Published in final edited form as:

J Am Acad Audiol. 2014 September ; 25(8): 737–745. doi:10.3766/jaaa.25.8.4.

Recording and Evaluation of an American Dialect Version of the Four Alternative Auditory Feature Test

Jingjing Xu and Robyn M. Cox

University of Memphis, Memphis, TN, USA

Abstract

Background—The Four Alternative Auditory Feature test (FAAF) is a word-based closed-set speech recognition test. Because the original test materials were recorded in British English dialect, it is not appropriate for use in the United States.

Purpose—To produce an American dialect FAAF (AFAAF).

Research Design—The AFAAF materials spoken by a native American-English speaking male were recorded and digitally edited. In the validation study, the AFAAF was administered monaurally at five SNRs in both ears for each listener.

Study Sample—Twenty young adults with normal hearing participated in the validation study.

Data Collection and Analysis—For each participant, speech recognition scores were collected in one session. The speech level was fixed at 70 dB SPL and the steady-state talker-matched noise level was varied, resulting in five SNRs from –15 to –5 dB. One full list (80 words) was used for each SNR. For each subject, a performance-intensity (PI) function was fit to the discrete mean percent correct scores for the five SNRs according to a best-fit, three-parameter sigmoid function. In addition, scores for the left and right ears were compared to examine test-retest reliability.

Results—Results show that the slope of the PI function is 6%/dB; the mean test-retest difference scores for the five SNRs are within 3 rationalized arcsine units (rau); and the 95% critical difference for the 80-word scores is 12 rau. Compared to the FAAF, the slope of the PI function for the AFAAF is slightly less steep. Test-retest reliability of the AFAAF is at least equal to that of the FAAF.

Conclusions—It is concluded that the AFAAF is similar but not identical to the FAAF. The AFAAF is now available for measuring speech recognition performance in listeners who use American English as a native language.

Keywords

Speech intelligibility; Speech perception; Dialect

Corresponding Author: Jingjing Xu, University of Memphis School of Communication Sciences & Disorders, 807 Jefferson Ave, Memphis, TN 38105, Phone: 901-678-5800, FAX: 901-525-1282, jxu3@memphis.edu.

A part of the results of this project was presented at the Annual Meeting of the American Auditory Society, Scottsdale, Arizona, U.S.A. (March 2010).

Introduction

Closed-set tests have been shown to be useful for evaluating speech recognition performance (Gelfand, 2009). In the United States, a number of word-based closed-set tests have been developed for use in both audiology clinics and research laboratories. Some of the most well-known such tests are the Modified Rhyme Test (MRT; Fairbanks, 1958; House, Williams, Hecker, & Kryter, 1965; Kreul et al., 1968), the California Consonant Test (CCT; Owens & Schubert, 1977), the University of Oklahoma Closed Response Speech Test (UOCRST; Pederson & Studebaker, 1972), and the Speech Pattern Contrast (SPAC) Test (Boothroyd, 1984, 1988). Another closed-set test, which has been in wide use in the United Kingdom for research in hearing science and audiology since 1987, is the Four Alternative Auditory Feature (FAAF) test. This test was originally developed at the MRC-Institute of Hearing Research, United Kingdom. Since its development, the FAAF has been used in numerous studies in the UK and has been shown to be an effective test for assessment of speech recognition performance in hearing research. In this paper, we introduce an American dialect version of the FAAF. We believe that an American dialect version of this test will be a useful addition to the battery of closed-set speech recognition tests available for use in the United States.

The development and evaluation of the original version of the FAAF is thoroughly described in Foster and Haggard (1987). The FAAF is a closed-set monosyllabic word recognition test designed to evaluate a listener's speech recognition performance. As an objective measurement method, the FAAF has been used for different purposes, for example, (1) evaluating aided and unaided speech recognition performance in different background noises and reverberation (e.g., Shields & Campbell, 2001); (2) comparing the benefits of different hearing aids (e.g., Davies, John, & Jones, 1990); (3) comparing the different settings of a hearing aid feature (e.g., Gatehouse, Naylor, & Elberling, 2003); and (4) evaluating the effect of frequency bandwidth of speech on cochlear implant users (e.g., Milchard & Cullington, 2004).

The FAAF has a number of advantages. First, its learning effect is minimal. The FAAF is a monosyllabic word test with 80 key words accompanied by a consistent carrier phrase. The fairly large pool of key words and a contextually unrelated carrier phrase make the listener unlikely to learn the test words. As a result, the same words in different orders can be used repeatedly for testing the same listener in different test conditions.

A second advantage noted with the FAAF is its closed-set testing structure. Compared to open-set tests, closed-set tests can reduce the effects of word frequency in the language (Gelfand, 2009). Moreover, results from a closed-set test are easier to score and such a test format allows incorporation into a computer software program for controlling the test procedure. The details of the advantages of using closed-set speech recognition tests have been thoroughly addressed by Boothroyd (1985).

