

Robert H. Margolis<sup>\*,§</sup>  
George L. Saly<sup>§</sup>

<sup>\*</sup>Department of Otolaryngology,  
University of Minnesota, USA  
<sup>§</sup>Audiology Incorporated, Arden Hills,  
Minnesota, USA

## Key Words

Hearing  
Hearing loss  
Audiogram  
Audiometry  
Audiometer  
AMCLASS™

## Abbreviations

AMCLASS™: Audiogram  
classification system  
HL: Hearing level  
dB: Decibel  
Hz: Hertz

# Toward a standard description of hearing loss

## Abstract

Hearing losses are frequently described by categories that characterize the configuration, severity, and site of lesion from a pure-tone audiogram. Although many category descriptors are in common use, there are no standard definitions of those terms, nor have the category definitions been validated against current clinical practice. The development and validation of AMCLASS™ is described. To validate the classification method, five expert judges selected configuration, severity, and site of lesion categories for 231 audiograms that varied widely in audiometric configuration. Interjudge comparisons indicated that expert judges frequently disagree on how they describe an audiogram. Category definitions were adjusted to maximize agreement between AMCLASS™ and the consensus of the judges. The final set of category definitions produced categories that agreed with the consensus more often than the average agreement between pairs of judges.

## Sumario

Las hipoacusias son frecuentemente descritas a partir de un audiograma, caracterizado por su configuración, severidad y sitio de lesión. Aun cuando hay varias categorías descriptivas de uso común, no existen definiciones estándar de tales términos, ni se han validado las definiciones de las categorías con respecto a la práctica clínica actual. Se describe el desarrollo y la validación de AMCLASS™. Para validar el método de clasificación, cinco jueces expertos seleccionaron las categorías de configuración, severidad y sitio de lesión de 231 audiogramas que variaban ampliamente en su configuración audiométrica. Las comparaciones entre jueces indicaron que ellos frecuentemente discrepaban en la descripción del audiograma. Se modificaron las definiciones de las categorías para maximizar el acuerdo entre AMCLASS™ y el consenso de los jueces. El conjunto final de las definiciones de las categorías produjo categorías que concordaban con el consenso más a menudo que el consenso promedio entre pares de jueces.

There is nothing more fundamental to the role of the audiologist than evaluating hearing, determining the nature of a hearing loss, and communicating that determination to the patient and other professionals. The results of the study reported here indicate that expert audiologists vary widely in how they describe a given hearing loss. A solution is provided in the form of a validated, automated classification scheme that can form the basis for a standardized description of the pure-tone audiogram.

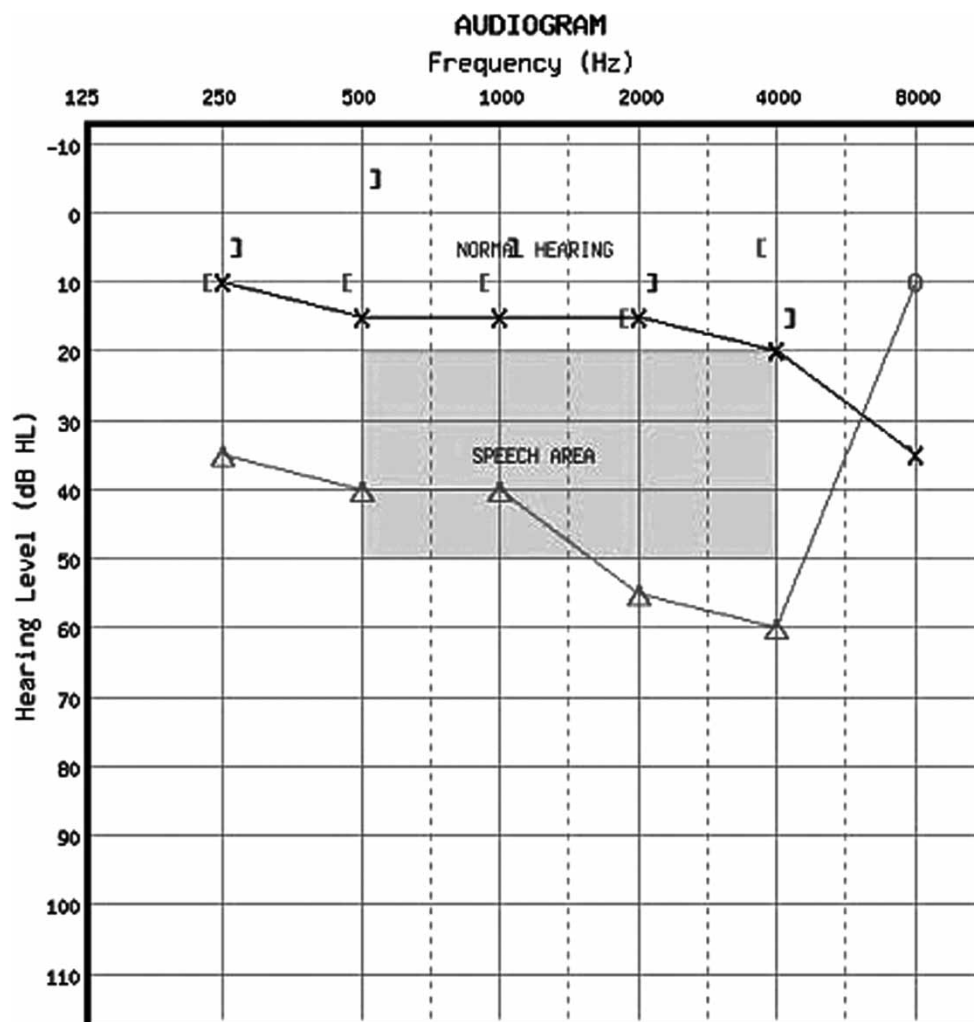
With the introduction of electric pure-tone audiometers in Europe and North America in the 1920s, and with their increasing popularity for clinical testing, it was quickly recognized that the pure-tone audiogram can take a dizzying variety of forms that are difficult to categorize by a concise and practical classification system. Guild (1932) and Carhart (1945), using the technology of their times, developed systems that classified audiograms according to configuration, severity, and interaural asymmetry. Because bone-conduction testing was not yet routinely performed, classification by site of lesion was not part of their schemes but would be added later. Guild used a punch-card system for sorting audiograms into groups, and Carhart used transparencies to match categories with individual audiograms. Carhart validated his method with experienced judges and modified the classification scheme to achieve the best agreement with the judges. Both systems could be implemented today with software, which could render them useful for their intended purpose, but because of their numerous categories,

subcategories, labels, subscripts, and superscripts, these systems are somewhat unwieldy for clinical application.

The purpose of the Guild and Carhart methods was to group test results into categories to facilitate the study of relationships between audiogram characteristics and ear disease, and to compare various clinical populations. Several studies of this type have been conducted to describe, for example, audiometric findings in Meniere's Disease (Lee et al, 1995; Paparella et al, 1982; Savastano et al, 2006), and acoustic neuroma (Neary et al, 1996). Pittman and Stelmachowicz (2003) compared audiograms of children and adults with sensorineural hearing loss based on audiometric configuration, asymmetry, and progression. They reported differences in the distributions of configurations for children and adults, with sloping and trough-shaped losses more prevalent in adults. In addition, children were more likely to have asymmetrical hearing loss.

