SPEECH INTELLIGIBILITY TESTING FOR NEW TECHNOLOGIES

Susan L. Hura

Multimedia Perception Assessment Center
Lucent Technologies
101 Crawfords Corner Road, Room 1L-506
Holmdel, NJ 07733
slhura@lucent.com

1. ABSTRACT

There are several tests of speech intelligibility currently available which employ a variety of methods. The most appropriate method for testing intelligibility of speech transmitted via telephony is a forced choice task in which listeners hear speech samples and identify what they hear from among a set of alternatives displayed onscreen. This methodology allows tests to be run quickly and scored automatically. A major flaw in existing forced-choice intelligibility tests is the use of unfamiliar words, nonwords, and proper names along with common real words. A stimulus set that is mixed in this way may introduce response biases into the test and therefore produce results that are less predictive of actual intelligibility performance. The Intelligibility of Familiar Items Test (IFIT) ameliorates several methodological flaws found in earlier tests. The IFIT uses a stimulus set composed of high familiarity real English words and tests consonants in initial and final word position and vowels in word medial position.

2. BACKGROUND

Speech intelligibility testing for telephony has a long history. Since the early 1900's researchers have attempted to define a representative language sample for intelligibility testing and an appropriate method of assessing listeners' perceptions of these samples. The ultimate goal of intelligibility testing is a realistic estimate of how easy it will be, under normal conversational conditions, to understand a talker's voice transmitted via telephony. Intelligibility tests employ a variety of stimulus materials and require different tasks of subjects. Tests which use full sentence stimuli are a realistic approximation of real world conditions; however, it is difficult to control for effects of context and predictability within the stimulus set. Some intelligibility tests require spoken or written responses from subjects. This sort of test may provide valuable information, but given the large amount of time required to run subjects and score responses, such tests do not allow for the extremely rapid testing necessary in today's telecommunications industry. Many new technologies such as digital transmissions, speech coders, and Internet telephony are emerging quickly and show audio impairments not present in traditional analog systems. Therefore, there is a need for an intelligibility test that is quick to run and score, but that provides comprehensive coverage of the phonetic inventory of the language. The Intelligibility of Familiar Items Test (IFIT) de is aimed at meeting these goals and at ameliorating some problems found in existing intelligibility tests.

2.1. Existing Intelligibility Tests

The Diagnostic Rhyme Test (DRT, [1]) and the Minimal Pairs Intelligibility Test (MPI, [2]) are two popular intelligibility tests that share many characteristics of the proposed IFIT. However, the stimulus words in each of these tests are not optimal for subjective testing of the intelligibility of processed speech. The DRT and MPI use a 2-item forced choice format and use minimal pairs of words as stimuli. Because the MPI test was designed to test the intelligibility of synthetic speech, its stimulus list is composed of multi-syllabic words with differing phonetic content and stress patterns. These items therefore do not constitute an appropriate set for the current purpose. The DRT stimulus set is composed of real words, proper names, and nonwords, and both the DRT and MPI use very unfamiliar words. The problem with using a mixed stimulus set such as this is that listeners process words differently from nonwords (see for example [3, 4]), and that proper names involve yet other perceptual processes. Furthermore, the familiarity of words can also significantly affect subjects' responses to the stimuli [5, 6, 7] such that they are biased to select more familiar words as responses. That is, listeners may make errors on particular items in an intelligibility test because the test items are unfamiliar, rather than because they are unintelligible. Therefore a mixed stimulus set may introduce error into the testing process and produce data that is less reliable. To improve the quality of intelligibility results, the proposed IFIT has a more uniform stimulus set, which will likely produce more reliable data and may be sensitive to finer differences in intelligibility. Such fine distinctions are increasingly important in today's competitive telecommunications industry.

2.2. Goals Of The New Test

The stimuli in the new IFIT test are minimal pairs of words differing by one phoneme, (e.g., *type* vs. *tight*). A 2-item forced choice method is used to determine subjects' ability to discriminate the contrasting phonemes. The 2-item forced choice methodology is important in creating a test that can be easily implemented on a desktop computer, and that can be run rapidly and scored automatically. To meet the goal of comprehensive coverage of the phonetic inventory of the language, IFIT stimuli represent three sorts of phoneme

distinctions: 1) consonants in the initial position in the word (*toes/nose*); 2) consonants in final position in the word (*type/tight*); 3) and vowels in medial position in the word (*lake/lack*). The set of phoneme distinctions tested was generated without appeal to a specific theory of distinctive features, a choice that is in contrast to existing intelligibility tests. The words chosen to represent phoneme distinctions are all common, reasonably familiar real words of English likely to be known by the average subject. The familiarity and consistency of the items in the test is the most unique feature of the new intelligibility test.

