Journal of Speech and Hearing Research, KILLION, TILLMAN, Volume 25, 15-25, March 1982

EVALUATION OF HIGH-FIDELITY HEARING AIDS

MEAD C. KILLION TOM W. TILLMAN
Auditory Research Laboratories
Northwestern University
Evanston, Illinois

An essential building block for any high-fidelity hearing aid is an amplifier-transducer-coupling combination that does not audibly degrade the sound, that is, provides high-fidelity sound reproduction as judged by someone with normal hearing. To demonstrate that such a combination is possible, two binaural pairs of hearing aids were assembled using available hearing aid transducers and electronic components, one pair of Over-The-Ear hearing aids with 8-kHz bandwidth and one pair of In-The-Ear hearing aids with 16-kHz bandwidth. Objective insertion-gain measurements on these aids, obtained with a KEMAR manikin in a diffuse sound field, revealed a frequency-response accuracy comparable to that available in expensive high-fidelity loudspeakers. Subjective fidelity ratings obtained from three groups of listeners judging prerecorded A-B-A comparisons (made from equalized eardrum-position microphones in a KEMAR manikin) produced a similar conclusion. We conclude that the important question for hearing aid research is no longer "What *can* a hearing aid be designed to do?" but "What *should* a hearing aid be designed to do for the hearing impaired?"

For quite some time, a common assumption has been that hearing aids are inherently low-fidelity sound reproducers. This assumption has had an inevitable impact on hearing aid research, much of which appears to have been directed toward making the best of a bad situation. In the meantime, transducer and amplifier technology has progressed to the point that high-fidelity sound reproduction is readily achievable in headworn hearing aids (although it may or may not be desirable in a given instance).

This paper describes the results of experiments undertaken to demonstrate that high-fidelity hearing aids are now practical using available transducers and electronic components. These experiments were performed on two pairs of experimental headworn hearing aids, one binaural pair of Over-The-Ear (OTE) hearing aids with 8-kHz bandwidth and one binaural pair of In-The-Ear (ITE) hearing aids with 16-kHz bandwidth.

EXPERIMENTAL HEARING AIDS

The ITE aids were assembled with BT-1759 microphones (Killion & Carlson, 1974), BP-1712 earphones (Carlson, Mostardo, & Diblick, 1976), and BF-1921 acoustic damping-resistance elements (Carlson & Mostardo, 1976), all manufactured by Knowles Electronics. The "16KM" earmold construction (16 kHz earmold developed by E. Monser) was employed (Killion, 1979a).

The OTE aids were assembled using (a) experimental EA-type microphones (XD-1116) coupled to the hearing aid sound inlet with 10 mm of 1.5-mm diameter tubing, and (b) BP-1712 earphones compliantly mounted in commercial OTE hearing aid cases with a 10-mm length of 1.1-mm diameter rubber tubing coupling the earphone to the earhook whose sound channel was 23 mm long and had a 1.2-mm internal diameter. The earmolds were of the "8CR" construction (Killion, 1981).

The overriding importance of proper earmold acoustics has been discussed by Knowles and Killion (1978). The novel earmold construction used with these experimental hearing aids was pivotal to achieving the design goals.

The amplifiers used with the experimental hearing aids were designed as practical hearing aid amplifiers and were operated on 1.5-V "S76" hearing aid cells. They were assembled using "breadboard" construction mounted in pocket-sized cases: Reducing a discrete-component breadboard amplifier to subminiature dimensions is a feat regularly accomplished by hearing aid designers and was not considered an important part of this investigation.

A more complete description of the experimental hearing aids, as well as an extensive set of design guidelines for high-fidelity hearing aids, can be found in Killion (1979a).

OBJECTIVE PERFORMANCE MEASURES

Frequency Response

Coupler response. The Zwislocki coupler response of the completed OTE aids is shown in Figure 1 as a solid curve. The response peak near 2700 Hz was included by design to compensate for the loss of external-ear resonance which occurs when the ear canal is occluded by an earmold (Knowles, 1968, Note 1).

It is useful at this point to introduce formally the term *insertion gain*, which is the ratio of eardrum pressure produced by a hearing aid to the eardrum pressure pro-

0022-4685/82/2501-0015$01.00/0

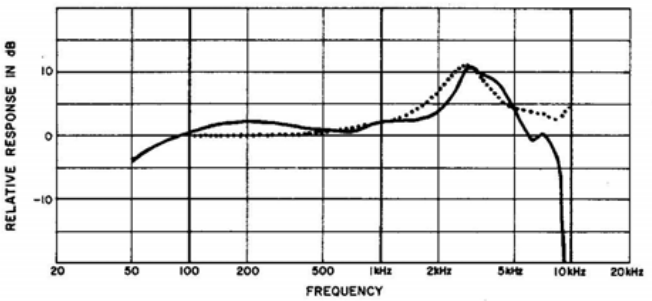

FIGURE 1. Zwislocki coupler response of experimental Over-The-Ear hearing aid (—) compared to OTE CORFIG (·····).

duced without the hearing aid (Dalsgaard & Jensen, 1974). Expressed in dB, the insertion gain of a hearing aid is the difference between aided and unaided eardrum sound pressure levels. [Similar terms are *orthotelephonic gain, etymotic gain,* and *functional gain,* with functional gain generally reserved for subjective measurements of insertion gain. See Burkhard (1978) for further discussion of these terms.]

The design goal, which was the estimated Zwislocki Coupler Response required for Flat Insertion Gain (CORFIG), is shown as a dotted curve in Figure 1. This curve applies to OTE hearing aids and is based on measurements with a KEMAR manikin (Burkhard & Sachs, 1975) in a diffuse (random-incidence) sound field, as described by Killion and Monser (1980). Note that the measured result shown in Figure 1 agrees with the design goal within ±3 dB up to nearly 8 kHz, which was the design cutoff frequency.

The Zwislocki-coupler response of the completed ITE aids is shown in Figure 2 (solid curve) compared to the estimated random-incidence CORFIG response goal for ITE aids. Here it is clear that the simple amplifier equalization used with the ITE aids did not adequately compensate for the loss of external-ear resonance. (Simple equalization was adequate with the OTE aids because the compensation was designed into the 8CR earmold response characteristics.) Since the time these aids were designed, however, the Knowles ED-series earphone has become available. When coupled with the 16KM earmold, that earphone produces a Zwislocki-coupler frequency response which more nearly duplicates the estimated ITE CORFIG (Knowles Electronics, Note 2).

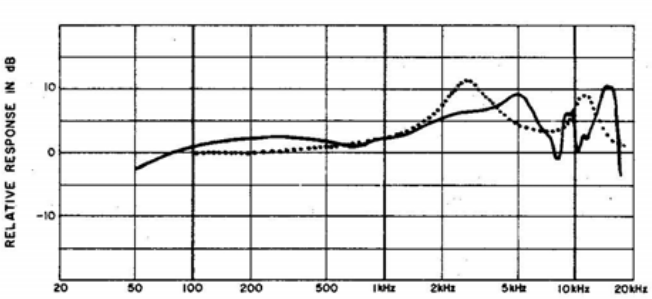

FIGURE 2. Zwislocki coupler response of experimental In-The-Ear hearing aid (—) compared to ITE CORFIG (·····).