The problem is a little more complex than perhaps even Guild and Carhart realized. The wide dynamic range of hearing, frequency selectivity of hearing loss, behavioral variability, and measurement error conspire to create an enormous number of possible configurations. We calculated the number of possible audiogram configurations for six air-conduction frequencies and five bone conduction frequencies with the following constraints.

1. Air-conduction thresholds can take any value between –10 and 110 dB HL (except at 250 Hz, where the upper limit is 90 dB HL);



**Figure 1.** The configuration for the right ear was described by five expert judges as flat, sloping, rising, trough, and 'other'. (X = left ear unmasked air conduction; Δ = right ear masked air conduction; O = right ear unmasked air conduction; [ = right ear masked bone conduction; ] = left ear masked bone conduction).

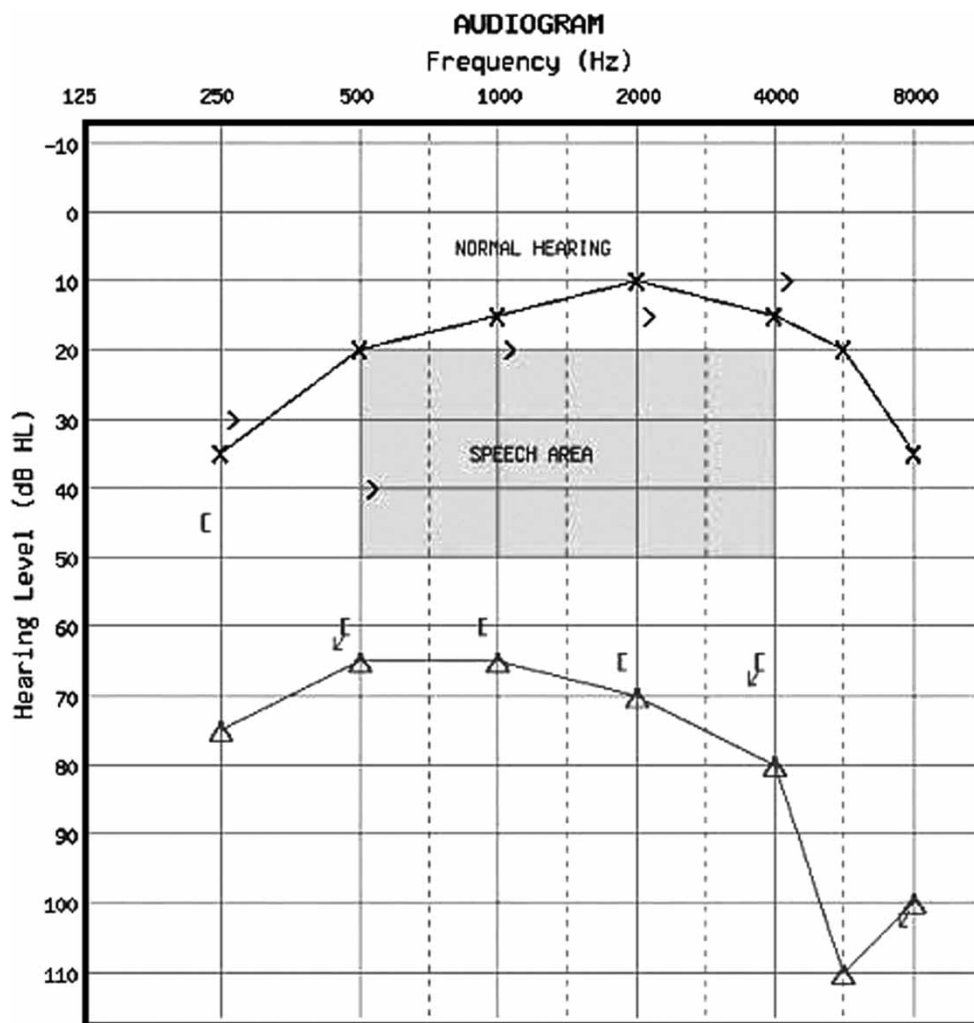
2. An air-conduction threshold must be within 30 dB of the threshold at the next lowest frequency;
3. Bone-conduction thresholds can take any value between -10 and 60 dB HL except at 250 Hz where the upper limit is 40 dB HL;
4. A bone-conduction threshold must be between -50 and 10 dB relative to the air-conduction threshold at that frequency.

Constraints 2 and 4 are intended to eliminate slopes that are physiologically unlikely. With these constraints there are more than 376 billion possible air- and bone audiograms for a single ear. For air-conduction-only audiograms there are 3.62 million possibilities. In the validation study reported here with real audiograms obtained from a hospital-based audiology clinic, we were frequently amazed at the number of audiograms that were difficult to categorize. Figure 1 illustrates the difficulty. This real audiogram was obtained anonymously from the records of the University of Minnesota Hospital Audiology Clinic. The right ear configuration was categorized by five expert judges with five different descriptions of the hearing loss: flat, sloping, rising,

trough, and 'other'. All are reasonable descriptions. Carhart (1945) dealt with the situation of multiple possible categories by using the one that is closest to 'flat'.

One approach to reducing the large number of possible audiograms to a manageable characterization of a hearing loss is to calculate a percent impairment. Methods of this type have been proposed by Fowler and Wegel (1922), Fletcher (1929), the American Medical Association (AMA Council on Physical Therapy, 1942), the American Academy of Ophthalmology and Otolaryngology (Davis, 1965), the Department of Health and Social Security in the United Kingdom (Tempest, 1976), and the National Acoustics Laboratories in Australia (Macrae, 1975). These methods are useful for several purposes such as medicolegal matters and compensation decisions. But percentage scales do not retain important characteristics that have audiologic and medical importance. For example, audiometric configuration, which is not conveyed in a percentage calculation has important implications for diagnosis and hearing-aid fitting.

There has been little further development of categorical methods since the early attempts of Guild and Carhart, but



**Figure 2.** The audiogram in the right ear could be ‘flat’ or ‘sloping’. (X = left ear unmasked air conduction; Δ = right ear masked air conduction; [] = right ear masked bone conduction; Δ = left ear masked bone conduction).

textbooks and research articles continue to indicate an interest in methods of this type. Some textbooks offer a classification based on severity, with categories such as normal, slight (minimal), mild, moderate, severe, and profound (Martin, 1986; Bess & Humes, 1990; Kaplan et al, 1993; Yantis, 1994; Keith, 1996; Stach, 1998). Some also provide a classification of audiogram configuration with categories like flat, sloping (falling), rising, trough (scoop), and ‘miscellaneous’ (Kaplan et al, 1993; Stach, 1998; Roeser et al, 2000). Kaplan et al (1993) point out that audiograms additionally can be classified by site of lesion, but we are not aware of a formula that has been proposed for that purpose. In general, classification systems provide general rules for placing audiograms in categories, but do not deal with the practical issue of assigning a category when there are local irregularities that experienced judges learn to ignore. Nor do they provide rules when there is no response at a particular frequency at the limit of the audiometer, or when an audiogram can be reasonably described by more than one configuration.

Clark (1981) pointed out that because of a lack of standardized, rigorous definitions, it is common for two audiologists to describe the same audiogram differently. We found that the rule set necessary to categorize audiograms in a manner that is consistent with expert judges is extraordinarily complex. Programming the resulting rule set in software provides a practical solution to an otherwise unwieldy task.