3. DESIGN OF THE NEW TEST

3.1. The Problem With Distinctive Features

There are two outstanding problems with 2-item forced choice tests of speech intelligibility: defining the contrasts to be tested and finding words to represent these contrasts. Some tests, including the Diagnostic Rhyme Test and the Minimal Pairs Intelligibility test, generate segment contrasts by varying +/- feature values within a theory of distinctive features. Both the MPI and DRT use a perceptually based feature system, roughly modeled after that of Jakobson, Fant and Halle [8]. The rationale for this choice in the design of the DRT is that these distinctive features capture the relevant acoustic/perceptual properties of the speech signal as well as the relevant differences between segments. In the design of the MPI, van Santen uses distinctive features as a method for generating segment contrasts but makes clear that his choice was made on practical grounds: distinctive feature theory simply provides a framework for developing a reasonable set of segment contrasts.

One potential problem with relying on a set of features to generate segment contrasts is deciding how to select the appropriate set of features. Like any theory, distinctive feature theory is under debate and the optimum set of features for describing speech sounds has not been universally agreed upon. Since the 1960's the majority of phonologists have favored a distinctive feature theory based on articulatory rather than perceptual characteristics of speech sounds. Although perceptual features may seem more relevant for an auditory test of speech intelligibility, Jakobsonian features are certainly not the most accepted in the field. A more basic issue with the use of distinctive feature theories is that the contrasts to be tested are limited by the specific properties of the theory. Because distinctive feature theories are usually exclusively articulatory or exclusively perceptual/auditory, the contrasts tested will similarly be skewed. Because both articulation and perception certainly play a role in discriminating speech segments, an intelligibility test based on a particular distinctive feature theory may omit likely confusions or include unlikely ones because of the properties of the theory itself.

All this said, given the limited phonetic inventory of English, the set of contrasts developed by most distinctive feature theories will be in large part the same. It is clear that [p] vs. [b] is a likely confusion and that [θ] vs. [w] is not, whatever the bent of a particular theory. In large part, a set of likely segment confusions can be defined independent of any particular theory of distinctive features. Given the information in the literature about segment confusability (e.g. [9]), the acoustic characteristics of speech sounds, and speech perception in general, it is relatively straightforward to select appropriate segment pairs for intelligibility testing. In fact, this is the solution I have chosen to employ in the new test. Here segment contrasts were chosen in a common sense manner that does not strictly adhere to any single DF theory, although it agrees in most instances with contrasts generated according to distinctive feature theories.

3.2. Defining Segment Contrasts

The first step in developing the new test of intelligibility was to devise rules for generating segment contrasts for consonants and vowels. The most likely and logical confusions were included; no attempt was made to conform to any single distinctive feature theory. The development of the set of consonant contrasts began by dividing obstruents from approximants. Throughout this test, stops, fricatives and affricates are contrasted with one another and liquids and glides are contrasted separately amongst themselves. Each obstruent are contrasted with:

- all segments sharing the same manner of articulation and voicing (i.e., all voiceless fricatives);
- the segment sharing the same place and manner of articulation but opposite voicing; and
- all other obstruents at the same place of articulation across manner classes.

Because English stops and fricatives do not occur at identical places of articulation, the correspondences are rough: bilabial stops are contrasted with labiodental fricatives, alveolar stops are contrasted with both alveolar and interdental fricatives, and velar stops are contrasted with palato-alveolar fricatives. The palato-alveolar affricates are contrasted with alveolar stops. Both stops and fricatives are contrasted with the nasal stop at the same place of articulation, irrespective of voicing. The four approximant consonants are contrasted only with each other.

Vowel contrasts began within backness categories: all front vowels are contrasted with each other, and all back vowels are contrasted with each other. Tense/lax vowel pairs are contrasted, and all lax vowels are contrasted with [ə]. Finally, each vowel is contrasted with the segment at the same height with the opposite backness value (e.g., [i] is contrasted with [u]). Note that the mid back lax rounded vowel [ɔ] is included in the stimulus set; care should be taken in interpreting responses to [ɔ] because this vowel is not distinct from [ɑ] in many dialects of American English.