Insertion gain. Actual insertion-gain measurements, obtained during the listening-test recording sessions described below, were performed using one-third-octave bands of noise and a KEMAR manikin. The results (shown later in Figure 5) provide an estimate of the frequency response a user of these hearing aids would experience listening to a live concert performance or a stereo high-fidelity system at home.

Because one of the intents of the study was to demonstrate that a basic "building block" hearing aid arrangement with high-fidelity performance was possible, the ability of the experimental aids to provide useful gain without feedback problems was verified in separate experiments by increasing the electrical gain of the preamplifier. Objective measurements using the KEMAR manikin verified the ability of the experimental OTE aids to provide 30 to 40 dB of insertion gain while maintaining an 8-kHz bandwidth. Similar amounts of "full on" gain were obtained in direct listening tests with well-fitted earmolds.

Calculated accuracy scores. A procedure based on loudness calculations was adopted recently by Consumers Union for rating the frequency-response accuracy of high-fidelity loudspeakers ("How CU's Audio Lab," 1977). The accuracy scores for 16 models of "low-priced" ($100–$200 per pair) high-fidelity loudspeakers ranged between 63% and 93%, with a median value of 80%. Listening tests were said to have borne out the utility of the accuracy scores, although "experience has taught us that a group of listeners won't readily agree on which of two speakers is more accurate when the speaker's scores differ by eight points or less" ("Low-priced loudspeakers," 1977, p. 406).

More recently, a group of expensive ($600–$1000 per pair) "State of the Art" loudspeakers was tested ("High-priced loudspeakers," 1978). The median accuracy score for those loudspeakers was 89%.

By applying a procedure comparable to that used by Consumers Union, it was possible to calculate an accuracy score corresponding to the insertion-gain frequency-response curves measured on the OTE and ITE hearing aids (Killion, 1979a). The results of that process yielded an accuracy score for the OTE aids of 82% and an accuracy score for the ITE aids of 91%. Each score exceeded the median of the inexpensive and expensive (respectively) groups of high-fidelity loudspeakers discussed above, leading to the conclusion that the frequency response accuracies of the experimental aids fell in the "high fidelity" category by that measure.

Maximum Undistorted Output

The instantaneous peak sound pressure level (SPL) which the experimental OTE hearing aids could produce without distortion is shown in Figure 3 (solid curve) compared to the peak eardrum SPL required to reproduce a full symphony orchestra in live performance (dotted curve) as estimated elsewhere (Kil-

lion, 1979a). Substantially greater undistorted output could have been obtained at the expense of an increase over the .7-mA battery drain of the experimental OTE hearing aids, as illustrated by the earphone-overload (dashed) curve in Figure 3.

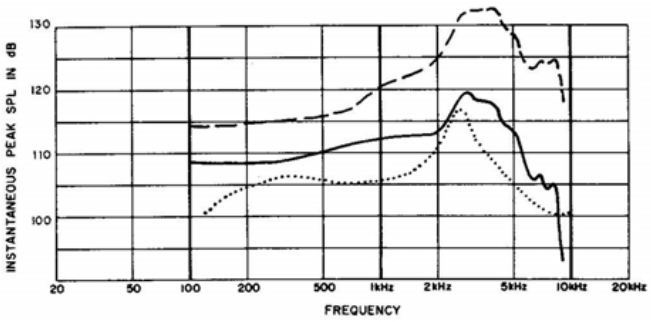

FIGURE 3. Peak eardrum pressure requirements (• • • • •) compared to maximum linear output of BP-1712 earphone with 8CR earmold, as limited by:
earphone overload (— —)
clipping in .7-mA experimental amplifier (——).

Total Harmonic Distortion

A large amount of negative feedback and a low output impedance were designed into the experimental amplifiers so that nonlinear distortion was not expected to be a problem in the experimental hearing aids. (The distortion of the BP-series earphones themselves is normally low compared to that produced by typical amplifiers.) This expectation was confirmed by a series of swept-frequency measurements of second- and third-harmonic distortion, obtained for inputs of 60, 70, 80, 90, 100, 105, and 110 dB SPL. At no frequency did hearing aid distortion—measured in a Zwislocki coupler—exceed 1% for inputs of 100 dB SPL or less. Plots of CCIF-intermodulation distortion, obtained for a 200-Hz difference frequency, showed a similar result.

Data on total harmonic distortion versus output—measured in a Zwislocki coupler—also were obtained at a fixed 500-Hz input frequency. Those data are shown plotted in Figure 4 (solid curve). Below 105-dB-SPL output, the measured total harmonic distortion is roughly one-fourth the maximum inaudible hearing aid distortion for music and speech (dashed curve) estimated by Killion (1979a). The estimate of the maximum inaudible distortion levels for speech and music may seem high to readers accustomed to seeing high-fidelity amplifier distortion ratings below .005%, but it is consistent with recent psychoacoustic (listening-test) evidence as reported by Gabrielsson, Nyberg, Sjögren, and Svensson (1976), Milner (1977), and Davis (1978) when the available data are referred to eardrum pressure levels.

The abrupt increase in measured distortion above 105 dB SPL corresponds to the onset of amplifier clipping, which was determined by the choice of earphone impedance and battery drain used in the experimental OTE aids. A battery drain of .7 mA was sufficient to meet

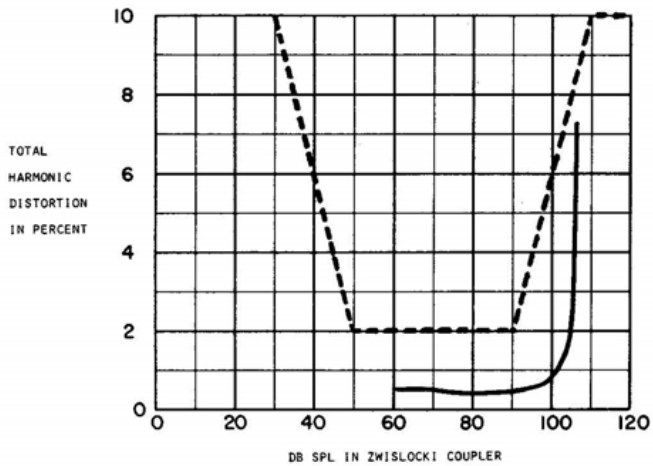

FIGURE 4. Total harmonic distortion at 500 Hz versus output for the OTE hearing aids (——) compared to estimated maximum inaudible distortion (– – –).

the undistorted 105-dB instantaneous-peak SPL design goal. By way of reference, a .7-mA battery drain corresponds to nearly 2 weeks of continuous 16-hour-per-day operation with a 1.5-volt S76 hearing aid battery.