Figure 2 provides an example. The audiogram for the right ear was categorized by three judges as ‘flat’ and by two judges as ‘sloping’. By the Stach (1998) system it would be ‘sloping’ (‘thresholds for high frequencies are at least 20 dB poorer than for low frequencies’, p. 107). Apparently three judges were willing to overlook the thresholds at 6000 and 8000 Hz in favor of the more important 250–4000 Hz range. Either judgment is defensible, but the case illustrates that judges do not weight all frequencies equally and they are not consistent in the weights they ascribe to a particular frequency for a particular audiogram. The classification system described in this report,

AMCLASS™, was devised to offer a standard, although complex, set of rules for maximizing the likelihood that the selected configuration agrees with expert judges.

The site of lesion determination is also not as simple as it might seem. While there is general agreement on the definitions of conductive, sensorineural, or mixed, implementing those definitions requires more complex rules than simply detecting the presence or absence of air-bone gap. The bone-conduction threshold at 250 Hz in the right ear of the audiogram shown in Figure 2 would probably be interpreted by most clinicians as a vibrotactile response that does not reflect a true conductive component. The 'gap' at 4000 Hz is the result of the lower output limit for bone conduction than for air conduction. Most expert judges would characterize the hearing loss as sensorineural. But a mixed hearing loss cannot be completely ruled out. The left-ear bone-conduction threshold at 500 Hz presents another type of challenge that any practical classification scheme has to deal with.

The purpose of this project was fourfold. First, like Guild and Carhart, we sought to provide a classification scheme that would be useful for grouping audiograms to facilitate studies of relationships between audiogram configurations and ear disease. Second, we would like the method to produce a concise verbal description of the hearing loss to facilitate communication among professionals, between clinicians and patients, and between teachers and students. Third, after the method of Carhart, we sought to validate the method against the opinions of expert judges. And fourth, we wished to develop a software implementation of the validated classification system that could be incorporated into audiometer software or serve as a stand-alone or web-based tool.

### AMCLASS™: Automated classification of audiograms

Categories were based on the previous work cited above. The classification system provides a unique category for any audiogram that includes air-conduction thresholds at six octave frequencies (250–8000 Hz) and inter-octave frequencies if available; and bone-conduction thresholds at four octave frequencies (500–4000 Hz). Bone-conduction thresholds at 250 Hz were excluded from the analysis because of the contaminating influence of vibrotactile responses and the typically low audiometric maximum output at that frequency. Threshold measurements may be defined at a specific level or may be 'no response' at the output limit of the audiometer. The classification includes a configuration, severity, and site of lesion. If only air-conduction thresholds are available, a configuration and severity is determined. If results for both ears are available, a symmetry category is provided. The classification scheme is summarized in Table 1. Although the details of the calculations are proprietary, we provide the general principles used to produce the categorizations. The algorithms are designed to ignore local irregularities that expert judges overlook. AMCLASS™ consists of a total of 161 rules including 23 for configuration, 45 for severity, 56 for site of lesion, and 37 for symmetry.

#### Configuration

In selecting the number of configuration categories, it is desirable to capture a large majority of patterns seen in clinical popula-

tions without resorting to an unwieldy number, which, of course, is a subjective judgment. We chose seven configuration categories, including an *Other* category for cases that do not fit the criteria of the other categories.

1. *Normal* is generally thresholds better than or equal to 20 dB, with some local deviations allowed.
2. *Flat* is a hearing loss where all thresholds are generally within a 20-dB range.
3. *Sloping* is a hearing loss that has a generally-downward trend of any slope. A sloping hearing loss may be flat over a portion of the frequency range or even rising if the general trend is downward.
4. *Rising* is similar to *Sloping* but in the reverse direction.
5. *Trough* is a hearing loss that is most severe in the middle frequencies. The mid-frequency dip must be a clear trend and not a local deviation.
6. *Peaked* is similar to trough but with best hearing in the middle frequencies.
7. *Other* is a category for audiograms that are not consistently placed in the other categories by a panel of expert judges, or when there is no consensus.

#### SEVERITY

Severity categories are typical of those that have been suggested in the literature with the exception that a 'slight' or 'minimal' category was not used because of the difficulty of distinguishing it from surrounding categories. In *Sloping* and *Rising* configurations, two severities are determined, one for the low frequency region and one for the high frequency region. The boundaries between categories were initially selected and then adjusted to maximize agreement with the judgments of the expert panel. In general, the boundaries are

1. Mild: >20 and ≤40 dB HL
2. Moderate: >40 and ≤60 dB HL
3. Severe: >60 and ≤90 dB HL
4. Profound: >90 dB HL

The effects of local irregularities are avoided by a system of overlapping averages in adjacent frequency regions. This produces better agreement with judges, but creates more complex boundaries between categories.

#### SITE OF LESION

Site of lesion categories are based on the presence of air-bone gaps at octave frequencies between 500 and 4000 Hz. Note that the audiology literature is inconsistent in the use of terminology related to these categories. Some authors prefer the term 'type' of hearing loss. Roeser (2000) for example uses 'type' and relates each type to 'anatomic site of involvement' (Figure 11–7, p. 237). Similarly, Kaplan et al (1993) describes 'types of organic hearing impairment' and then relates each type to the 'site of lesion' (Figure 1–9, p. 12). However, Stach (1998) uses 'type' to distinguish between hearing impairments characterized by 'hearing sensitivity loss', and 'auditory nervous system disorders'. Conductive, sensorineural, and mixed are classified under the hearing sensitivity loss type with no designator of their own. Bess and Humes (1990) use 'site of lesion' similarly to the usage in this manuscript. On the other hand, some authors use 'site of

**Table 1.** AMCLASS™



**AMCLASS™ - AUDIOGRAM CLASSIFICATION SYSTEM**

Configuration	Severity	Site of Lesion	Symmetry
Normal Hearing	Mild	Conductive	Symmetrical Hearing Loss
Flat Hearing Loss	Moderate	Sensorineural	Asymmetrical Hearing Loss
	Severe	Mixed	
	Profound	Sensorineural or Mixed	
Sloping Hearing Loss	Normal-Mild		
	Normal-Moderate		
	Normal-Severe		
	Mild-Moderate		
	Mild-Severe		
	Moderate-Severe		
	Severe-Profound		
	Profound		
Rising Hearing Loss	Mild-Normal		
	Moderate-Normal		
	Moderate-Mild		
	Severe-Normal		
	Severe-Mild		
	Severe-Moderate		
	Profound-Severe		
	Profound		
Trough-shaped Hearing Loss	Mild		
	Moderate		
	Severe		
Peaked Hearing Loss	Mild		
	Moderate		
	Severe		
Other	Mild		
	Moderate		
	Severe		

lesion' in reference to tests for distinguishing cochlear from retrocochlear hearing-loss (Kaplan et al, 1993, p. 158–159; Lonsbury-Martin et al, 1999, p. 181). We prefer 'site of lesion' to 'type' because it communicates the basis for the classification. Clearly there is a need for standardized usage of these terms.

The size of the air-bone gap that indicates a conductive component is dependent on the number of frequencies at which air-bone gaps occur. In general, a 10-dB air-bone gap at three or more frequencies, or a 15-dB air-bone gap at any one frequency, indicates a conductive component. A hearing loss is judged to be sensorineural if the configuration is not 'normal' and there is no significant air-bone gap. The 'Sensorineural or Mixed' site of lesion was included to describe cases where there is no response by bone conduction at the audiometer limit. In these cases it is





not possible to rule out a conductive component. It is possible for the configuration to be normal, and the site of lesion to be conductive if air-conduction thresholds are judged to be in the normal range and there is an air-bone gap.

