5). Use of Eq. (5) implicitly assumes that phoneme recognition follows the exponential transfer function derived by Fletcher and Galt (1950). Other IF measurements have used the curve-bisection approach of French and Steinberg (1947), in which crossover frequencies for multiple LPF/HPF curve pairs at varied intensities and/or SNRs are used to define the IF and transfer function. The need for multiple curves would increase the number of trials required to measure the IF. Although Fletcher and Galt claimed that their method and the curve-bisection method provided similar results for CVC syllables, it's not clear that differences between the methods have ever been rigorously evaluated. It is also likely that source materials with more linguistic redundancy will follow different transfer functions (Studebaker *et al.*, 1987). In such cases, the proposed method could be extended to measure the additional curve pairs needed for the curve bisection method at specified performance levels through adaptive variation of intensities and/or SNRs. The ability to specifically target performance levels would also eliminate the need for extensive pilot testing and/or unplanned changes to test conditions (Wong *et al.*, 2007). Further investigation is needed to determine the relationship between linguistic content and testing time.

D. Advantages of adaptive-mode measurement

The similarities between adaptive-mode and fixed-mode results for normal hearing patients confirm the validity of adaptive-mode measurements. When weighed against the large number of trials used in this report, however, the motivation for using an adaptive approach may be unclear. Certainly, for experiments where stimuli are restricted to a range bracketing the threshold, fixed-mode approaches can provide consistent, accurate threshold estimates that are as good or better than adaptive-mode estimates (Simpson, 1988). Such restrictions require pilot testing (Watson and Fitzhugh, 1990), which may only be feasible in some research conditions. For other conditions where the appropriate ranges are *unknown a priori*, the flexibility of the adaptive-mode approach is preferred. This is exemplified in the present study, where adaptive-mode measurements were used as pilot tests to determine the (unknown) bandwidths required for fixed-mode measurements. While the bandwidths measured here could be used as preliminary norms for subjects with normal hearing (a uniform group appropriate for testing validity and reliability), it's unlikely that they would be appropriate for individual hearing-impaired patients. (As an example, consider the futility of repeatedly using fixed-mode 4889 Hz high-pass filtered speech to test patients with extensive high frequency losses.) Adaptive-mode testing eliminates the need for preliminary pilot testing to determine appropriate bandwidths for each patient and (as shown in Fig. 10) can converge rapidly on those bandwidths in a short number of trials when appropriate starting frequencies and stopping rules are applied. Determining these rules is a topic for future research.

V. SUMMARY AND CONCLUSION

The AI and SII prediction algorithms rely on importance functions, which vary widely among source materials and can be time-consuming and complicated to measure. Previous research suggests that the accuracy of AI and SII predictions might be improved if individually-measured importance functions were available. The present work evaluated the feasibility of an adaptive method of measuring importance functions for individual subjects. Results indicate that the proposed method produces reliable results that are consistent with importance functions measured using a constant-stimuli method.

ACKNOWLEDGMENTS

The authors would like to thank the two anonymous reviewers for their feedback, which improved the paper considerably. The authors would also like to thank Dr. Richard Freyman, Dr. Karen Helfer, and Dr. Sarah Poissant for their support and helpful suggestions with previous drafts of this paper. Funding for this research was provided by the National Institutes of Health (NIDCD Grant No. R03 DC7969).