SUBJECTIVE FIDELITY RATINGS

Method

The technique with the greatest face validity for rating the fidelity of a sound reproduction system is to compare the reproduced sound with the original sound. This approach was used by Olson (1957) in his famous 1947 demonstration in which the Boston Symphony Orchestra was compared with a phonograph recording of the orchestra before an overflow audience in the music shed at Tanglewood, Massachusetts.

Although a true live-versus-recorded listening test has excellent face validity, it becomes impractically cumbersome when several different sound reproduction systems are to be tested. Villchur (1962) used prerecorded stimuli in a "simulated live-versus-recorded" technique, where the source material was itself a reproduction of previously recorded material. To use this technique for comparisons employing musical reproductions, for example, a loudspeaker with good dispersion (output nearly the same in all directions) is chosen as a "reference" loudspeaker. Anechoic chamber recordings of that loudspeaker reproducing musical selections from a master tape are then obtained, just as if that reference loudspeaker were itself a group of live musicians. The simulated-live-versus-recorded comparisons are subsequently presented between (a) the reference loudspeaker reproducing the original master tape recording (the simulated live source) and (b) the loudspeaker under test reproducing the anechoic-chamber recording of that simulated live source.

If the loudspeaker system chosen for the "surrogate

live source" is found to have a sensibly flat frequency response, the assumption can be made that the anechoic-chamber recording of that speaker's *output* will be sensibly equivalent to its electrical *input,* in which case the anechoic-chamber rerecording may be dispensed with. That approach was chosen for the present experiments. AR3a loudspeakers were selected for the reference loudspeakers because they have successfully passed true live-versus-recorded listening tests (Villchur, 1964).

A KEMAR manikin was used as a "surrogate listener," with the output of its eardrum-position microphone fed through a pair of bridged-T filters (Killion, 1979b). These filters provided equalization accurate to within ±3 dB of a flat frequency response for KEMAR manikin recordings made in a diffuse sound field.

The equalization filters were present under all experimental conditions so that the recordings could be subsequently reproduced over either loudspeaker or conventional headphones without introducing the "duplicate ear canal resonance problem." To explain: Commonly available loudspeakers are designed to produce a relatively flat frequency response in the sound field, and the better high-fidelity headphones are designed to produce a flat sound-field-referenced response (Martin & Anderson, 1947). As a result, the eardrum-pressure frequency response which they produce will exhibit a peak of roughly 15 dB at 2.7 kHz due to the effect of external-ear resonances (Shaw, 1980). When added to the roughly 15-dB peak introduced by the external-ear resonances in the manikin, a duplication of resonances occurs. (The subjective result of such a duplication is a single 15-dB peak because the peak introduced by the subject's own external-ear resonances is a normal part of his experience.) Although the same peak would be added to both the reference and comparison sounds, such a large peak is likely to introduce a bias in favor of systems with compensating deficiencies and should be avoided in listening-test experiments of this sort.

Preparation of Prerecorded Comparison Materials

Stimuli: Master stimulus tape. Six selections of program material, chosen to best expose a variety of potential deficiencies in the systems under test, were spliced together to form a "master stimulus tape." One selection was an anechoic chamber recording of repeated nonsense sentences ("Joe took father's shoebench out; she was sitting at my lawn.") spoken by one of the writers. One selection was 15 seconds of "speech spectrum noise," that is, broadband noise filtered to provide approximately the long-term average spectrum of speech.

The remaining four selections were musical passages dubbed from virgin pressings of commercial recordings. Two of the passages were taken from a New York Philharmonic recording of the Beethoven Violin Concerto in D (Columbia stereo record M33587) and two from an Oscar Peterson piano trio recording of Peterson's cheerful blues "The Smudge" (Mercury stereo record EMC-2-405). One of the orchestral passages was fortissimo, the other was forte. All passages were chosen as relatively unchanging through the switchover region from the reference system A to the comparison system B to allow the most sensitive A-B comparisons.

Comparison systems. A range of popular high-fidelity systems was included in the fidelity-rating experiment to serve as benchmarks against which to compare the fidelity rating given the experimental hearing aids.

One system was a pair of popular high-efficiency two-way studio monitor loudspeakers. A second system was a popular stereo headphone designed to produce a wide bandwidth with (intentionally) exaggerated bass response, a design which presumably accounts for its popularity in "hi-fi" dealers' showrooms.

A third system was a simulated speech audiometer obtained by using a pair of TDH-39 earphones (in MX-41/AR cushions) which were factory selected to have a frequency response nearly identical to the published "typical" response curve. [A commercial speech audiometer was not used because of the ±5-dB frequency-response tolerance and the 5–10% equivalent total harmonic distortion at ±6 dB VU (no other distortion test is specified) allowed by ANSI Standard S3.6 (1969). Rather, the same amplifier and tape reproducer used with the reference system were used to simulate an essentially flawless speech audiometer.]

As a representative from the low end of the range of systems advertised as *high fidelity*, an inexpensive stereo phonograph (typically sold at discount department stores) was included in the comparisons.

Finally, a transistor pocket radio (purchased in 1976 for $4.95) was included to serve as a low-fidelity anchor for the fidelity-rating scale.

The abbreviated designations for these seven comparison systems are as follows:

1. Pocket radio (PR)
2. Discount stereo (DS)
3. Simulated audiometer (SA)
4. Popular head phones (PP)
5. Monitor speakers (MS)
6. In-the-ear hearing aids (ITE)
7. Over-the-ear hearing aids (OTE)

The relative frequency responses of six of the seven comparison systems are shown in Figure 5. Each response was obtained during the course of the recording sessions by subtracting the manikin-sensed reference curve from the manikin-sensed response of the sound system under test. All curves were obtained using one-third octave bands of noise. Note that the hearing aid response curves (ITE and OTE) in Figure 5 are simply insertion-gain curves of those hearing aids on the KEMAR manikin. The remaining response curves represent *difference* curves and reflect only the accuracy to which the system under test could duplicate the room response of the AR3a loudspeakers used in the reference high-fidelity system. By oversight, no frequency response was obtained for the discount stereo system.

Comparison recordings: Four-track master comparison tapes. The master stimulus tape was reproduced on a 2-track Ampex 440 professional tape recorder whose output was fed through a 125-watts-per-channel Marantz

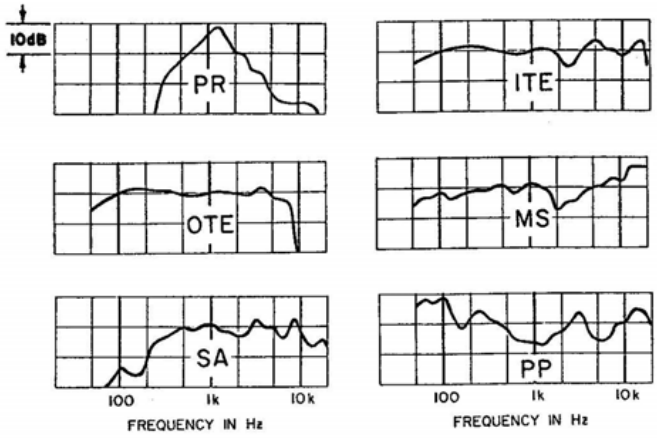

FIGURE 5. Frequency responses of sound systems used in listening tests. Note: Hearing aid responses are insertion gains; all others are relative to reference loudspeaker system response.

