Objective Measurement of Intelligibility

Jared Bernstein
Ordinate Corporation, Menlo Park, California
Stanford University, California, USA
E-mail: jared@ordinate.com

ABSTRACT

Intelligibility is not often reported in language testing because measurement procedures have practical problems. Intelligibility is sensitive to the speaking skill of the candidate as well as to the predictability of the material and the intelligence and experience of the listeners. A given nonnative speaker will be more intelligible when saying predictable material to intelligent listeners who have experience with nonnative speech. Also, in a real language assessment procedure, after a listener has heard one candidate speak on a particular subject, that listener is no longer "naive" and will be better able to understand the next candidate's speech sample. An experiment conducted at Indiana University attempted to establish a stable intelligibility scale for nonnative speech. Utterances from 485 nonnative speakers of English were presented to 141 naive native listeners in such a way that no listener heard the same item from more than one speaker. This produced over 27,000 individual listener responses. Analysis of these data allows the nonnative speakers to be placed on an intelligibility scale that is reliable and clearly interpretable in terms of percent of words correctly heard for materials at a known level of predictability.

1. INTRODUCTION

Assessments of the spoken language proficiency of second-language learners often presume that a rater can roughly judge the intelligibility of the learner's speech on an approximate ordinal scale with only two anchor points: completely understandable and completely unintelligible. These rough, holistic judgments sometimes appear in language test score reports. People who use language test scores might prefer a numeric estimate of intelligibility (e.g., that a naive native listener would understand 60% of the candidate's words in a particular domain of discourse) which would be easier to interpret without any specific expertise. Because judgments of pronunciation quality typically correlate well with intelligibility, pronunciation quality judgments are sometimes used and reported directly in language test scores, even though the score user might prefer an intelligibility score, as such.

The principal reason that actual objective measures of intelligibility are not commonly used in language testing is because precise procedures for measuring intelligibility have many practical problems. Intelligibility is sensitive not only to the speaking skill of the candidate, but also to the predictability of the material, and to the intelligence and language experience of the listeners. Intelligibility can be schematized as:

(1) **Intelligibility = F(pron., material, context, listener)**

where *pron.* is *pronunciation quality.* A given nonnative speaker will be more intelligible when saying predictable material to listeners of high intelligence who have extensive experience with nonnative speakers. Also, in a real language assessment procedure, after a listener has heard one candidate read a particular passage or speak on a particular subject, that listener is no longer "naive"and will typically be better able to understand the next candidate's speech sample.

2. COMMUNICATIVE COMPETENCE

People who need to interpret language test results and use scores as part of a decision process want scores in a form that can be understood. Many standardized language proficiency tests (e.g., TOEFL or PhonePass SET-10) report performance on a numeric scale that needs to be interpreted and understood in relation to the required activities that a candidate will be expected to participate in.

Tests that are framed in terms of communication functions hope to produce scores that are self-explanatory. However, when one considers the structure of communicative competence as currently conceived in applied linguistics (e.g., Bachman, 1990 [1]), it may seem that the task of transparent score reporting would present some difficulty.

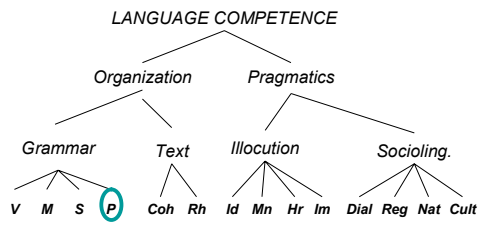

Figure 1: *Taxonomy of Language Competence in communication (after Bachman, 1990). P (circled) is phonology, or pronunciation – one factor among many.*

Figure 1 just presents a taxonomy of phenomena, or factors,

ISBN 1-876346-48-5

that make up communicative language competence; it does not relate them to each other in terms of relative weight or their mode of combination in predicting successful communication. Nor does the taxonomy in Figure 1 relate these aspects of competence to any speaker-external factors such as the skills or experience of the interlocutor.

In order to make any prediction of how well a speaker might actually perform in communication, a person would need to provide a combination rule and scales for the various factors or elements in Bachman's model. Based on data presented in Bernstein et al. (1999) [2], we can suggest that one reasonable structure within which to place these factors would be an equation of the form:

(2) **comm. ≅ pron.• lex. • (1 + syn.+ rhet.+ illoc.+ soc.)**

or in words, communication skill can be approximated as a product of pronunciation skill (intelligibility) times vocabulary (lexical control) times one plus all the other factors summed. The intuitive and observational basis of the combination structure in equation (2) is that a person's communication is absolutely limited by pronunciation and vocabulary, but deficits in all the other factors (syntax, rhetorical form, pragmatics, and sociolinguistics) can be compensated for in some way. For many populations of nonnative speakers of English, either a pronunciation score alone or a vocabulary score alone will predict communicative effectiveness with a correlation of about 0.7.

An excellent understanding of vocabulary measurement for human tasks has evolved over the past decades (see, e.g., Read, 2000 [3]). In the same period, psycholinguists and speech engineers have developed performance measures for spoken language transmission to computers and human listeners. These measures are typically based on intelligibility, which is measured as word error rate (see e.g., Miller & Isard, 1963, [4] or Fourcin et al., 1989 [5]). In applied linguistics, it would be useful to have a numeric scale of intelligibility for nonnative speech that can be directly interpreted in standard circumstances, especially if the numeric scores are anchored to a known population in an understandable way.