Last, the FAAF is sensitive to consonant recognition. The key words used in the FAAF were carefully selected with regard to consonants, including most of the commonly used plosives, fricatives, affricates, and nasals. Each key word and its corresponding alternatives have the

same vowel but differ on either the initial or the final consonants. The FAAF can produce scores not only for whole-word recognition performance, but also for both initial and final consonant recognition performance.

There are two major processing strategies that are heavily involved in processing speech information: bottom-up and top-down processing. With bottom-up processing, speech understanding is carried out according to the analytical results of information that has entered into the peripheral auditory systems (Nakagawa, Shikano, & Tohkura, 1995). Top-down processing is based on higher level information, such as previous knowledge and expectations. Such higher level information can somewhat compensate for missing or unclear speech segments and facilitate speech understanding. Because the key words and the carrier phrase of the FAAF are not contextually related, speech understanding depends primarily on audibility. Thus, bottom-up processing presumably is engaged. In addition, as a closed-set test, the displayed alternatives provide some predictive clues (e.g., phonological cues) and assist in identifying the key word. From this perspective, top-down processing is involved in the FAAF to some extent.

In summary, the FAAF is a closed-set word recognition test with minimal learning effect. It involves mostly bottom-up but some top-down processing. Research has shown that the FAAF is an effective test for evaluating speech recognition performance in various test conditions.

Because the speech material of the original FAAF test is recorded in British English dialect, this test is most suited for listeners who are native speakers of British English. Evidence suggests that the presentation of dialectally unfamiliar words may negatively affect speech recognition performance (Weisleder & Hodgson, 1989; Wilson & Moodley, 2000). Considering the difference between British dialect and the American dialect, the original FAAF is of questionable value for listeners who are native speakers of American English. Therefore, its application has been constrained in the United States. In order to use the FAAF in hearing research in the United States, a new set of FAAF speech materials recorded in American dialect, is needed.

The purpose of this project was to produce an American dialect FAAF and evaluate its validity and reliability. The resulting test was dubbed the American Four Alternative Auditory Feature test (AFAAF). Henceforth, the FAAF will refer to the original FAAF recorded in British dialect. For this project, the test materials were recorded by a talker using general American English. Then, a validation study was conducted in listeners with normal hearing to evaluate the validity and reliability of the AFAAF. Because the AFAAF in this project and the FAAF (Foster & Haggard, 1987) were administered on the same type of listeners in very similar test conditions, speech recognition scores and performance-intensity (PI) functions obtained from the two recordings were compared. We hypothesized that the AFAAF has psychometric characteristics similar to those of the FAAF.

Production of the AFAAF

The test comprises 80 key words and five practice words. The 80 key words are partitioned into 20 sets of four, according to the characteristics of the initial consonants or the final

consonants. Taken together, the 80 words contain plosives, fricatives, affricates, and nasals. In each test utterance, the key word is embedded in a carrier sentence: "Can you hear ____ clearly?" The test subject listens to the presentation and selects the word he or she hears from four displayed alternatives that are very similar in pronunciation. For example, a test utterance is: "Can you hear BAIL clearly?" and the four alternatives displayed are: MAIL, BAIL, NAIL, and DALE. The test can be administered in quiet or in the presence of masking noises. The scoring can produce percentage correct scores for initial consonants, final consonants, or whole words.

Permission was obtained from the FAAF developers in the United Kingdom to produce the AFAAF. The two major tasks in the present project included: (1) recording speech samples for the AFAAF, and (2) evaluating the validity and reliability of the scores. All procedures were carried out in the Hearing Aid Research Laboratory (HARL) at the University of Memphis. Recordings were made in a double-walled sound room (Width×Length×Height: 1.9 meters × 1.83 meters × 1.98 meters). Ambient noise level in the sound room was 40 dB SPL Leq.

Talker

The FAAF was spoken by a male talker in a clear speaking style. To promote equivalence, a native American-English speaking male without a regional accent and able to produce speech clearly was chosen for the AFAAF. This talker was 65 years old with a background in audiology. He grew up in the Northeast and Midwest regions of the United States and had been living in the Mid-South region for 40 years. The talker was instructed to speak clearly. However, differences between his clear and conversational speech production were not measured in this study.

All the FAAF word sets that rhyme in British English dialect also rhyme in American English dialect except one. This set of four words (FOUGHT, PORT, TAUGHT, and THOUGHT) is a rhyming set in British English. However, it is no longer a rhyming set in American English, which possibly makes the words in this set distinguishable more from each other in the AFAAF than in the FAAF.