250 stereo amplifier to a pair of Acoustic Research AR3a loudspeakers spaced along one wall of a 170 m^3 (6000 cu. ft.) room in the Auditory Research Laboratories of Northwestern University. The room dimensions were 6.1 m × 7.6 m × 3.7 m high. The sound absorption treatment on the walls and floor of that room was adjusted to eliminate audible flutter echos and to provide the .3–.5 sec reverberation time typically recommended for recording studios of that volume (e.g., see Olson, 1957).

With the master stimulus tape reproduction as source, binaural master comparison tapes were recorded from the output of eardrum-position microphones in the KEMAR manikin, as just discussed. The manikin was placed 1 m to the right of the room midline and 3.3 m from the wall along which the AR3a reference loudspeakers were located.

The hearing aid comparison recordings were obtained under exactly the same reproducing and recording conditions used for the reference recordings except that the OTE or ITE hearing aids were placed on the manikin and set to unity insertion gain.

The loudspeaker comparison recordings (MS and DS) were obtained with the loudspeakers substituted for (placed on the same 1-m-high stands previously occupied by) the AR3a reference loudspeakers. The monitor loudspeakers had "high-frequency roll-off" controls which were set to the position marked "flat."

The headphones were adjusted on the KEMAR manikin—with the help of tape and discs of closed-cell foam—to produce as close as could be estimated the equivalent of a real-ear seal and/or pinna deformation. For the TDH-39/MX-41AR headphones, the low-frequency attenuation due to the well-known leak around the ear cushions was made equal to the average obtained from probe-tube measurements on real ears as given by Shaw (1966) and confirmed by the authors and L. Young.[1] Between 200 and 10,000 Hz, the resulting "eardrum pressure" response measured on the unequalized KEMAR manikin fell within 2–4 dB of the predicted real-ear response calculated from Shaw's data for a typical TDH-39/MX-41AR earphone.

The pocket radio was located in the pocket of a shirt worn by the manikin. With the exception of the discount stereo (DS) and pocket radio ((PR) systems, all loudspeakers and headphones were driven from the output of the same stereo amplifier used with the reference loudspeakers. The headphones were driven through a 20-db passive attenuator with 10-ohm output impedance, an attenuator required to bring the 125-watt amplifier outputs down to suitable earphone-drive levels. The amplifiers in the DS and the PR were included in the listening-test recordings of those two systems. Both amplifiers produced noticeable distortion at high levels.

The creation of the final A-B-A comparisons was simplified by using an Ampex 440 4-track recorder to record the output from the manikin. The reference loudspeaker reproduction was recorded on one pair of tracks, the tape rewound, and the comparison reproduction subsequently recorded in synchrony on the other pair of tracks. To preclude the possibility of high-frequency tape overload, a 15-ips tape speed and Ampex 456 mastering tape were used throughout, with the OVU recording level (200 nW/m) set 20 dB below tape saturation. Under those conditions, the A-weighted noise level on the tape was approximately 60 dB below OVU.

Listening-test recordings: Binaural listening comparison tapes. The A-B-A comparisons were recorded on a 2-track Ampex 440 recorder from the 4-track master comparison tape by switching between track pairs at appropriate times. For the four musical passages, therefore, a continuous musical passage—the middle portion of which had been reproduced over the comparison system—was recorded. (Within each musical-selection block, we attempted to hold the switchover points to the same beat of the same measure for all comparisons.) For the live voice, the same "Joe . . . lawn" nonsense sentence was recorded three times, the second time representing the comparison-system reproduction.

The A-B-A comparisons were organized into six program-selection blocks, each containing seven system-comparison units. Each unit consisted of a spoken comparison-identification number, a 5-sec (approximately) segment from the reference system (A) recording, a 5-sec segment of the comparison system (B) recording, and another 5-sec segment of the reference (A). The same A-B-A comparison was repeated to permit a "second listen" to each comparison. Including pauses, each complete unit occupied about 40 seconds. (The total of 6 × 7 = 42 comparison units occupied just under 30 minutes after the program-selection block announcements were included.)

Within each of the six program-selection blocks, the first comparison unit was always for the low-fidelity pocket radio, but the remaining six comparisons were randomized according to a Latin-square form of randomized-block design. (Thus, each of the six nominally high-fidelity sound systems was represented once in every position in the presentation order.)

[1]Unpublished probe-tube data obtained on six subjects at Northwestern University, 1976.

Subjects and Experimental Procedures

Subject groups. The three subject groups of Untrained Listeners, Golden Ears, and Trained Listeners are described next.

To obtain as close as possible a "man-on-the-street" jury, a group of 24 Untrained Listeners was selected by the personnel department of a manufacturing concern to meet only the following criteria: equal male-female representation, approximately rectangular age distribution between ages 20 and 60, and as wide a distribution of occupations as obtainable. (The final criterion was included to avoid the possibility of a heavy technical representation on the listening jury.) The resulting jury contained 12 men and 12 women, with 8 subjects in their twenties, 5 in their thirties, 4 in their forties, 6 in their fifties, and 1 61-year-old subject.

A second group of Golden Ears subjects was enlisted, consisting of 5 individuals (Alf Gabrielsson, Julian Hirsch, Hugh Knowles, Bruno Staffen, and Edgar Villchur). Each had devoted a large amount of time at some point in his life to the subjective evaluation of high-fidelity loudspeaker systems.

A third group of Chicago-area Trained Listeners was also included. This group consisted of 6 individuals (Elmer Carlson, Richard Peters, Daniel Queen, Eugene Ring, Robert Schulein, and Frederic Wightman), each of whom had considerable training in listening experiments, although not necessarily in high-fidelity-system evaluations.

Method of presentation. The Untrained Listeners were made available for two 1-hr. sessions on successive days. On each day, the subjects rated nine blocks of seven comparisons. The first three blocks were practice comparisons, although the subjects were not so informed. The order of the last six blocks was randomized differently for each day's presentations.

The comparisons were reproduced on a 1-track Ampex AG500 reproducer and presented over Electro Voice Sentry V loudspeakers driven by Crown D75 amplifiers in a cafeteria area having only minimal sound treatment. To eliminate obvious flutter echos, we placed five sheets of 2.5-cm thick acoustical foam along three walls.

On the first day, the gain of the reproducing system was set for peak sound-level meter readings of 93 dB on the fortissimo Beethoven passage; the background noise level was 52 dB(A). For the second day's session, the left and right channels feeding the loudspeakers were reversed to provide some counterbalancing for seating position, which remained the same both days. To obtain an estimate of the effect of different *S/N* ratios during the comparison presentations, we increased the level on the second day to 95-dB peaks and shut down the air-conditioning system, thereby reducing the background noise level to 46 dB(A). The *S/N* ratio during the second day was thus 8 dB higher than on the first day.