**SYMMETRY**





Symmetry categories (symmetrical and asymmetrical) are based on interaural air-conduction threshold differences for octave frequencies (250–8000 Hz). Audiograms were judged to be asymmetrical if there were three or more interaural differences of 10 dB or more, two interaural differences of 15 dB or more, or one interaural difference of 20 dB or more. Validation data were not obtained in this study for symmetry because only single audiograms were presented to the judges so that selection of







**Table 2.** Interjudge agreement is shown for pairs of judges and for all judges combined (ALL). Values are given in percent, and as Kappa statistics (in parentheses). The CONSENSUS row shows agreement between each judge and the consensus category for all judges. ALL refers to the rate of agreement across all judges (all judges chose the same category). MEAN OF PAIRS is the average agreement between pairs of judges. MEAN CONSENSUS is the average of the CONSENSUS row.

Judge	Configuration				
	1	2	3	4	5
1		65.4 (0.60)	58.0(0.51)	62.8 (0.57)	62.8 (0.57)
2			75.3(0.71)	68.4 (0.63)	78.8 (0.75)
3				62.8 (0.57)	74.5 (0.70)
4					66.7 (0.61)
CONSENSUS	73.2 (0.69)	92.6 (0.91)	82.3 (0.79)	78.4 (0.75)	89.2 (0.87)
ALL	43.0 (0.43)				
MEAN OF PAIRS	67.6 (0.62)				
MEAN CONSENSUS	83.1 (0.80)				

Judge	Severity				
	1	2	3	4	5
1		80.5 (0.76)	78.8 (0.73)	84.8 (0.81)	75.8 (0.70)
2			83.5 (0.79)	86.6 (0.83)	84.4 (0.81)
3				82.3 (0.78)	87.0 (0.84)
4					82.3 (0.78)
CONSENSUS	86.6 (0.83)	92.2 (0.90)	87.5 (0.84)	88.7 (0.86)	86.2 (0.83)
ALL	61.0(0.61)				
MEAN OF PAIRS	82.6 (0.78)				
MEAN CONSENSUS	88.2 (0.85)				

Judge	Site of Lesion				
	1	2	3	4	5
1		69.7 (0.62)	60.6(0.51)	65.8 (0.57)	68.0 (0.60)
2			81.0(0.76)	79.7 (0.75)	87.4 (0.84)
3				73.6 (0.67)	80.1 (0.75)
4					77.9 (0.72)
CONSENSUS	73.2 (0.64)	94.8 (0.93)	84.9 (0.80)	84.4 (0.81)	93.1 (0.91)
ALL	49.8 (0.33)				
MEAN OF PAIRS	74.4 (0.68)				
MEAN CONSENSUS	86.1 (0.82)				

configuration, severity, and site of lesion categories were not biased by knowledge of the other ear. An informal evaluation of the symmetry decisions of AMCLASS<sup>TM</sup> was performed by one judge (the first author) who viewed 200 paired audiograms. There was excellent agreement between AMCLASS<sup>TM</sup> and that judge. A validation study for asymmetry has been completed and will be reported separately.

#### AMCLASS<sup>TM</sup>: Development and validation

An initial set of rules was developed by the first author to define the categories shown in Table 1, and encoded in a software program by the second author. AMCLASS<sup>TM</sup> outcomes were determined for a library of 3686 audiograms mined (with IRB approval) from the digital database of the University of Minnesota Hospital Audiology Clinic. The only criteria for inclusion were that thresholds (including 'No response' indications) must be present for both ears for air conduction at octave frequencies over the 250–8000 Hz range (with or without

interoctave thresholds), and for bone conduction at octave frequencies over the range 500–4000 Hz. AMCLASS<sup>TM</sup> determined a configuration, severity, and site of lesion for each ear (n = 7372) and a symmetry category (symmetrical or asymmetrical) for each case (n = 3686).

This began an iterative process with which categories were determined and examined, and the rules were revised to achieve the outcome that was thought to be most appropriate by the first author. It immediately became clear that the original set of rules was not sufficient to categorize the large number of possible audiograms with the desired accuracy. The rules were revised many times until the agreement between AMCLASS<sup>TM</sup> and the judge was thought to be satisfactory, and an interim set of rules was established.

To validate the interim rules, four additional judges were recruited. Judges were selected with experience in a variety of employment settings. Three were Ph.D. audiologists who have worked in academic health centers, private practices, ENT offices,

**Table 3.** Agreement pairs. Severity was considered to be in agreement for these pairs of judgments.

<i>Agreement Pairs</i>	
normal	normal
mild	mild
moderate	moderate
severe	severe
profound	profound
normal-mild	normal
normal-mild	mild
normal-moderate	normal-moderate
normal-moderate	mild
normal-severe	normal-severe
normal-severe	moderate
mild-moderate	mild-moderate
mild-moderate	mild
mild-moderate	moderate
mild-severe	mild-severe
mild-severe	moderate
moderate-severe	moderate-severe
moderate-severe	moderate
moderate-severe	severe
severe-profound	severe-profound
severe-profound	severe
severe-profound	profound
profound	profound

and industry; one is an Au.D. audiologist who has worked primarily in private practice; and one is an otologist who practices at the University of Minnesota Hospital. Each judge has at least 20 years of clinical experience. A subset of 231 single-ear audiograms (only the right or left ear) was selected that had approximately equal distribution of the categories shown in Table 1. By a web-based form, each judge viewed each audiogram and selected a configuration, severity, and site of lesion. No rules were given to the judges for categorizing the audiograms. They were told to use the definitions that they use in their clinical practices for interpreting audiograms. Only one ear was shown so that their judgments would not be biased by the other ear. Thus, symmetry categories were not selected by the judges.

Based on the results of the five judges, a consensus configuration was determined. The consensus was the category that was chosen by the largest number of judges. There were 19 cases for which there was a tie between two configuration categories. For these cases, AMCLASS™ was judged to be in agreement with the judges if either of the two categories was indicated. The

case shown in Figure 1 (right ear) was eliminated from the analysis of configuration because there was no consensus.

The initial agreement between AMCLASS™ and the consensus was 68%, 81%, and 73% for configuration, severity, and site of lesion, respectively. A second iterative process was undertaken to maximize agreement with the judges. This process continued until further adjustments did not increase agreement between AMCLASS™ and consensus. At that point all remaining cases for which there was disagreement between AMCLASS™ and consensus were judged to be acceptable disagreements. That is, in the subjective judgement of the first author, the AMCLASS™ category, although it disagreed with the consensus, was a reasonable description of the audiogram.