Recording procedures

The recording setup was similar to that used in recording the FAAF. The talker was seated in the center of the sound room, facing a computer monitor (about one meter away). The key words were displayed on the computer monitor one at a time. The talker read each word including the carrier sentence "Can you hear ____ clearly?" The key-word display was controlled by the investigator who was monitoring the recording process outside the sound room. This method has the advantage of avoiding potential sound reflections from a written text, held by the talker. In addition, this method prevents the talker from knowing the number of key words remaining. The talker was instructed to speak each utterance clearly but without emphasizing the key word. Key words were naturally produced in the carrier phrase. Also, all utterances were spoken with normal effort at a constant voice level. A digital recorder (Marantz PMD660) with an external Marantz condenser microphone was used to record speech produced by the talker. The microphone was held on a microphone

stand and oriented towards the talker's mouth. The recording format was mono wav audio file with a sampling rate of 44100Hz and a sampling resolution of 16 bits. The recording level was adjusted so that the talker's voice record level was as high as possible without peak clipping. All sounds, including speech, silent gaps, throat clearing etc., were recorded in one wav file.

Signal editing

Speech—The utterances (both the carrier phrase and key words) including the 80 test words and 5 practice words were extracted from the recorded audio file via a digital audio editor: Adobe Audition, version 1.0. The duration of each utterance was about 2 seconds. Digital RMS levels of the wav files of these utterances were measured. The mean absolute deviation (MAD) of the 85 RMS values was 0.67 dB, which was similar to that of the FAFAF wav files (MAD=0.63 dB). The 85 utterances were concatenated without any silent gaps and saved into one wav file. The spectrum parameters of the AFAAF speech material were derived from this wav file using a signal analysis software program (SpectraPLUS 5.0, Pioneer Hill Software LLC, Poulsbo, WA. USA). These spectrum parameters were used later for shaping masking noises. Each of the 85 speech utterances was bandpass-filtered from 70Hz to 10000Hz. Finally, the total root-mean-square (RMS) level of each utterance (including both the carrier phrase and key word) was digitally adjusted so that they were equal. The amount of adjustment was between 0.01 and 2.34 dB with the mean value 0.67 dB.

Masking noises—One unmodulated and two modulated masking noises were created. Figure 1 depicts the long-term average spectra of the speech and the unmodulated noise. The spectrum of the AFAAF speech was used to shape: (1) a steady-state random noise, (2) a two-talker babble modulated noise from the International Collegium of Rehabilitative Audiology (ICRA) compact disc track 6 (Dreschler, Verschuure, Ludvigsen, & Westermann, 2001), and (3) a six-talker babble modulated noise from the ICRA compact disc track 7 (Dreschler et al., 2001). Then, the three noises were bandpass-filtered from 70Hz to 10000Hz. Each noise was 15 seconds long. The RMS level of each masking noise was digitally adjusted to the same level as the speech materials.

The purpose of creating talker-matched unmodulated and modulated masking noises was to suit different research needs. In the validation study reported here, only the unmodulated masking noise was used.

Calibration signals—A 1000Hz pure tone and a 1000Hz octave-band calibration noise were created. Their durations were both 15 seconds. As with the masking noises, RMS levels of the calibration signals were also digitally adjusted to the same level as the speech materials.

An audio file with an example from the test in the presence of the speech-shaped random noise at +5 dB SNR is attached as supplementary material.

Validation study

Following the completion of the test stimuli editing for the AFAAF, a study was carried out to validate the newly recorded speech materials. Results were compared to those reported by Foster and Haggard (1987) for the FAAF.

In Foster and Haggard (1987), validity and reliability of the FAAF were evaluated in 16 listeners with normal hearing. Speech recognition performance in the presence of a steady-state talker-matched noise was measured monaurally in each ear under eight SNRs from -12.5 to 5 dB with a step size of 2.5 dB.

Subjects

Twenty young adults with normal hearing volunteered to participate in this experiment. The majority of the subjects were audiology and speech-language pathology graduate students at the University of Memphis. There were five males and 15 females, ranging in age from 21 to 33 years (mean = 25.7 , $SD=3.7$). Inclusion criteria were: air conduction pure-tone thresholds better than 25 dB HL bilaterally for octave frequencies from 250 to 8000 Hz; a type "A" tympanogram with the compliance above 0.3 cc; no history of congenital or fluctuating hearing loss, no chronic outer or middle ear pathological conditions, no history of ear surgery; and American English as the native language. The subjects did not receive any payment for their participation.