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

- ANSI (1997). ANSI S3.5-1997, *American National Standard Methods for the Calculation of the Speech Intelligibility Index* (American National Standards Institute, New York).
- Ardoit, M., and Lorenzi, C. (2010). "Effects of lowpass and highpass filtering on the intelligibility of speech based on temporal fine structure or envelope cues," *Hear. Res.* **260**, 89–95.
- Beranek, L. (1947). "The design of speech communication systems," *Proc. IRE*, 880–890.
- Ching, T. Y. C., Dillon, H., and Byrne, D. (1998). "Speech recognition of hearing-impaired listeners: predictions from audibility and the limited role of high-frequency amplification," *J. Acoust. Soc. Am.* **103**, 1128–1140.
- Eisenberg, L. S., Dirks, D. D., Takayanagi, S., and Martinez, A. S. (1998). "Subjective judgments of clarity and intelligibility for filtered stimuli with equivalent Speech Intelligibility Index predictions," *J. Speech Lang. Hear. Res.* **41**, 327–339.
- Fletcher, H., and Galt, R. (1950). "The perception of speech and its relation to telephony," *J. Acoust. Soc. Am.* **22**, 89–151.
- French, N. R., and Steinberg, J. C. (1947). "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.* **19**, 90–119.
- Garcia-Perez, M. A. (2001). "Forced-choice staircases with fixed step sizes: asymptotic and small-sample properties," *Vision Res.* **38**, 1861–1881.
- Garcia-Perez, M. A. (2001). "Yes-no staircases with fixed step sizes: psychometric properties and optimal setup," *Optom. Vision Sci.* **78**, 56–64.
- Hall, J. W., Buss, E., and Grose, J. H. (2008). "Spectral integration of speech bands in normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.* **124**, 1105–1115.
- Henry, B. A., McDermott, H. J., McKay, C. M., James, C. J., and Clark, G. M. (1998). "A frequency importance function for a new monosyllabic word test," *Austr. J. Audiol.* **20**, 79–86.
- Hirsh, I. J., Reynolds, E. G., and Joseph, M. (1954). "Intelligibility of different speech materials," *J. Acoust. Soc. Am.* **26**, 530–538.
- Kaernbach, C. (1991). "Simple adaptive testing with the weighted up-down method," *Percept. Psychophys.* **49**, 227–229.
- Kamm, C. A., Dirks, D. D., and Bell, T. S. (1985). "Speech recognition and the Articulation Index for normal and hearing-impaired listeners," *J. Acoust. Soc. Am.* **77**, 281–288.
- Kollmeier, B., Gilkey, R. H., and Sieben, U. K. (1988). "Adaptive staircase techniques in psychoacoustics: a comparison of human data and a mathematical model," *J. Acoust. Soc. Am.* **83**, 1852–1862.
- Kryter, K. (1962). "Methods for the calculation and use of the articulation index," *J. Acoust. Soc. Am.* **34**, 1689–1697.
- Lam, C. F., Dubno, J. R., and Mills, J. H. (1999). "Determination of optimal data placement for psychometric function estimation: a computer simulation," *J. Acoust. Soc. Am.* **106**, 1969–1976.
- Leek, M. R. (2001). "Adaptive procedures in psychophysical research," *Percept. Psychophys.* **63**, 1279–1292.
- Levitt, H. (1971). "Transformed up-down methods in psychoacoustics," *J. Acoust. Soc. Am.* **49**, 467–477.
- Li, F., and Allen, J. B. (2009). "Multiband product rule and consonant identification," *J. Acoust. Soc. Am.* **126**, 347–353.
- Ma, J., Hu, Y., and Loizou, P. (2009). "Objective measures of speech intelligibility in noisy conditions based on new band-importance functions," *J. Acoust. Soc. Am.* **125**, 3387–3405.
- Miller, G., and Nicely, P. (1955). "An analysis of perceptual confusions among some English consonants," *J. Acoust. Soc. Am.* **27**, 338–352.
- Nilsson, M., Soli, S. D., and Sullivan, J. A. (1994). "Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise," *J. Acoust. Soc. Am.* **95**, 1085–1099.
- Noordhoek, I. M., Houtgast, T., and Festen, J. M. (1999). "Measuring the threshold for speech reception by adaptive variation of the signal bandwidth. I. Normal-hearing listeners," *J. Acoust. Soc. Am.* **105**, 2895–2902.
- Noordhoek, I. M., Houtgast, T., and Festen, J. M. (2000). "Measuring the threshold for speech reception by adaptive variation of the signal bandwidth. II. Hearing-listeners," *J. Acoust. Soc. Am.* **107**, 1685–1696.