These rules generate 31 vowel pair contrasts and 65 consonant pair contrasts. All vowel contrasts are tested in word-medial position; consonant contrasts are tested in both initial and final position, if possible. Because consonant

segments such as [ŋ] and [h] are disallowed in certain word positions, 12 initial and 9 final contrasts are omitted from the test. Additionally, there are 6 word-final consonant contrasts, each including the segment [ʒ], and 1 initial contrast [θ]/[ð] which are omitted because there are no appropriate common English words containing those segments. The IFIT test contains 133 pairs of words overall: 31 vowel contrasts and 102 consonant contrasts (52 with contrasting segments in initial position and 50 with contrasting segments in final position). The DRT omits 32 of the consonant contrasts tested in the IFIT and all vowel contrasts; the MPI omits 21 of the consonantal contrasts and 13 of the vowel contrasts.

3.3. Selecting words to represent segment contrasts

A pair of one-syllable real English words was selected to represent each contrast. All stimulus words have a consonant-vowel-consonant shape; 10 items testing vowel contrasts or consonant contrasts in initial word position have the shape CVCC. The primary goal in selecting items was to choose reasonably common words, likely to be known by an average adult native speaker of American English. Proper names and phonologically possible nonwords were not used, but there were no constraints on part of speech, spelling, or formality. That is, some items are slang terms (e.g., *yuck, coke*), others are morphologically complex (*taught, pays*), other items were function words (*that, with*). Pairs of test items did not necessarily share a common spelling, so long as they formed a minimal pair (e.g., *type* vs. *tight*). Although controlling test items for these factors would yield a more homogeneous test, doing so necessitates other undesirable choices, such as mixing proper names with common nouns, or using highly unfamiliar words or nonsense syllables. The overriding concern in choosing items for this test was to have each item be quite familiar to the subjects, so as not to distract them from the perceptual task with confusion over word meaning.

Some additional phonetic constraints were used when selecting words to represent phoneme contrasts. For vowel contrasts, words ending in nasal and liquid consonants were avoided because these consonants significantly affect the typical format frequency patterns of vowels. For consonant contrasts, a different vowel was used when testing a consonant contrast in initial and final positions (i.e., if the initial contrast was represented by the word pair *pat/bat*, the final contrast would not be *cap/cab*). Moreover, an attempt was made to select words for the consonant stimuli containing a wide variety of vowels, and to select vowel stimulus words containing a variety of consonants. Recall that each phoneme contrast occurs only once for each vowel pair and once or twice for each consonant pair in the IFIT test. Therefore, complete diphone balance could not be achieved within the stimulus set; that is, consonant contrasts are not tested in each possible vowel environment and vowel contrasts are not tested in all possible consonant environments. To achieve this type of balance would require a huge increase in the number of stimuli and would greatly increase testing time. Instead, the IFIT stimuli contain a wide variety of consonant-vowel combinations across the test which should prove representative of general pattern in the language.

3.4. Published Word Familiarity

The published word frequency and familiarity of stimulus items was obtained from the online MRC database [10]. The central tendencies for frequency and familiarity are shown in Table 1. Ninety-five of the 266 stimuli in the IFIT have no published familiarity score; the column labeled "Published Familiarity" represents the familiarity scores for the remaining 171 items. The test items are, as a group, relatively infrequent, but highly familiar words. There are only 13 words whose Kucera-Francis [11] frequency is reported at 2000 or higher (all are function words: *that, with, his, but, have, has, some, could, these, then, did, must, such*). Among the 171 items with a published familiarity score, the majority had scores of 5 or better on a 7 point scale, indicating the words are highly familiar. There are several items which are both high frequency and high familiarity (e.g., *thought, that, his, with)*. However, there are also several items that are low frequency and high familiarity (e.g., *wash, shop, thumb, sheet)*.

	IFIT Word Frequency	IFIT Published Familiarity	DRT Published Familiarity
Mean	241.9	542.1	525.6
Median	21	555	542
Mode	0	585	541
SD	990.92	58.9	68.2
Range	10595 – 0	643 - 295	632 – 300
n	266	171	103

Table 1: Central tendencies for word frequency and published familiarity distributions.

As a comparison, the familiarity scores for the items in the DRT were also obtained from the MRC Database. As with the IFIT items, many of the stimuli had no published familiarity score, so the figures in Table 1 represent data on 103 of the 192 DRT stimulus items. Considering only the published familiarity scores, the distributions of the IFIT and the DRT do not appear to be very different from one another. However, a t-test for differences between the means showed that the IFIT items are significantly more familiar than the DRT items as a group ($t(190)=2.046$, $p<.04$).

4. A FAMILIARITY EXPERIMENT

To verify the differences in word familiarity between the stimuli included in the IFIT versus those in the DRT, a test of word familiarity was conducted.

4.1. Stimuli

The stimuli for the experiment were the stimulus words from the two intelligibility tests. Each stimulus was a one-syllable word, nonword, or proper name. There were a total of 421 distinct words (266 IFIT items + 192 DRT items – 37 items appearing on both lists).