Procedures

The subject was seated in a single-wall sound room facing a computer monitor that displayed the response alternatives. Each ear was tested separately with stimuli presented using an Etymotic Research ER-2 insert earphone. This type of earphone provides a flat frequency response up to approximately 10000 Hz when measured in a Zwislocki-type ear simulator. The speech recognition test was administered using the test administration software designed for the original FAAF. The steady-state speech-matched noise was used as the masking signal, which was gated on and off with each speech utterance. A 100 -ms ramp was applied to the rising and falling portion of the masking noise. For each utterance, speech and noise were digitally merged within the administration software. Five SNRs, from -15 to -5 dB with a step size of 2.5 dB, were created by varying the masking noise level for a constant speech level of 70 dB SPL RMS. The subject was instructed to listen to the test signals carefully and use a computer mouse to choose the word that he or she heard from the displayed alternatives. Calibration of the playback system for this experiment was achieved by delivering the steady-state speech-spectrum noise from the ER-2 earphone to a Zwislocki-type ear simulator that was connected to a sound level meter (Larson Davis 800B).

The FAAF has five pre-determined randomizations for the 80 key words. In the present study, these five randomizations were used for testing each ear, one for each SNR. In order to familiarize the subject with the upcoming SNR condition, a practice set containing the five practice words was administered before administering each 80-word test list. The SNRs and the word randomizations were counterbalanced. Half of the subjects were tested on their

right ears first and then on their left ears. The remaining subjects were tested in the opposite order. The performance measure used in this experiment was the percentage of the key words that were correctly identified.

This study was approved by the Institutional Review Board of the University of Memphis.

Results

After completing the AFAAF test, each subject had five scores for each ear, one for each SNR. The score range is from 0% to 100%. However, because the AFAAF is a four alternative force choice test, random selection of one word out of the four alternatives theoretically results in a chance performance of 25%. Figure 2 depicts a box-and-whisker plot that displays the speech recognition scores for each ear at each SNR. This figure illustrates the decrease in speech recognition scores as the SNR decreased (more negative). Scores obtained at lower SNRs (−12.5dB and −15 dB) are dispersed more widely than those obtained at higher SNRs (−5 dB and −10 dB), indicating greater variability in speech recognition performance under more adverse SNR conditions.

All statistical data analyses in the present study were performed using Statistical Package for the Social Sciences (SPSS) Version 16 software.

Validity—Table 1 shows the mean performance scores and the corresponding SDs for the AFAAF scores and the FAAF scores. It is important to note that the calibration strategy that was used in the present study was different from that was reported in Foster and Haggard (1987). In our study, presentation levels of both the speech and the masking noise were determined based on the long-term RMS level. In Foster and Haggard (1987), the speech level was determined by the level of a 1000Hz tone which equaled the mean peak level of the test words, while the masking noise level was calibrated using long-term RMS level. A comparison was carried out between the calibration strategies used for determining speech and masking noise levels for the AFAAF and the FAAF. Results revealed that the calibration strategy used in Foster and Haggard (1987) resulted in a 4 dB higher nominal SNR than that used in the present study when the same speech and masking noise were used. Therefore, the SNRs reported in Foster and Haggard (1987) were adjusted to an SNR scale which was equivalent to the present study to allow for a more direct comparison. Five SNRs used in Foster and Haggard (1987), which were similar to those used for the AFAAF, were selected for further comparisons. These equivalent SNRs for the FAAF and the SNRs for the AFAAF are shown in Table 1.

PI functions for the two recordings were constructed for comparison (Figure 3). Each PI function was fit to the mean percent correct scores for the five SNRs according to a best-fit, three-parameter sigmoid function (Equation 1), which had been modified for data collected from a four alternative force choice test. The equation is expressed as:

$$y = \frac{a}{1 + e^{-\frac{x-b}{c}}} + 25 \quad (\text{Equation 1})$$

where x was SNR; y was the speech recognition performance in percent correct; a , b , and c were parameters of the function. The values of a , b , and c were adjusted differently for AFAAF data and FAAF data to produce the best fit using Equation 1. The two PI functions were monotonic and they were not forced to reach 100% correct. From Figure 3, it is seen that the PI functions for the AFAAF and the FAAF are similar. However, the PI function for the AFAAF is somewhat shallower than that for the FAAF. The slope of the PI function for each recording was determined using the first derivative of the fitting equation at the SNR where the performance level was 60% correct. The rationale for choosing the 60% correct performance was that this performance level was located close to the middle of the steepest portion of the PI functions for both the AFAAF and the FAAF scores. By this method, the slope of the PI function for the AFAAF data is 6%/dB, while the slope of the PI function for the FAAF data is 7 %/dB.