The Golden Ear and Trained Listener groups received copies of the instructions and listening-test tapes (the tape copies were made by a professional recording studio) and were asked to use their best headphones during their evaluations. They were further instructed to set the headphone levels for the equivalent of an 84-dB sound field while reproducing a calibration segment of 84-dB SPL speech spectrum noise. Each listener received different instructions as to the order in which he was to listen to the tapes, providing a modicum of counterbalancing in presentation order.

Instructions and fidelity rating scale. The instructions to all of the subjects were taken with minimal change from those used by Gabrielsson and Sjögren (1976). The principal instructions for the present experiments are reproduced below.

> You are about to help rate some loudspeakers, stereo headphones, and hearing aids on their ability to accurately reproduce music and speech. You will hear a series of comparisons in the form of A-B-A presentations, where the reference sound system is heard in segment A, the system under test is heard in segment B, and then the reference sound system is heard again in the final segment A. This A-B-A presentation is then repeated so that you have two chances to judge each sound system. Your task is to judge how accurately the system under test duplicates the sound of the reference system. Your judgments should be made on a 0 to 100% scale as follows: A 100% rating means you cannot hear any difference between the reference system (A) and the system under test (B). The meanings of the 90%, 70%, 50%, 30%, and 10% ratings are illustrated in the figure at left (Figure 6). The rating of 0% should be assigned if you hear practically no similarity between the two sounds; a still worse reproduction would be hard to imagine. The fact that certain numbers are given definitions does not mean that they should be used more than others. You may use any number from 0 to 100 which you think best describes the accuracy of the reproduction.

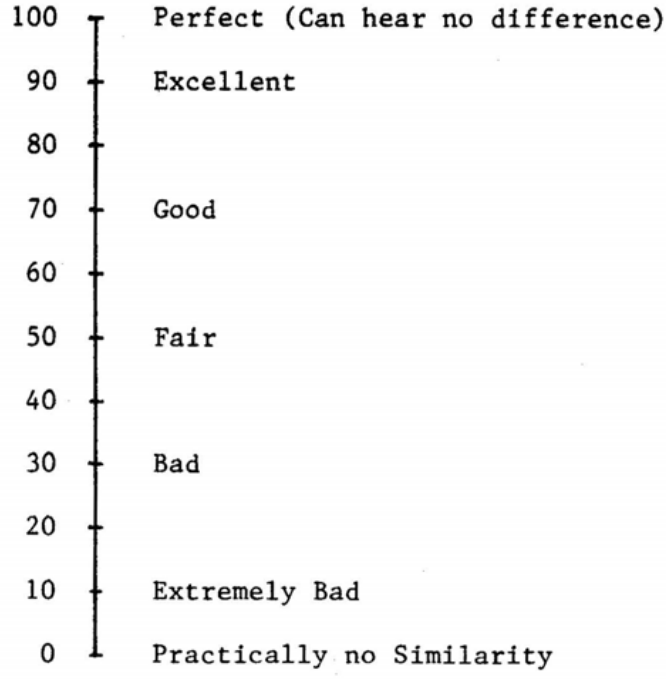

FIGURE 6. Fidelity (similarity) rating scale.

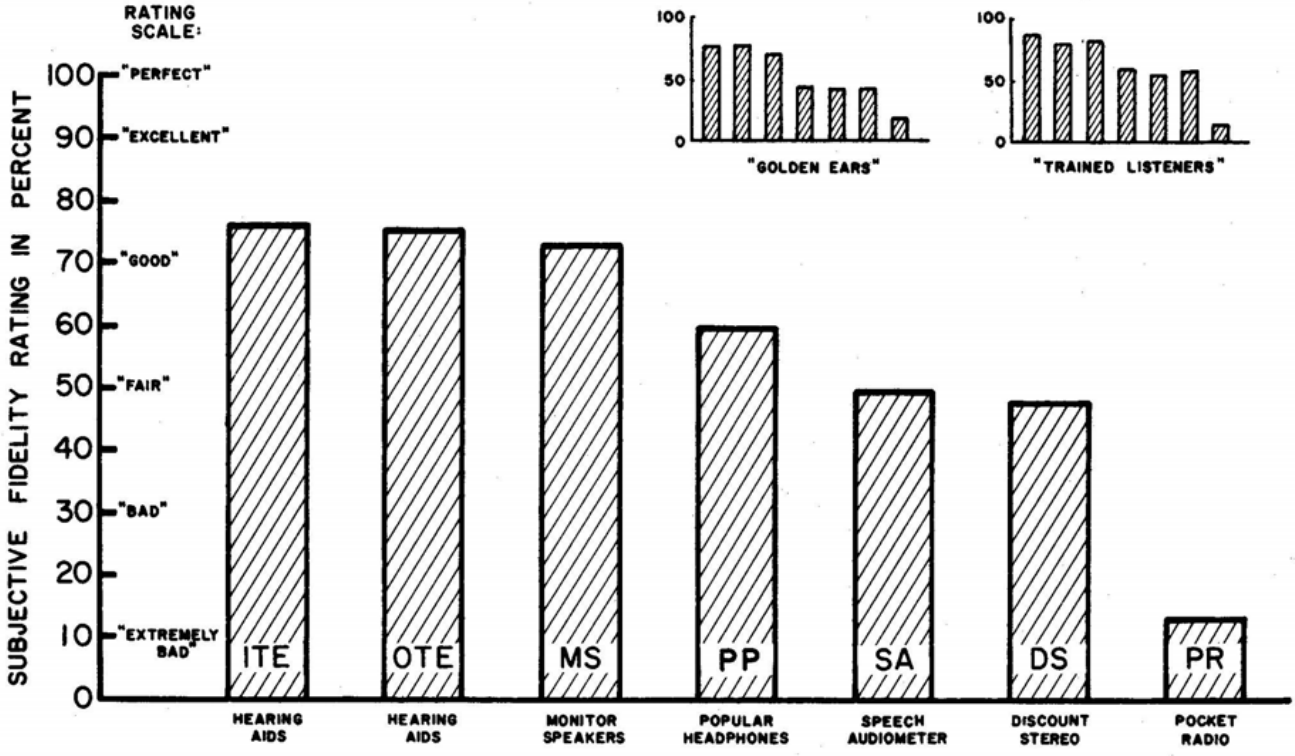

FIGURE 7. Average fidelity (similarity) ratings for six program selections from main experiment using 24 Untrained Listeners. Results from five Golden Ears and six Trained Listeners shown inset.