- Noordhoek, I. M., Houtgast, T., and Festen, J. M. (2001). "Relations between intelligibility of narrow-band speech and auditory functions, both in the 1-kHz frequency region," *J. Acoust. Soc. Am.* **109**, 1197–1212.
- Oron, A. P. (2007). "Up-and-down and the percentile-finding problem," Ph.D. thesis, University of Washington, Seattle, WA.
- Pavlovic, C. V. (1984). "Use of the articulation index for assessing residual auditory function in listeners with sensorineural hearing impairment," *J. Acoust. Soc. Am.* **75**, 1253–1258.
- Pavlovic, C. V. (1994). "Band importance functions for audiological applications," *Ear Hear.* **15**, 100–104.
- Pavlovic, C. V., Studebaker, G. A., and Sherbecoe, R. L. (1986). "An articulation index based procedure for predicting the speech recognition performance of hearing-impaired individuals," *J. Acoust. Soc. Am.* **80**, 50–57.
- Plomp, R., and Mimpen, A. M. (1979). "Improving the reliability of testing the Speech Reception Threshold for sentences," *Audiology* **18**, 43–52.
- Rhebergen, K. S., and Versfeld, N. J. (2005). "A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *J. Acoust. Soc. Am.* **117**, 2181–2192.
- Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (2006). "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise," *J. Acoust. Soc. Am.* **120**, 3988–3997.
- Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (2008). "Prediction of the intelligibility for speech in real-life background noises for subjects with normal hearing," *Ear Hear.* **29**, 169–175.
- Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (2010). "Modelling the speech reception threshold in non-stationary noise in hearing-impaired listeners as a function of level," *Int. J. Audiol.* **49**, 856–865.
- Saberi, K., and Green, D. M. (1996). "Adaptive psychophysical procedures and imbalance in the psychometric function," *J. Acoust. Soc. Am.* **100**, 528–536.
- Schlauch, R. S., and Rose, R. M. (1990). "Two-, three-, and four-interval forced-choice staircase procedures: estimator bias and efficiency," *J. Acoust. Soc. Am.* **88**, 732–740.
- Simpson, W. A. (1988). "The method of constant stimuli is efficient," *Percept. Psychophys.* **44**, 433–436.
- Studebaker, G. A. (1985). "A 'rationalized' arcsine transform," *J. Speech Hear. Res.* **28**, 455–462.
- Studebaker, G. A. and Sherbecoe, R. L. (1993). "Frequency-importance functions for speech recognition," in *Acoustical Factors Affecting Hearing Aid Performance*, 2nd ed., edited by G. A. Studebaker and I. Hochberg (Allyn and Bacon, Boston), Chap. 11, pp. 185–204.
- Treutwein, B. (1995). "Adaptive psychophysical procedures," *Vision Res.* **35**, 2503–2522.
- van Schijndel, N. H., Houtgast, T., and Festen, J. M. (2001). "The effect of intensity perturbations on speech intelligibility for normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.* **109**, 2202–2210.
- van Schijndel, N. H., Houtgast, T., and Festen, J. M. (2001). "Effect of degradation of intensity, time, and frequency content on speech intelligibility for normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.* **110**, 529–542.
- Wang, M. D., Reed, C. M., and Bilger, R. C. (1978). "A comparison of the effects of filtering and sensorineural hearing loss on patterns of consonant confusions," *J. Speech Hear. Res.* **21**, 5–36.
- Watson, A. B., and Fitzhugh, A. (1990). "The method of constant stimuli is inefficient," *Percept. Psychophys.* **47**, 87–91.
- Whitmal, N. A., Poissant, S. F., Freyman, R. L., and Helfer, K. S. (2007). "Speech intelligibility in cochlear implant simulations: effects of carrier type, interfering noise, and subject experience," *J. Acoust. Soc. Am.* **122**, 2376–2388.
- Wichmann, F. A., and Hill, N. J. (2001). "The psychometric function: I. Fitting, sampling, and goodness of fit," *Percept. Psychoph.* **63**, 1293–1313.
- Wichmann, F. A., and Hill, N. J. (2001). "The psychometric function: II. Bootstrap-based confidence intervals and sampling," *Percept. Psychoph.* **63**, 1314–1329.
- Zychaluk, K., and Foster, D. H. (2009). "Model-free estimation of the psychometric function," *Percept. Psychophys.* **71**, 1414–1425.