Foster and Haggard (1987) examined the score differences between adjacent SNRs that differed by 2.5 dB for the FAAF scores. They reported that each two adjacent SNRs provided significantly different performance. This examination also was carried out in the present study. For statistical analyses, the percentage correct scores were converted into rationalized arcsine units (rau; Studebaker, 1985) to homogenize the variances. The score range of rau is in theory from -23 to $+123$. A repeated-measures analysis of variance (ANOVA) was used to compare the scores obtained in the five SNRs. The main effect of SNR was significant ($F[4,76] = 419.307$, $p = 0.000$). Preplanned contrasts between each two adjacent SNR conditions were tested to explore this main effect. Results revealed that speech recognition scores obtained in adjacent SNRs were all significantly different (Table 2). This finding bolstered the validity of the PI function for the AFAAF recording.

Reliability—In Foster and Haggard (1987), test-retest reliability was assessed using scores collected from the right and the left ears. In the present study, each participant was tested in the five SNR conditions in each ear. We planned also to use the left and right ear scores to assess test-retest reliability of the AFAAF test. Although this is not a typical approach to evaluate test-retest reliability, it was reasonable to analyze the scores for the two ears as long as any order/fatigue effect was ruled out and the scores from the two ears were similar.

As described earlier, the order of the tested ear was counterbalanced for all the participants. To assess the possible order/fatigue effect, scores from the first tested ears and the second tested ears were compared. The mean difference score and the corresponding SD for each SNR are shown in Table 3. The difference scores range from -0.12 to -2.4 rau. The scores from the first and the second tested ears were subjected to a within-subject repeated ANOVA to test whether there was an order/fatigue effect. No overall order/fatigue effect was observed across the five SNR conditions ($F[1,19] = 2.769$, $p = 0.113$). The interaction effect between order and SNR was not statistically significant ($F[4,76] = 0.667$, $p = 0.617$). Therefore, it was concluded that there was no order/fatigue effect.

Since there was no order/fatigue effect, the speech recognition scores collected from the left and the right ears could potentially serve as test-retest reliability measurement if the scores from the two ears were similar. For each participant, the difference score between the two ears for each SNR was calculated by subtracting the right ear score from the corresponding

left ear score. There were 100 difference scores between the two ears (20 participants and five SNRs). The mean difference scores were listed in Table 3. The mean difference scores (left minus right) for the five SNR conditions were between -0.67 rau and -3 rau. The scores from the left and the right ears were subjected to a within-subject repeated ANOVA to test whether the scores for the two ears were significantly different. A significant main effect of ear was obtained ($F[1,19] = 6.779, p = 0.017$), which revealed that the scores for the right ears were overall significantly higher than the scores for the left ears. Interaction effect between ear and SNR was not significant ($F[2.913, 55.354] = 0.651, p = 0.582$, using the Greenhouse-Geisser correction for a significant test of Sphericity). Although the main effect of ear was statistically significant, the difference scores between the two ears in each SNR condition were small (Table 3). Therefore, it is arguable that the scores for the left and the right ears were still able to serve as the assessment of test-retest reliability.

Figure 4 depicts the test and retest scores of the AFAAF. Each symbol on the figure was plotted based on the speech recognition scores for the left and right ears for each participant at each SNR. One method to evaluate test-retest reliability is to perform correlation analyses between test and retest scores. Because the participants recruited in this study were all young listeners with normal hearing, the between-subject variability of the speech recognition scores was small. As a consequence, the data points for each SNR were clustered together (see Figure 4). Therefore, correlation analyses between scores obtained from the right and the left ears were inappropriate for assessing test-retest reliability in this study. An alternative method is to compute critical difference (CD).

CD is used to examine whether two speech recognition scores from the same listener are significantly different (see Thornton & Raffin, 1978). There are a number of ways to calculate CDs (e.g., Cox, Alexander, & Gilmore, 1987; Demorest & Walden, 1984; Thornton & Raffin, 1978). In the present study, CDs were calculated for the AFAAF scores (in rau) based on the distribution of the difference scores between right and left ears. A 95% CD and a 90% CD are calculated using the following equations:

$$95\%CD = 1.96 \times SD \quad (\text{Equation 2})$$

$$90\%CD = 1.65 \times SD \quad (\text{Equation 3})$$

where SD is the standard deviation of the difference scores. In this study, when analyzing the 100 difference scores, the SD of the difference scores is 5.9 rau. Thus, the 95% CD was computed by multiplying the SD of the distribution of test-retest differences by 1.96. This yielded a value of 12 rau. Therefore, when using the steady-state speech-matched noise as the masker, the 95% CD of the 80-item AFAAF test scores for listeners with normal hearing is 12 rau. In Figure 4, the 95% CD is demonstrated by the two dashed lines drawn 12 units in parallel with the main diagonal line. If a 90% CD is more appropriate for a particular application, the corresponding CD is computed by multiplying the SD of the distribution of test-retest differences by 1.65. The 90% CD of the 80-item AFAAF test scores for listeners with normal hearing is 10 rau.