RESULTS AND DISCUSSION

Untrained Listener Ratings

Average fidelity ratings. The mean fidelity ratings yielded by the Untrained Listeners for the six program selections are shown in Figure 7. The values in this figure represent the combined average data for the two days' sessions, as shown in Table 1. Both the OTE and ITE hearing aids obtained higher ratings than any of the other high-fidelity systems. A three-way analysis of variance was performed on the 2016 individual-subject ratings. Application of the F test to the results indicated (a) significant differences among sound systems, program materials, and subjects; and (b) statistically significant interactions between each. Only the three-way system-program-subject interaction was not significant at the .01 level. Essentially identical results were obtained from a comparable analysis applied to arcsin-transformed data. All results in this report were obtained from the intuitively simpler, untransformed data.

The standard error of the mean ratings—based on the system-subject interaction obtained from the three-way analysis of variance—was less than 1.6%. A t test applied to the differences between the hearing aids and the other system indicated no significant difference between either of the hearing aids and the monitor speakers. All other differences between the hearing aids and the other systems were significant at well beyond the .001 level (the smallest of those differences was 10 times the standard error of the mean), based on a multiple-comparisons analysis using the Bonferroni inequality (Miller, 1966).

TABLE 1. Overall fidelity ratings for seven sound systems obtained from 24 Untrained Listeners.

	Fidelity ratings		
Sound system	*First day*	*Second day*	*Average*
ITE hearing aids	74.2	77.3	75.7
OTE hearing aids	74.7	76.0	75.4
Monitor speakers	73.1	72.3	72.7
Popular head phones	59.2	59.5	59.3
Speech audiometer	50.0	48.4	49.2
Discount stereo	47.3	46.8	47.1
Pocket radio	12.5	12.8	12.6

Regarding instructions to the subjects, these results indicate that the change in sound quality caused by interposing either pair of hearing aids in the sound path between the reference loudspeakers and the eardrum-position microphones in the KEMAR manikin was rated *comparable* to the change in sound quality caused by changing from the AR3a reference loudspeakers to a different pair of high-quality loudspeakers. The change in sound quality caused by interposing the hearing aids was judged to be significantly *less* than that caused by changing from the reference loudspeakers to (the amplifier and speakers from) a discount stereo phonograph, the popular phones, a speech audiometer, or (not surprisingly) a pocket radio.

These results may appear surprising to those familiar with the design compromises found in conventional hearing aids, although they are entirely consistent with the objective data presented under Objective Performance Measures. Recall, for example, that the calculated accuracy score for both the OTE and ITE hearing aids fell in the upper half of the range of scores obtained by inexpensive and high-priced (respectively) high-fidelity loudspeakers tested recently at Consumers Union. Given that result, it is not surprising that the hearing aids rated significantly higher than (a) the popular phones with their exaggerated bass response or (b) the simulated speech audiometer with the severe bass loss produced by the well-known cushion leak. Both defects were readily apparent in the frequency response curves in Figure 5. Although an objective measure of the frequency response of the discount stereo system was not obtained, subject comments (optional) indicated that it had a "high frequency roll-off," a "mid-frequency dip," a "hollow sound," and a "lack of bass response." (In one writer's judgment, it also had a "boomy" mid bass, and it distorted badly on the fortissimo orchestral passage.)

The fact that the OTE aids with only 8-kHz bandwidth rated as well as the ITE aids with 16-kHz bandwidth was presumably due to the comparable importance of the different defects in their frequency response. The OTE aids had a limited bandwidth but an extremely smooth insertion-gain frequency response, whereas the ITE aids had a sensibly unlimited bandwidth but a dip in response near 2.7-kHz due to their imperfect compensation for loss of normal external-ear resonances. The high rating of the OTE aids came as a surprise to us, although it was entirely consistent with Fletcher's (1942) conclusion that "substantially complete fidelity (for) . . . orchestral music is obtained (with) . . . a frequency range of from 60 to 8000 cycles per second" (p. 266). The average rating for the OTE aids on the two orchestral passages was 85%. Snow (1931) reported a value of 91%, obtained in A-B-A-B- quality-rating comparisons using orchestral music for a system with an 8-kHz upper cutoff frequency and no other defects. To the extent that the two experiments are comparable, the out-of-ear microphone location with the OTE aids appears not to have been a major defect.

Interestingly, the 73% average rating for the high-quality monitor loudspeakers (and the 75 and 76% ratings for the experimental hearing aids) almost exactly equaled the 74 and 75% equivalency of the 7.4 and 7.5 decimal ratings for two "high" fidelity loudspeakers by Gabrielsson, Rosenberg, and Sjögren (1974). They used a similar rating scale, but the subject's memory of how a live performance sounded was the reference.

Test-retest reliability. Because all comparisons were repeated on the second day of testing, it was possible to obtain an estimate of statistical reliability, as well as indication of the importance of learning, seating position (recall that the two loudspeaker channels were reversed for the second day's comparisons), and *S/N* ratio during comparison presentation.

Figure 8 graphically compares the two-days' ratings, based on the data in Table 1. The calculated correlation (Pearson's product-moment) coefficient between the two sets of mean ratings was .998. These results indicate that the ratings were relatively independent of the factors listed above.

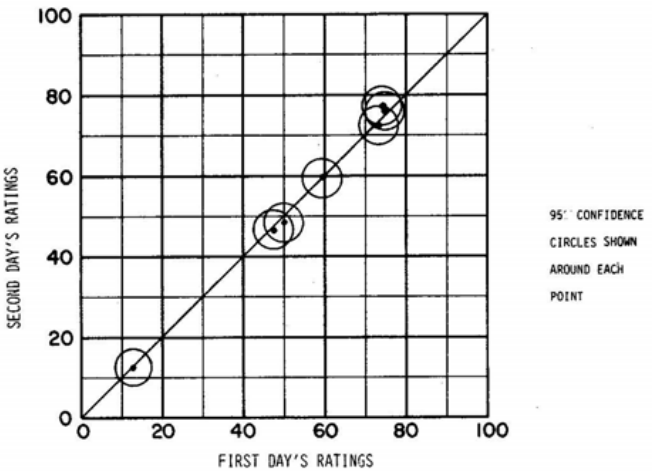

FIGURE 8. Comparison between fidelity ratings obtained from Untrained Listeners on two different days.

Indeed, additional data were obtained from nine of one writer's relatives who were imposed upon to "take the listening test" during visits to his home. These comparisons were reproduced over an old (relatively low-fidelity) hi-fi, at levels estimated to range between 5 and 15 dB below those used in this experiment. The correlation (.997) between those ratings and the average ratings from this experiment was as good as the test-retest correlation discussed above.

These results are not surprising when taking into account that the fidelity ratings obtained in our experiments were basically *similarity* ratings, as stated in the instructions to the subjects. Thus, the *constant* abberations in sound quality introduced by any reasonable sound reproduction system might be expected to have little effect on a subject's ability to detect *changes* in sound quality between two segments of a prerecorded comparison.

Trained Subject Ratings

The average ratings obtained from the five Golden Ear subjects and the six Trained Listener subjects are given in Table 2, as well as inset for comparison in Figure 7. Note that the average ratings are qualitatively quite similar across subject groups.