## Results and Discussion

### *Interjudge Agreement*

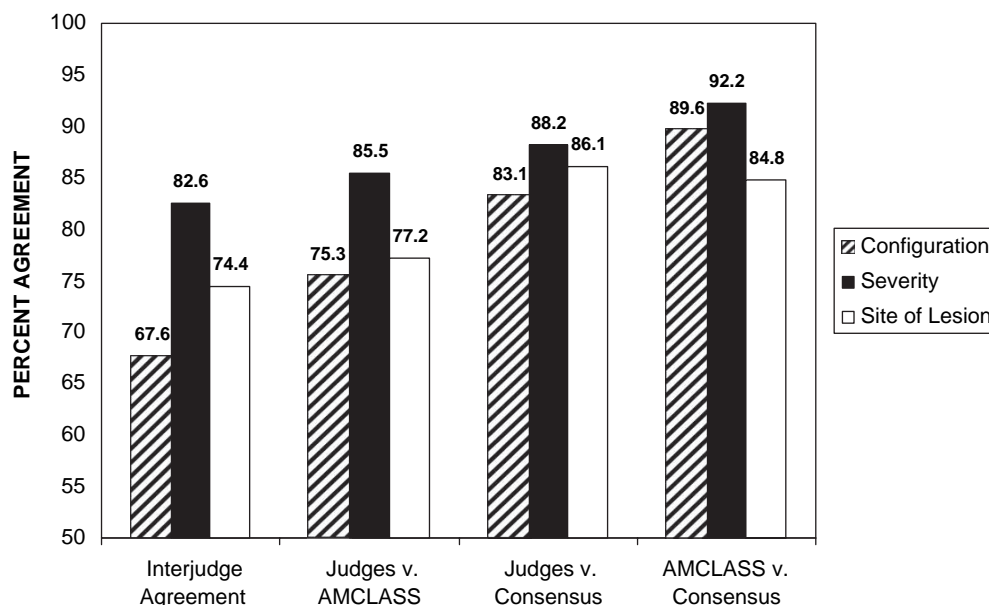
A determination of agreement among judges is important because the goal is that AMCLASS™ should agree with expert judges similarly to the agreement among the judges. If this goal is met, we can claim that AMCLASS™ performs similarly to an expert judge. Agreement between pairs of judges and among all the judges are shown in Table 2. Results are given as percent of cases on which there was agreement and as a Kappa statistic. The Kappa statistic, introduced by Cohen (1960) and described in most statistics texts, is a measure of agreement between categorical data sets that takes into account the probability of agreement due to chance. In the case of the seven AMCLASS™ configuration categories, for example, the likelihood of agreement between a pair of observations due to chance is 1/7. For agreement among all judges, the likelihood of a chance occurrence is  $(1/m)^{n-1}$  where  $m$  is the number of categories and  $n$  is the number of judges.

Agreement between pairs of judges on configuration ranged from 58 to 79% (mean = 68%). There was agreement among all five judges for only 43% of cases. These somewhat surprisingly low levels of agreement may reflect: (1) a large proportion of ambiguous audiograms in the sample, such as those in Figures 1 and 2; and (2) the lack of standard definitions and common understandings of the configuration categories.

It was necessary for severity categories to be different for different configurations (see Table 1) because we wished to characterize the low versus high frequency severity in sloping and rising hearing loss. However, this complicated the comparison of severity judgments when there was disagreement on configuration. We wished to consider two judgments in agreement when they reflected similar hearing loss magnitudes, even when they were judged to have different configurations.

**Table 4.** Agreement percent and Kappa values (in parentheses) between categories selected by each judge and by AMCLASS™. The Consensus column is agreement between the consensus category (selected by the largest number of judges for each case) and AMCLASS™.

	<i>Agreement between judges and AMCLASS</i>						<i>Consensus</i>
	<i>Judge 1</i>	<i>Judge 2</i>	<i>Judge 3</i>	<i>Judge 4</i>	<i>Judge 5</i>	<i>Mean</i>	
Configuration	66.7 (0.61)	78.4 (0.75)	73.2 (0.69)	78.4 (0.75)	80.1 (0.77)	75.3 (0.71)	89.6 (0.88)
Severity	85.7 (0.83)	86.6 (0.84)	86.1 (0.84)	87.4 (0.85)	81.4 (0.78)	85.5 (0.83)	92.2 (0.91)
Site of Lesion	65.4 (0.60)	81.0 (0.78)	73.2 (0.69)	86.6 (0.84)	80.1 (0.77)	77.2 (0.73)	84.8 (0.82)



**Figure 3.** Average comparisons between judges (interjudge agreement), between judges and AMCLASS<sup>TM</sup>, between judges and the consensus of judges, and between AMCLASS<sup>TM</sup> and the consensus of judges.

Accordingly, the pairs of severity categories shown in Table 3 were considered to be in agreement.

Agreement on severity tended to be higher than on configuration. Agreement between pairs of judges ranged from 76% to 87% (mean = 83%). There was agreement among all judges for 61% of cases.

Agreement on site of lesion was higher than on configuration but lower than on severity. Agreement between pairs of judges ranged from 61% to 87% (mean = 74%). There was agreement among all judges for 50% of cases. The CONSENSUS rows in Table 2 indicate the agreement between each judge and the category chosen by the largest number of judges. Agreement between judges and consensus ranged from 73% to 95% with means of 83%, 88%, and 86% for configuration, severity, and site of lesion, respectively.

Studies of disagreement among expert judges in the interpretation of medical tests have found levels of disagreement that are similar to those reported here. Elmore et al (1994), for example, reported on interjudge agreement on the interpretation of mammograms by radiologists. Average agreement between pairs of judges was 78% (Kappa = 0.47). Potchen (2006) reported interjudge agreement among radiologists in judging chest X-rays as normal or abnormal. Agreement between pairs of judges averaged 80%. These interjudge agreement levels are similar to those shown in Table 2 (Mean of Pairs). Ringsted et al (1978) reported that the average agreement on interpretations of cervical biopsies by pathologists with the consensus of a panel of experts was 87%, very similar to the average agreement between individual judges and the consensus of judges reported

here (83%, 88%, and 86% for configuration, severity, and site of lesion, respectively; see Table 2, Mean Consensus).

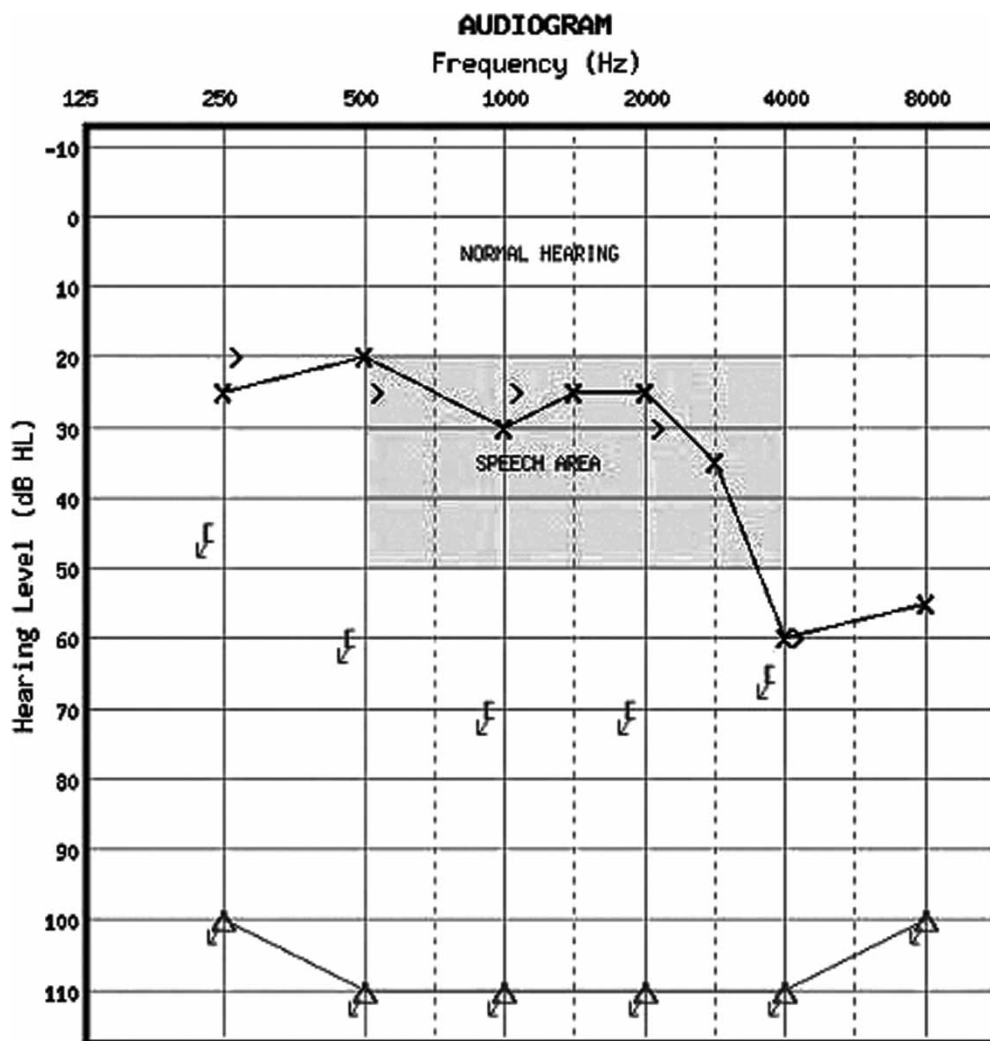
#### Agreement with AMCLASS<sup>TM</sup>