General discussion

The goal of this project was to produce an American dialect recording of the FAAF test. In this paper, the rationales and procedures for recording and evaluating the AFAAF were reported. The 80 test words and five practice words were recorded and digitally edited. Masking noises with spectra identical to the speech signals were created. Validity and reliability of the newly produced materials were assessed in 20 young listeners with normal hearing. The results from the validation study show that the AFAAF is similar to the FAAF. Although efforts have been made to produce a recording that is identical to the FAAF in a different English dialect, the two recordings are not equal.

As shown by the PI functions for the AFAAF and the FAAF (Figure 3), speech recognition performance for the two recordings is similar at favorable SNR conditions (better than -7.5 dB SNR). However, speech recognition performance drops more rapidly with the FAAF than it does with the AFAAF as SNR decreases. As a result, the slope of the PI function for the FAAF is somewhat steeper than that for the AFAAF. This implies that the speech recording of the FAAF is slightly less intelligible than that of the AFAAF in the poorer SNR conditions.

The speech materials for the AFAAF and the FAAF were both spoken by male talkers in a clearly articulated manner. The observed difference in the PI functions between the two recordings suggests that the speech produced by the AFAAF talker is slightly more intelligible than that produced by the FAAF talker. Previous research concerned with clear speech found that clear speech was more intelligible than less clear speech (e.g., conversational speech) for normal hearing listeners (Payton, Uchanski, & Braida, 1994). Moreover, the difference in speech intelligibility performance between clear and less clear speech increased as the masking noise levels increased (Payton et al., 1994). Speaking rate and inherent acoustic properties of speech also contribute to the advantage in speech intelligibility with clearer speech (see Picheny, Durlach, & Braida, 1986). Difference in speaking rate between the two talkers was assessed in the current study. Since the same 80 utterances were spoken for the AFAAF and the FAAF recordings, speaking rate for each recording could be evaluated by measuring the mean duration across all the utterances. The mean duration for the AFAAF is 1.8 seconds ($SD = 0.6$), while the mean duration for the FAAF is 1.6 seconds ($SD = 0.5$). This difference in duration indicates that the talker for the AFAAF spoke somewhat more slowly than the talker for the FAAF. This is consistent with a finding that the speech produced for the AFAAF is a little more intelligible than that for the FAAF, especially at poorer SNRs.

Test-retest reliability for the AFAAF scores was found to be similar to that for the FAAF scores. Foster and Haggard (1987) used the left and right ear speech recognition scores for evaluating test-retest reliability of the FAAF. They reported that the mean left-right differences were within 2.5%. The same strategy was employed in the present study and the mean left-right ear differences for the five SNRs were no greater than 3 rau (Table 3). In addition, the 95% and 90% CDs were computed for the AFAAF scores as a measure of reliability and their values were 12 rau and 10 rau, respectively. CDs were not reported by Foster and Haggard (1987). However, SDs for difference scores between the left and the

right ears were reported and these were used to compute CDs for the FAAF scores. According to Studebaker (1985), the difference between percentage scores and the corresponding rau scores are almost the same at all percentage values between 15% and 85%. In Foster and Haggard (1987), mean performance scores for five SNRs, from -12.5 to -2.5 dB nominally, were from 31% to 85.1%, respectively. Thus, the individual listener's scores in percentage for these five SNRs were considered identical to rau scores. The mean SD in rau was then calculated across these five SNRs and the value was 8.3 rau. Therefore, the estimated 95% and 90% CDs of the 80-item FAAF scores for normal hearing listeners are 16 rau and 14 rau, respectively. Comparing the CDs for the AFAAF scores and the estimated CDs for the FAAF scores, it is seen that the CDs for the AFAAF scores are 4 rau smaller than those for the FAAF scores. This suggests that test-retest reliability of the AFAAF is at least equal to that of the FAAF test.

As noted before, one expected advantage of the FAAF test format is that it has minimal learning effect (Foster & Haggard, 1987). This was examined in the current study by comparing the scores for the first tested ears and the second tested ears (Table 3). The relatively small mean difference value for each SNR is consistent with the expectation that the AFAAF has minimal learning effect.

Based on the above discussion, the AFAAF maintains the advantages of the FAAF. Despite the great care taken to produce the AFAAF recording, the two recordings were still not entirely identical, although they are very similar. The AFAAF is thus a valid and reliable test to evaluate speech recognition performance for listeners who are native speakers of American English.

According to Foster and Haggard (1987), the FAAF was not designed as a speech recognition test of disability. The minimal pair structure of the AFAAF makes it a consonant level test that is more sensitive to high frequency characteristics of the test signal. Because of this, the FAAF has been intensively used in hearing aid research in the United Kingdom (e.g., Gatehouse et al., 2003). It is reasonable to expect that the AFAAF also will be widely used in the same type of research in the United States. It has proven to be a useful tool for evaluating outcomes of amplification.