An analysis of variance applied to the Golden Ear and Trained Listener data produced the same conclusions as stated above for the Untrained Listener data, with the following exceptions and observations:

1. The error variance (estimated from the three-way interaction between systems, programs, and subjects) for both the Golden Ear and Trained Listener subjects was nearly four times

TABLE 2. Overall fidelity ratings for seven sound systems obtained from three subject groups.

Sound system	Subject group: *Untrained Listeners (n=24)*	*Golden Ears (n=5)*	*Trained Listeners (n=6)*
ITE hearing aids	75.7	74.6	86.2
OTE hearing aids	75.4	75.9	77.9
Monitor-speakers	72.7	67.5	79.9
Popular phones	59.3	43.5	57.5
Speech audiometer	49.2	42.2	54.5
Discount stereo	47.1	42.5	56.6
Pocket radio	12.6	17.7	14.5
Average	56.0	52.0	61.0

smaller than that for the Untrained Listener subjects. Not surprisingly, highly trained listeners are much more consistent in making subjective judgments than are untrained listeners.

2. The system-subject interaction was not significant for the Trained Listener subjects, indicating a high degree of homogeneity in that group. All were known to have spent an appreciable amount of time listening to and/or performing music (an observation which may or may not be relevant).
3. The standard error of the mean, estimated from the variance due to system-subject interaction, was 2.5% for the five Golden Ears subjects and 1.6% for the six Trained Listener subjects. The greatly reduced variance exhibited by the two trained-subject groups meant that the reliability of their single-session average ratings was comparable to that obtained from two sessions with the much larger (n = 24) group of Untrained Listeners. Thus, in those instances where population sampling is not a major concern, one trained subject appears to be worth as many as eight untrained subjects. This hardly surprising result is qualitatively similar to that obtained by Gabrielsson and Sjögren (1976).
4. The variance due to system-program interaction was almost 20 times smaller for the Golden Ear subjects and nearly 10 times smaller for the Trained Listener subjects than for the Untrained Listener subjects. Successful Golden Ear professionals presumably have found it useful to train themselves to "listen through" the particular musical selection used for system evaluation. Although the program selections were considered as "fixed effects" in the statistical analysis of these experiments, the calculated reliability of the trained-subject ratings would have suffered relatively little if the program selections had been considered a random sample. In other words, essentially similar ratings might be expected from trained subjects using any reasonable cross section of program material.

Comparison of Untrained Listener and Trained Listener Results

The application of Welch's t-test approximations (Brownlee, 1965, p. 299) to the differences between the overall average ratings obtained from the Untrained Listener, Golden Ear, and Trained Listener subjects indicated the differences were not significant at the .05 level.

Application of the Bonferroni inequality and t statistics (Miller, 1966) to obtain confidence intervals for the seven individual system ratings from each group, however, indicated that some of the between-group differences in individual system ratings were significant. The most striking was the roughly 15% lower rating given the popular phones by the Golden Ear subjects compared to the two other subject groups. This seemed reasonable in light of the comment of one Untrained Listener subject, who ignored instructions and gave the popular phones a 100% rating because he "liked them much better" (than the reference). In particular, anecdotal market evidence indicates that those who have not spent much time professionally evaluating high-fidelity systems are much more tolerant of an excessive bass response than of a deficient bass response.

The correlations between the Golden Ear and Untrained Listener ratings (r = .956) and the Trained Listener and Untrained Listener ratings (r = .984) were both high, further evidencing the stability—under different listening conditions and subject selections—of the relative ratings produced in the present experimental design. The good correlation between Trained and Untrained Listener ratings is consistent with the findings of Gabrielsson, Rosenberg, and Sjögren (1974) and Gabrielsson and Sjögren (1976).

In comparing the high correlation coefficient (the Pearson product-moment correlation coefficient r has been used throughout) to the obvious differences among ratings from the different subject groups, recall that the correlation indicates the degree to which the least-squares, best-fit linear relationship ($y = mx + b$) accounts for the dependent-variable data. After accounting for the differences between Untrained Listener and Trained Listener ratings (by applying the optimum linear transformation from one to the other), for example, all but $1-(.984)^2 = .03$ (3%) of the variance is accounted for. Simply stated, the two groups appear to measure essentially the same thing using slightly different subjective scales.

CONCLUSIONS

The most important conclusion of this study is that current hearing aid amplifier and transducer technology does, in fact, permit the construction of practical high-fidelity hearing aids as judged by someone with normal hearing. Not surprisingly perhaps, at least one high-fidelity hearing aid design became commercially available shortly after this study was undertaken (Toepholm, 1979).

At least three reasons exist for demonstrating that it is possible to design a hearing aid which is judged high-fidelity by someone with *normal* hearing:

1. Such a design provides a base to which electronic signal processing can conveniently be added.
2. A hearing aid which provides gain only for low-level signals (i.e., is a unity-gain, high-fidelity, sound-reproduction system for high-level signals) may prove useful to a large number of individuals.
3. The demonstration supports the following conclusion: that the important question for hearing aid research is no longer "What *can* a hearing aid be designed to do?" but "What *should* a hearing aid be designed to do for the hearing impaired?"

The lack of a satisfactory answer to this latter question is a major barrier to vastly improved hearing aid design. That question can be restated: What hearing aid characteristics will prove to be optimum (or even somewhere near optimum) for a given individual as he goes about his daily life? More specifically: Will a substantial number of hearing aid users with mild-to-moderate hearing impairments prefer a high-fidelity hearing aid (as defined earlier under Experimental Hearing Aids) to a more conventional hearing aid? The answer to this question is crucial, because the main goal of hearing aid use—improvement in communicative capacity—can only be achieved if the individual actually wears the hearing aid. There is much anecdotal evidence suggesting that many people who could benefit from a hearing aid refuse to wear it in many or most circumstances if it doesn't sound pleasing to them.

In a recent study by Pascoe (1975), his hearing-impaired subjects understood speech as well under his "uniform amplification" condition as they did at comparable intensity levels in the sound-field condition. The frequency response of Pascoe's aid in the uniform amplification condition was very similar to that of the high-fidelity aids in this experiment.

Thus, there is indirect evidence that a hearing aid judged to be *high fidelity* by normal listeners will allow hearing-impaired listeners to understand speech as well as they could in the (high-intensity) sound-field condition. Pascoe demonstrated that this latter level of understanding could be further improved via frequency selective amplification. Thus, the next question to be answered is "Will the sound produced via frequency selective amplification be sufficiently pleasant to entice the individual to wear the aid for long periods of time?" If the answer to this question is "No," then a wide-band hearing aid that is judged to be *high fidelity* by normal listeners may represent the best compromise for the hearing-impaired user. As Barfod (1979) discussed, such a hearing aid could be "especially suited for hearing impaired subjects having difficulty in adapting to a new speech code" (p. 436).