Agreement between AMCLASS<sup>TM</sup> and each judge, the average agreement between AMCLASS<sup>TM</sup> and all judges, and the agreement between AMCLASS<sup>TM</sup> and consensus categories are shown in Table 4. These data provide a basis for determining the relative performance of AMCLASS<sup>TM</sup> and expert judges. For all three audiogram characteristics (configuration, severity, and consensus) the average agreement between AMCLASS<sup>TM</sup> and the judges was higher than the average interjudge agreement (from Table 2, Mean of Pairs), indicating that AMCLASS<sup>TM</sup> agreed more consistently with the judges than the five judges agreed among themselves.

Figure 3 summarizes comparisons among judges (interjudge agreement), between judges and AMCLASS<sup>TM</sup>, between judges and the consensus of judges, and between AMCLASS<sup>TM</sup> and the consensus of judges. For configuration and severity, the best agreement was between AMCLASS<sup>TM</sup> and the consensus of judges. For site of lesion AMCLASS<sup>TM</sup> v. consensus was slightly lower than the average agreement between the judges and consensus. Nevertheless, the agreement between AMCLASS<sup>TM</sup> and consensus for site of lesion was higher than the average interjudge agreement.

The agreement between AMCLASS<sup>TM</sup> and judges for site of lesion was affected by a lack of common understanding of one of the site of lesion categories. The 'sensorineural or mixed' category was intended to be used for audiograms like the one





**Figure 4.** The right ear site of lesion is categorized by AMCLASS™ as ‘sensorineural or mixed.’

shown for the right ear in Figure 4. Although many would judge the hearing loss to be sensorineural, a mixed hearing-loss cannot be ruled out because of the limitation in bone-conduction levels available for testing on clinical audiometers. However, the judges did not consistently use the ‘sensorineural or mixed’ category to describe audiograms of this type. Rather than adjust the rules to maximize agreement between AMCLASS™ and consensus, we feel it is prudent to use the ‘sensorineural or mixed’ category for cases such as the one shown in Figure 4. There were 18 cases of this type in the set of audiograms presented to the judges (8%). When these cases were removed from the comparison of AMCLASS™ and consensus, the agreement for site of lesion increased from 85% to 92%.

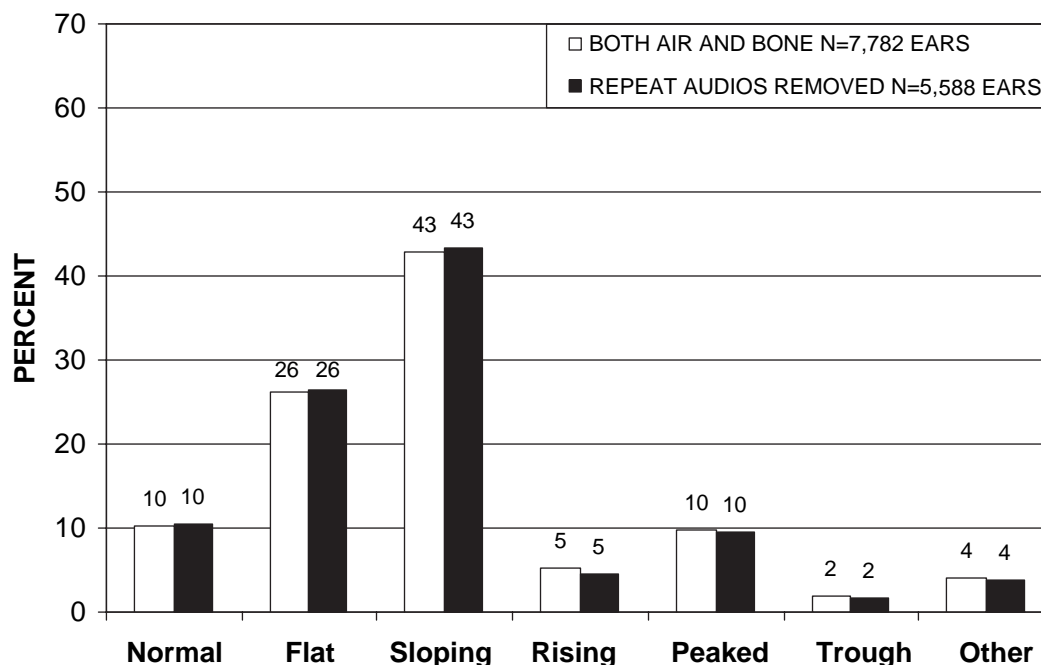
#### *Distribution of hearing-loss categories*

Two potential applications for audiogram classification systems are: (1) description of the proportion of hearing-loss categories in a set of audiograms, and (2) comparison of hearing-loss categories in different populations. There is very little information on the prevalence of hearing-loss categories in the general population or in clinical populations, due at least in part to the

lack of standard definitions. An example of such a description is provided in Figure 5 for 7372 ears of 3686 patients seen in the University of Minnesota Hospital Audiology Clinic over the period June 1989, to January 2003. This sample is a subset of 31,676 records saved to an electronic archive over that period. The subset includes all cases for which complete audiograms were obtained for both ears. A complete audiogram is defined as one that includes threshold values (including ‘no response’) for octave frequencies over the range 250–8000 Hz for air conduction, and 250–4000 Hz for bone conduction. From this sample, the 231 cases used in the validation study were drawn.

The distribution of configurations in Figure 5 is provided for all audiograms meeting the inclusion criteria, and for a smaller set with repeat audiograms removed. There was very little effect of removing the repeat audiograms. Sloping hearing loss was the most prevalent configuration followed by flat hearing loss. The distributions are quite different than those reported by Pittman and Stelmachowicz (2003) for clinical samples of six-year-old children and 60-year-old adults. They reported percentages of 50% and 33% sloping losses for adults and children, respectively, not too different from the 43% we show in Figure 5. But trough-

## Configuration - Both Ears Air and Bone



**Figure 5.** Distribution of configurations for audiograms drawn for a hospital-based clinic archive. Percentages are shown for all audiograms (7782 ears), and for patients with repeated tests omitted (5588 ears).

shaped hearing loss was the second most prevalent configuration with a lower prevalence for flat hearing loss. These trends are quite different from our results. Differences in the distributions may be due to the different category definitions. Differences in the patient samples may have contributed also.