4.2. Subjects

Seventy-three undergraduate students at a major university participated in the familiarity study for course credit. Data from seven subjects was discarded because the subjects were non-native speakers of English or because they failed to follow instructions, leaving 67 subjects.

4.3. Procedures

Subjects were run in two approximately equal sized groups, each group receiving a different randomization of stimuli. Subjects were given a packet containing instructions, the printed stimulus list, and several machine-readable answer sheets. Subjects were instructed to rate each item on the test according to how familiar they were with the word based upon how commonly or frequently they encounter it. Subjects marked their responses on a 1 – 7 scale where 1 represented "not familiar at all" and 7 represented "very familiar".

4.4. Results

Results were collapsed across the two randomizations, which were not significantly different from one another, and across subjects. The average familiarity score for the DRT stimuli was 3.97 (*SD*=1.06); the average score for stimuli on the IFIT was 4.63 (*SD*=1.33). Figure 1 shows the distribution of familiarity scores for items on the two tests. Both the IFIT and DRT show familiarity ratings across the entire scale; however, the IFIT ratings are skewed towards the high familiarity end of the scale while DRT items show a flat distribution across the scale. In fact, only 28% of IFIT items have a score of 3 or lower versus 50% of DRT items. Conversely, 42% of IFIT stimuli were rated 5 or better versus only 27% of DRT items.

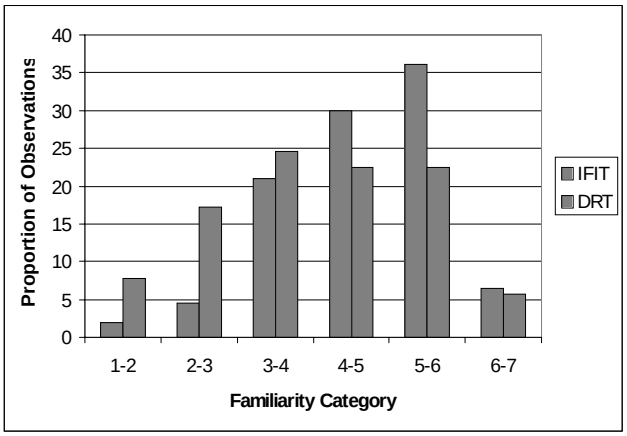

Figure 1: Familiarity distributions for DRT and IFIT stimuli.

Tests for differences between the means confirmed that the words included in the IFIT are significantly more familiar than the stimuli in the DRT ($t(456) = 5.88$, $p < .001$). These data suggest that the IFIT meets the goal of improving on existing intelligibility tests by using more familiar words as stimuli.

5. REFERENCES

1. van Santen, J. P. H. "Perceptual Experiments for Diagnostic Testing of Text-to-Speech Systems," *Computer Speech and Language*, 7: 49-100, 1993.

2. Voiers, W. D. "Evaluating Processed Speech Using the Diagnostic Rhyme Test," Speech Technology, Jan/Feb: 30-39, 1983.

3. Ganong, W. F. "Phonetic Categorization in Auditory Word Perception," *Journal of Experimental Psychology: Human Perception and Performance*, 6: 110-125, 1980.

4. Connine, C. M. and Clifton, C. "Interactive Use of lexical Information in Speech Perception," *Journal of Experimental Psychology: Human Perception and Performance*, 13: 291-299, 1987.

5. Gilhooly, K. J. and Logie, R. H. "Meaning-Dependent Ratings of Imagery, Age-of-Acquisition, Familiarity, and Concreteness for 387 Ambiguous Words," *Behavior Research Methods & Instrumentation*, 12: 428-450, 1980.

6. Paivio, A. V., Yuille, J. C., and Madigan, S. A. "Concreteness, Imagery, and Meaningfulness Valuesfor 925 Nouns," *Journal of Experimental Psychology Monograph, 76* (3, Pt. 2). 1968.

7. Toglia, M. P., and Battig, W. F. *Handbook of Semantic Word Norms.* Erlbaum, Hillsdale, N.J., 1978.

8. Jakobson, R., Fant, G. and Halle, M. *Preliminaries to Speech Analysis*, MIT Press: Cambridge, MA, 1952.

9. Miller, G. and Nicely, P. "An Analysis of Perceptual Confusions Among Some English Consonants," *Journal of the Acoustical Society of America*, 27: 338-352, 1952.

10. Coltheart, M. "The MRC Psycholinguistic Database," *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, 33A: 497-505, 1981.

11. Kucera, H. and Francis, W. *Computational Analysis of Present-Day American English*, Brown University Press, Providence, RI, 1967.