Summary

The FAAF is a four-alternative closed-set speech recognition test, which comprises 80 monosyllabic test words. This test has a number of advantages, such as minimal learning effect, sensitivity to consonant recognition, and automatic scoring. The original FAAF was developed in the United Kingdom and has been widely used in hearing research. However, the original test materials were recorded in British English dialect and it is not ideally applicable for use in the United States. In this project, the FAAF test materials were recorded in American English dialect. The newly recorded test was dubbed the AFAAF. The validity and reliability of the AFAAF data were compared to those for the FAAF data, showing that the two recordings are highly similar but not identical. The results support the use of the AFAAF for measuring speech recognition performance in listeners who have American English as a native language.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Manuscript preparation was supported by National Institute on Deafness and Other Communication Disorders grant R01DC011550. Many thanks to John Foster from the MRC-Institute of Hearing Research for permission to reproduce the FAAF (personal communication on 10/27/2008), and David J. Wark, Ph.D. from The University of Memphis for serving as the talker.

Abbreviations

CD	Critical Difference
FAAF	Four Alternative Auditory Feature test
HARL	Hearing Aid Research Lab
ICRA	International Collegium of Rehabilitative Audiology
RAU	Rationalized Arcsine Unit
RMS	Root Mean Square
SD	Standard Deviation
SNR	Signal-to-Noise Ratio

References

- Boothroyd A. Auditory perception of speech contrasts by subjects with sensorineural hearing loss. *Journal of Speech, Language and Hearing Research*. 1984; 27(1):134–144.
- Boothroyd A. Evaluation of speech production of the hearing impaired: Some benefits of forced-choice testing. *Journal of Speech, Language and Hearing Research*. 1985; 28(2):185–196.
- Boothroyd A. Perception of speech pattern contrasts from auditory presentation of voice fundamental frequency. *Ear and hearing*. 1988; 9(6):313–321. [PubMed: 3220183]
- Cox RM, Alexander GC, Gilmore C. Development of the connected speech test (CST). *Ear and Hearing*. 1987; 8(5):119s–126s. [PubMed: 3678650]
- Davies JE, John DG, Jones MJ. Hearing aids for high-frequency hearing loss. *Clinical Otolaryngology & Allied Sciences*. 1990; 15(4):321–326. [PubMed: 2225500]
- Demorest ME, Walden BE. Psychometric principles in the selection, interpretation, and evaluation of communication self-assessment inventories. *Journal of Speech and Hearing Disorders*. 1984; 49(3):226–240. [PubMed: 6748618]
- Dreschler WA, Verschuure H, Ludvigsen C, Westermann S. ICRA Noises: Artificial Noise Signals with Speech-like Spectral and Temporal Properties for Hearing Instrument Assessment: Ruidos ICRA: Se{ñ}ates de ruido artificial con espectro similar al habla y propiedades temporales para pruebas de instrumentos audit. *International Journal of Audiology*. 2001; 40(3):148–157.
- Fairbanks G. Test of phonemic differentiation: The rhyme test. *The Journal of the Acoustical Society of America*. 1958; 30:596–600.
- Foster JR, Haggard MP. The four alternative auditory feature test (FAAF)-linguistic and psychometric properties of the material with normative data in noise. *British journal of audiology*. 1987; 21(3):165–174. [PubMed: 3620751]
- Gatehouse S, Naylor G, Elberling C. Benefits from hearing aids in relation to the interaction between the user and the environment. *International Journal of Audiology*. 2003; 42:S77–S85. [PubMed: 12918613]