Figure 6 shows the distribution of severities for the same sample with and without repeat tests. Combination severities such as mild-moderate sloping hearing loss and moderate-mild rising hearing loss were grouped with one of the major categories according to the rules described in the figure legend. Mild and moderate hearing losses were the most prevalent. Again, removing repeat tests had very little effect.

Figure 7 shows the distribution of site of lesion for the same sample, with and without repeat tests. Sensorineural and mixed hearing losses were the most prevalent, followed by conductive. Again, removing repeat tests had a negligible effect on the distribution.

The negligible effect of removing repeat audiograms was a surprising result that indicates that no hearing-loss category is more or less likely than other categories to return for repeat evaluations.

The distributions shown in Figures 5–7 represent the characteristics of audiograms for patients tested in a large hospital-based audiology clinic when complete audiograms are obtained for both ears. There may be differences between the characteristics of this sample and other groups of patients seen in this type of facility. It is common for bone conduction to be tested only in one ear when the hearing loss is symmetrical and sensorineural. Almost 10,000 cases in the database received complete air-conduction tests in both ears, and complete bone-conduction in only one ear. Another subset of about 8000 cases received

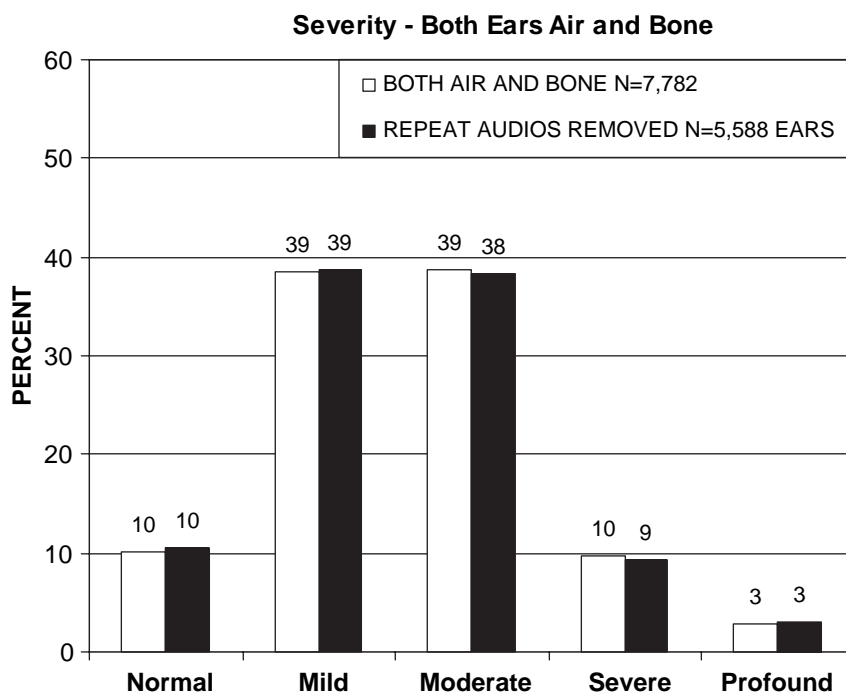
complete air-conduction testing but no bone-conduction testing. These are primarily patients who have been tested previously for whom there was no change in air-conduction thresholds.

### Interaural asymmetry

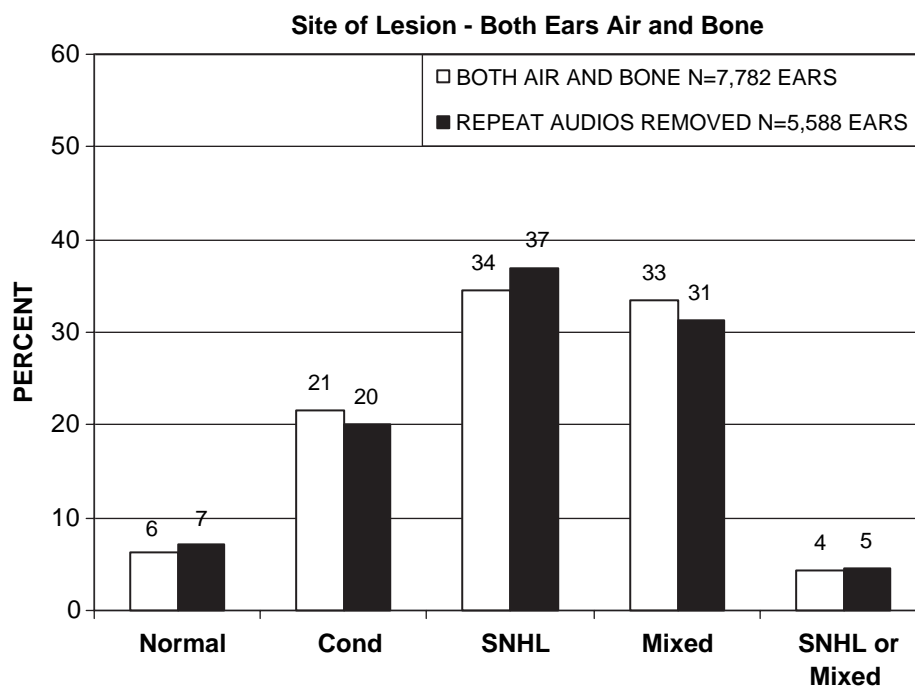
Although the rules for determining asymmetry were not validated for the reasons discussed previously, the results are striking. Eighty-three percent were judged to be asymmetric, whether or not repeat audiograms were omitted. An informal review by one judge (the first author) suggested that the asymmetry judgments made by AMCLASS™ are consistent with those of the judge. This result has important consequences because interaural asymmetry has been suggested as a criterion for referral for further evaluation for acoustic neuroma and other important otologic conditions (e.g. Mangham, 1991).

Pittman and Stelmachowicz reported the proportion of audiometric asymmetries in their samples of six-year-old children and 60-year-old adults. They defined asymmetry based on the number of frequencies at which interaural threshold differences of greater than 20 dB occurred. Forty-five percent of adults and 28% of children had interaural asymmetries greater than 20 dB at one or more frequencies. These are substantially lower than the 83% asymmetry rate in our sample. Differences in sampling methods and definitions of asymmetry likely contributed to these differences. A standard definition of audiometric asymmetry is needed to facilitate such comparisons.

One reason for the high rate of asymmetrical hearing loss may be the inclusion criteria for the sample of audiograms selected for analysis. The cases that were analysed were those who received complete air- and bone-conduction testing in both ears. As mentioned above, when the hearing loss is symmetrical and



**Figure 6.** Distribution of severities for audiograms drawn for a hospital-based clinic archive. Percentages are shown for all audiograms (7782 ears) and for patients with repeated tests omitted (5588 ears). For combination categories (e.g. mild-moderate) the following rules were used to group them into the five major categories. When the combination consisted of adjacent categories (e.g. mild-moderate) it was grouped with the less severe category, with the following exceptions. Normal-mild and mild-normal were grouped with mild. When the combination spanned three major categories it was grouped with the category in the middle (e.g. severe-mild was grouped with moderate).