Preliminary answers to the questions posed here could be reached in laboratory experiments such as the fidelity-rating experiment we described earlier, but we suspect that the final answers can only be obtained through the fairly clumsy process of trial and error in the marketplace, as dispensers discover which new hearing aid designs provide increased user satisfaction.

ACKNOWLEDGMENTS

This paper contains material presented at the 96th Meeting of the Acoustical Society of America, Honolulu, November 1978. This study benefited greatly from the insights and suggestions of Elmer Carlson, Hugh Knowles, Mahlon Burkhard, Richard Peters, and Edgar Villchur, whose contributions are hereby gratefully acknowledged.

REFERENCE NOTES

1. KNOWLES, H. S. *Physical aspects of hearing aids.* Unpublished paper presented at Allerton House, University of Illinois, 1959.
2. KNOWLES ELECTRONICS. *Receiver application* (Bulletin TB-20). Available from Knowles Electronics, Inc., 3100 N. Mannheim Road, Franklin Park, IL 60131, 1980.

REFERENCES

AMERICAN NATIONAL STANDARDS INSTITUTE (ANSI). *Specifications for audiometers* (S3.6-1969). New York: ANSI, 1969.

BARFOD, J. Speech perception processes and fitting of hearing aids. *Audiology,* 1979, *18,* 430-441.

BROWNLEE, R. G. *Statistical theory and methodology in science.* New York: Wiley, 1965.

BURKHARD, M. D. Gain terminology. In M. D. Burkhard (Ed.), *Manikin measurements—Conference proceedings* (Chap. 4). Elk Grove Village, IL: Industrial Research Products, 1978.

BURKHARD, M. D., & SACHS, R. M. Anthropometric manikin for acoustic research. *Journal of the Acoustical Society of America,* 1975, *58,* 214-222.

CARLSON, E. V., & MOSTARDO, A. F. *Damping element.* US Patent Office (Patent No. 3,930,560), 1976. (Filed July 1974.)

CARLSON, E. V., MOSTARDO, A. F., & DIBLICK, A. V. *Transducer with improved armature and yoke construction.* US Patent Office (Patent No. 3,935,398), 1976. (Filed July 1971.)

DALSGAARD, S. C., & JENSEN, O. D. Measurements of insertion gain of hearing aids. *Eighth International Congress on Acoustics* (Vol. 1, p. 205). London, 1974.

DAVIS, M. What's really important in loudspeaker performance? *High Fidelity,* June 1978, pp. 53-58.

FLETCHER, H. Hearing, the determining factor for high-fidelity transmission. *Proceedings of the Institute of Radio Engineers,* 1942, *30,* 266-277.

GABRIELSSON, A., NYBERG, P. O., SJÖGREN, H., & SVENSSON, L. *Detection of amplitude distortion by normal hearing and hearing impaired subjects* (Report TA No. 83). Stockholm: Karolinska Institutet, Technical Audiology, 1976.

GABRIELSSON, A., ROSENBERG, U., & SJÖGREN, H. Judgments and dimension analysis of perceived sound quality of sound-reproducing systems. *Journal of the Acoustical Society of America,* 1974, *55,* 854-861.

GABRIELSSON, A., & SJÖGREN, H. *Preferred listening levels and perceived sound quality at different sound levels in "high fidelity" sound reproduction* (Report TA No. 82). Stockholm: Karolinska Institutet, Technical Audiology, 1976, pp. 1-33 plus appendices.

High-priced loudspeakers. *Consumer Reports,* 1978, *43,* 592-599.

How CU's audio lab tests loudspeaker accuracy. *Consumer Reports* (TNG-3), 1977.

KILLION, M. C. *Design and evaluation of high-fidelity hearing aids.* Doctoral dissertation, Northwestern University, 1979. (a) (University Microfilms No. 7917816)

KILLION, M. C. Equalization filter for eardrum-pressure recording using a KEMAR manikin. *Journal of the Audio Engineering Society,* 1979, *27,* 13-16. (b)

KILLION, M. C. Earmold options for wideband hearing aids. *Journal of Speech and Hearing Disorders,* 1981, *46,* 10-20.

KILLION, M. C., & CARLSON, E. V. A subminiature electret-condenser microphone of new design. *Journal of the Audio Engineering Society,* 1974, *22,* 237-243.

KILLION, M. C., & MONSER, E. L. CORFIG: Coupler response for flat insertion gain. In G. A. Studebaker & I. Hochberg (Eds.), *Acoustical factors affecting hearing aid performance.* Baltimore: University Park Press, 1980.

KNOWLES, H. S. *Some considerations for hearing aid frequency response curves.* Paper presented at the 44th Annual Conven-

tion of the American Speech and Hearing Association, Denver, 1968.

KNOWLES, H. S., & KILLION, M. C. Frequency characteristics of recent broadband receivers. *Journal of Audio Technology* (Zeitschrift für Hörgeräte-Akustik), 1978, *17*, 86-89; 136-140.

Low-priced loudspeakers. *Consumer Reports*, 1977, *42*, 406-409.

MARTIN, D. W., & ANDERSON, L. J. Headphone measurements and their interpretation. *Journal of the Acoustical Society of America*, 1947, *19*, 63-70.

MILLER, R. G. *Simultaneous statistical inference*. New York: McGraw-Hill, 1966.

MILNER, P. How much distortion can you hear? *Stereo Review*, June 1977, pp. 64-68.

OLSON, H. F. *Acoustical engineering*. New York: Van Nostrand, 1957.

PASCOE, D. P. Frequency response of hearing aids and their effects on the speech perception of hearing-impaired subjects. *Annals of Otology, Rhinology, and Laryngology*, 1975, *84*(Suppl. 23).

SHAW, E. A. G. Earcanal pressure generated by circumaural and supra-aural earphones. *Journal of the Acoustical Society of America*, 1966, 39, 471-479.

SHAW, E. A. G. Acoustics of the external ear. In G. A. Studebaker & I. Hochberg (Eds.), *Acoustical factors affecting hearing aid performance*. Baltimore: University Park Press, 1980.

SNOW, W. B. Audible frequency range of music, speech, and noise. *Journal of the Acoustical Society of America*, 1931, 3, 155-166.

TOEPHOLM, C. Increasing demands on hearing aids and the ETYMOTIC GAIN. *Hearing Aid Journal*, January 1979, *32*(3), 8; 38-39.

VILLCHUR, E. A method of testing loudspeakers with random noise input. *Journal of the Audio Engineering Society*, 1962, *10*, 306-309.

VILLCHUR, E. Technique of making live versus recorded comparisons. *Audio*, October 1964, pp. 34-42; 120.

Received November 15, 1979
Accepted January 5, 1981

Requests for reprints and/or a 33-⅓ rpm stereo Soundsheet recording containing samples of the A-B-A comparisons used in this study should be addressed to Mead C. Killion, Industrial Research Products, Inc., 321 Bond Street, Elk Grove Village, IL 60007.

Reprinted from *Journal of Speech and Hearing Research*
March 1982, Vol. 25, No. 1