- Gelfand, SA. Essentials of audiology. 3. New York: Thieme; 2009.
- House AS, Williams CE, Hecker MHL, Kryter KD. Articulation-Testing Methods: Consonantal Differentiation with a Closed-Response Set. *The Journal of the Acoustical Society of America*. 1965; 37:158–166. [PubMed: 14265103]
- Kreul EJ, Nixon JC, Kryter KD, Bell DW, Lang JS, Schubert ED. A proposed clinical test of speech discrimination. *Journal of Speech, Language and Hearing Research*. 1968; 11(3):536–552.
- Milchard AJ, Cullington HE. An investigation into the effect of limiting the frequency bandwidth of speech on speech recognition in adult cochlear implant users. *International journal of audiology*. 2004; 43(6):356–362. [PubMed: 15457818]
- Nakagawa, S.; Shikano, K.; Tohkura, Y. Speech, hearing and neural network models. Amsterdam: IOS press; 1995.
- Owens E, Schubert ED. Development of the California consonant test. *Journal of Speech, Language and Hearing Research*. 1977; 20(3):463–474.
- Payton KL, Uchanski RM, Braida LD. Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. *The Journal of the Acoustical Society of America*. 1994; 95:1581–1592. [PubMed: 8176061]
- Pederson OT, Studebaker GA. A new minimal contrasts closed-response-set speech test. *Journal of Auditory Research*. 1972; 12:187–195.
- Picheny MA, Durlach NI, Braida LD. Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *Journal of Speech, Language and Hearing Research*. 1986; 29(4):434–446.
- Shields PW, Campbell DR. Intelligibility, subjective ratings and completion time scores using the FAAF test with hearing-impaired subjects and noisy reverberant environments. *British journal of audiology*. 2001; 35(4):237–245. [PubMed: 11694098]
- Studebaker GA. A “rationalized” arcsine transform. *Journal of Speech, Language and Hearing Research*. 1985; 28(3):455–462.
- Thornton AR, Raffin MJM. Speech-discrimination scores modeled as a binomial variable. *Journal of Speech, Language and Hearing Research*. 1978; 21(3):507–518.
- Weisleder P, Hodgson WR. Evaluation of four Spanish word-recognition-ability lists. *Ear and hearing*. 1989; 10(6):387–393. [PubMed: 2606290]
- Wilson WJ, Moodley S. Use of the CID W22 as a South African English speech discrimination test. *The South African journal of communication disorders. Die Suid-Afrikaanse tydskrif vir Kommunikasieafwykings*. 2000; 47:57–62. [PubMed: 11455824]

Figure 1. Long-term spectra of the AFAAF speech and the unmodulated talker-matched speech-spectrum noise. The spectra of the speech and noise were measured at the same digital RMS power using SpectraPLUS 5.0 software. Each signal was averaged for 60 seconds.

Figure 2.

Box-and-whisker plots show the speech recognition scores for each ear in each SNR condition. The bottom and top of the box are the 25th percentile and 75th percentile, respectively. The bar within each box is the 50th percentile. The top and bottom whiskers represent 10th percentile and 90th percentile, respectively. The circles and squares are data points that are outside 10th to 90th percentile range.

Figure 3. PI functions of the AFAAF and the FAAF. SNRs from the two studies were adjusted to equivalent calibration.

Figure 4.

Test-retest reliability of AFAAF scores (80 key words). Each symbol depicts the data collected from the right and the left ears for a subject at an SNR. The main diagonal line indicates equal performance between the two ears. The dashed lines are the 95% critical difference (12 rau) calculated from the test scores.

Table 1

Mean speech recognition scores for the AFAAF and the FAAF in five SNR conditions. In both cases, steady-state speech-shaped noises were used as masker. For the AFAAF, recognition scores in percent correct and rau are listed. For the FAAF, only the recognition scores in percent correct are listed. In addition, since the British FAAF was calibrated using a different scale, both the nominal SNRs (Foster & Haggard, 1987) and the SNRs equivalent to this study are given.

AFAAF (40 ears)				FAAF (32 ears)		
SNR (dB)	% (SD)	rau (SD)		Nominal SNR (dB)	Equivalent SNR (dB)	% (SD)
-15	41.8 (6.1)	42.3 (5.8)		-10	-14	40.5 (8.1)
-12.5	54.7 (6.8)	54.5 (6.4)		-7.5	-11.5	57.2 (9.7)
-10	69.6 (4.5)	68.9 (4.6)		-5	-9	72.4 (5.6)
-7.5	80.4 (3.5)	80.6 (4.2)		-2.5	-6.5	85.1 (6.5)
-5	86.9 (2.8)	88.8 (3.9)		0	-4	90.4 (3.8)

Table 2

Mean speech recognition score difference between adjacent SNRs. All the score differences are significantly different ($p < .001$).

	–15 dB	–12.5 dB	–10 dB	–7.5 dB
	–12.5 dB	–10 dB	–7.5 dB	–5 dB
Mean score difference (rau)	12.2	14.4	11.7	8.3
SD (rau)	5.2	5.7	5.0	5.5

Table 3

Mean speech recognition score difference in rau between (1) right and left ears; (2) first tested and second tested ears. For right-left ear comparison, a positive value indicates better performance with the left ear. For first-second tested ear comparison, a positive value indicates better performance with the first tested ear. No significant difference was observed ($p > 0.05$).

SNR (dB)	Mean score difference (N=20)	
	Right vs. Left ear (SD) *	First vs. Second tested ear (SD) **
-15	-0.92 (4.43)	-0.68 (4.47)
-12.5	-2.42 (5.83)	-2.40 (5.84)
-10	-3.01 (7.24)	-2.36 (7.49)
-7.5	-0.67 (5.61)	-0.94 (5.57)
-5	-2.32 (6.43)	-0.12 (6.86)

* Score difference = Left ear score – Right ear score

** Score difference = First tested ear score – Second tested ear score