**Figure 7.** Distribution of site of lesion for audiograms drawn for a hospital-based clinic archive. Percentages are shown for all audiograms (7782 ears), and for patients with repeated tests omitted (5588 ears).

sensorineural, it is common to test bone conduction in only one ear. Thus, the sample is probably biased toward a high proportion of asymmetrical hearing losses. To determine the proportion of asymmetrical hearing losses in the clinic population, an analysis of the larger database is necessary. In addition, rules for defining asymmetrical hearing loss require validation. An analysis of audiometric asymmetry for the larger database and validation study of the rules for identifying asymmetrical hearing loss will be reported separately.

### Limitations of the study

The method for validating AMCLASS™ utilized the responses of a panel of expert judges who viewed the results from one ear of a set of patients with a wide variety of hearing losses. The intent of providing results from one ear was to eliminate the bias that might occur if both ears were presented. For example a loss that is on the borderline between flat and sloping might be more likely to be judged as sloping if the configuration of the other ear was sloping. It could be argued that showing both ears is a more realistic situation but for the purpose of validating the method, we viewed it as a potential source of bias.

The method does not take into account other important information such as case history, previous audiograms, and other test results such as tympanometry and speech audiometry. All of the sources of relevant information must be utilized by the clinician and it may be appropriate to overrule the outcome of AMCLASS™ based on the totality of information available to the clinician.

It is possible that the results would be different with a different set of judges or a different set of audiograms. Audiograms of children, for example, may be interpreted differently by judges than those of adults. The cases for the validation study were selected from a database in which each age decade up to the eighth were equally represented.

### Summary and Conclusion

A set of rules (AMCLASS™) has been developed for categorizing the configuration, severity, site of lesion, and interaural asymmetry of an audiogram based on categories that are commonly found in the audiology literature and in clinical practice. AMCLASS™ was validated against the categories selected by a panel of expert judges. Surprisingly low agreement rates were found among judges, suggesting that there are no standard definitions of the categories employed, and judges have different understandings of descriptive terms commonly used for describing audiometric results. The validation data obtained from the panel of judges on 231 audiograms allowed the design of rules that maximize agreement with the judges. An analysis of agreement among judges and between judges indicates that agreement between AMCLASS™ and judges exceeds the average interjudge agreement in selecting categories. AMCLASS™ may provide a method for: (1) categorizing audiograms for research purposes, such as studying relationships between audiometric characteristics and ear disease; (2) providing a validated interpretation that can facilitate a more consistent approach to treatment; (3) facilitating consistent communication between clinicians and

patients, and between professionals; and (4) teaching audiogram interpretation to students.

### Acknowledgements

Our panel of expert judges created an enormously valuable database without which this study would be possible. We are grateful to Deborah Abel, Victor Berrett, Sam Levine, and Don Morgan for lending their expertise to this project. Three anonymous reviewers provided excellent suggestions for the revision of the manuscript.

AMCLASS™ is available for limited use at [www.audiologyincorporated.com](http://www.audiologyincorporated.com). For addition information, contact the authors.

### References

- AMA Council on Physical Therapy. 1942. Tentative standard for evaluating the percentage of useful hearing loss [sic] in medicolegal cases. *J Am Med Assoc*, 119, 1108–1109.
- Bess, F.H. & Humes, L.E. 1990. *Audiology: The Fundamentals*. Baltimore: Williams & Wilkins.
- Carhart, R. 1945. An improved method for classifying audiograms. *Laryngoscope*, 55, 640–662.
- Clark, J.G. 1981. Uses and abuses of hearing-loss classification. *Asha*, 23, 493–500.
- Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educ Psychol Meas*, 20, 37–46.
- Davis, H. 1965. Guide for the classification and evaluation of hearing handicap in relation to the international audiometric zero. *Trans Amer Acad Ophthalmol Otolaryngol*, 69, 740–751.
- Elmore, J.G., Wells, C.K., Lee, C.H., et al. 1994. Variability in radiologists' interpretation of mammograms. *New England Journal of Medicine*, 331, 1493–1499.
- Fletcher, H. 1929. *Speech and Hearing*. New York: Van Nostrand.
- Fowler E.P. & Wegel R.L. 1922. Audiometric methods and their applications. *Trans Amer Laryngol Rhinol Otol Soc*, 98–132.
- Guild, S.R. 1932. A method for classifying audiograms. *Laryngoscope*, 42, 821–836.
- Kaplan, H., Gladstone, V.S. & Lloyd, L.L. 1993. *Audiometric Interpretation: A Manual of Basic Audiometry*. (2nd edition) Boston: Allyn & Bacon.
- Keith, R.W. 1996. The audiological evaluation. In J.L. Northern (ed.), *Hearing Disorders*. (3rd edition) Boston: Allyn & Bacon.
- Lee, C.-S., Paparella, M.M., Margolis, R.H. & Le, C. 1995. Audiological profiles and Meniere's disease. *Ear Nose & Throat J*, 74, 527–532.
- Lonsbury-Martin, B.L., Martin, G.K. & Tedeschi, F.F. 1999. Otoacoustic emissions in clinical practice. In R.J. Roeser, M. Valente & H. Hosford-Dunn. (eds.) *Audiology: Diagnosis*. New York: Thieme.
- Macrae, J.H. 1975. A procedure for classifying degree of hearing loss. *J Otolaryngol Soc Aust*, 4, 26–34.
- Mangham, C.A. 1991. Hearing threshold difference between ears and risk of acoustic tumor. *Otolaryngol Head & Neck Surg*, 105, 814–817.
- Martin, F.N. 1986. *Introduction to Audiology*. Englewood Cliffs, NJ: Prentice-Hall.
- Neary, W.J., Newton, V.E., Laoide-Kemp, S.N., Ramsden, R.T., Hillier, W.F., et al. 1996. A clinical, genetic, and audiological study of patients and families with unilateral vestibular schwannomas. *J Laryngol Otol*, 110, 1120–1128.
- Paparella, M.M., McDermott, J.C. & deSousa, L.C.A. 1982. Meniere's disease and the peak audiogram. *Arch Otolaryngol*, 108, 555–559.
- Pittman, A.L. & Stelmachowicz, P.G. 2003. Hearing loss in children and adults: Audiometric configuration, asymmetry, and progression. *Ear Hear*, 24, 198–205.
- Potchen, E. 2006. Measuring observer performance in chest radiology: Some experiences. *J Amer Col Radiol*, 3, 423–432.
- Ringsted J., Amtrup F., Asklund C. et al. 1978. Reliability of histopathological diagnosis of squamous epithelial changes of the

- uterine cervix. *Acta Pathologica Microbiologica et Immunologica Scandinavica*, Section A: Pathology, 86, 273–278.
- Roeser, R.J., Buckley, K.A. & Stickney, G.S. 2000. Pure tone tests. In R.J. Roeser, M. Valente & H. Hosford-Dunn. (eds.) *Audiology: Diagnosis*. New York: Thieme.
- Savastano, M., Guerrieri, V. & Marioni, G. 2006. Evolution of audiometric pattern in Meniere's disease: Long-term survey of 380 cases evaluated according to the 1995 guidelines of the American Academy of Otolaryngology: Head and Neck Surgery. *J Otolaryngol*, 35, 26–29.
- Stach, B.A. 1998. *Clinical Audiology: An Introduction*. San Diego: Singular Publishing Group.
- Tempest, W. 1976. Medicolegal aspects of noise induced hearing loss. In S.D.G. Stephens (ed.), *Disorders of Auditory Function*. London: Academic Press.
- Yantis, P.A. 1994. Pure-tone air-conduction threshold testing. In J. Katz (ed.), *Handbook of Clinical Audiology*. (4th edition) Baltimore: Williams & Wilkins.