3. LISTENING EXPERIMENT

An experiment was conducted at the Speech Research Laboratory at Indiana University through the courtesy of Prof. David Pisoni. The experiment provided data with which to establish a stable intelligibility scale for a range of nonnative speech. In overview, utterances from 485 nonnative speakers of English were presented to 141 naive native listeners in such a way that no listener heard the same item from more than one speaker. This produced over 27,000 individual listener responses. Analysis of these data allows the nonnative speakers to be placed on a unidimensional intelligibility scale that is reliable and clearly interpretable in terms of percent words correctly heard for materials at a known level of predictability.

The experiment reported here in a preliminary form was designed to support the use of PhonePass SET-10 scoring to predict how intelligible a nonnative speaker will be to a particular population of listeners – undergraduate students attending a university in the United States.

3.1 MATERIALS

A balanced set of 461 test takers was assembled from a large database that archives SET-10 test performances (see www.ordinate.com). The test-taker set was balanced for gender and distributed over many native languages. Of the 461 test takers, 71 (15%) were native speakers of English. From these 461 test performances, 5,585 response tokens (about 12 per test-taker) were selected for presentation to naive native listeners. The response tokens were recordings of single-sentence utterances made in response to one of 246 test items selected from the first two parts of the SET-10 test (Readings or Repeats). Examples include "Larry took down five, but one at a time." and "The endless city has no coherent mass transit system."

3.2 PROCEDURE

The naive native-English listeners were a group of 150 undergraduate student volunteers at Indiana University. Each listener called into the PhonePass system and was presented with 200 response tokens. The listeners were instructed to listen carefully to each response and to repeat it verbatim into the phone. Listeners had no option to rehear the recording. The 200 test-taker recordings were selected for presentation to the nave listeners in such a way that the listeners did not hear more than one response token to a given SET-10 test item.

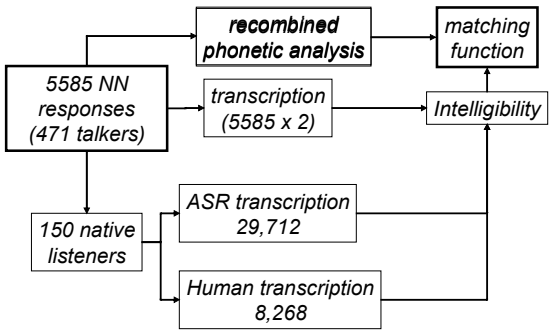

Figure 2: *Estimation Method: data flow in experiments to form an intelligibility reference set for a matching function.*

The naive listeners produced a set of 29,712 usable responses. These listener responses were then used to estimate the intelligibility of the test takers. All 29,712 listener responses were transcribed by automatic speech recognition (ASR), and a subset of 8,268 of the listener responses were also transcribed by human operators.

3.3 RESULTS

Intelligibility was measured as word error rate (WER). WER was calculated for a test taker by comparing the words found in the *listener's* spoken response to the words

in the test taker's original response token. WER was implemented as the minimum number of substitutions, deletions, and insertions needed to match two word strings, with leading and trailing material ignored.

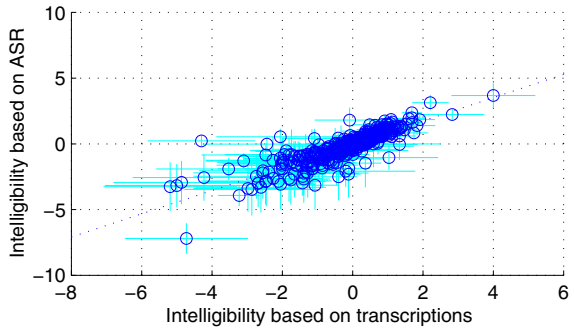

Figure 3: *Word Error Rate, in Logits, based on Human vs. Machine Transcriptions; N = 443; r = 0.86*

To establish that WER can be estimated accurately based on automatic recognition of the listener's responses, we compared the WER estimates for 443 test takers for whom we had sufficient naive listener responses that had been transcribed by both humans and by ASR. The reliability of the WER done by ASR was 0.80 and that done by human operators was 0.78, and the correlation between the two estimates was 0.86. Figure 3 shows a scatter plot of the two estimates for a set of 443 test-takers.

For a set of 459 test takers, the correlation between the SET-10 Overall scores with the WER intelligibilities is 0.61. Using the SET-10 pronunciation subscore only, the correlation with WER is 0.65. If we calculate the expected pronunciation score as a function of WER, then we get the curve shown in Figure 4. Note that the average WER for self-reported native speakers was 6%.

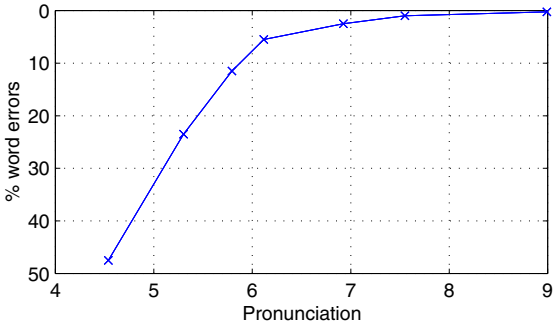

Figure 4: *Trend of SET-10 Pronunciation Score in relation to Word Error Rate*

The SET-10 Pronunciation subscore is based on a nonlinear combination of measures of the acoustic speech signal that has been optimized to match human judgments of pronunciation quality, not intelligibility. In trying a preliminary recombination of the base measures to predict WER, we have found, so far, that we can increase the correlation between the machine scores and the listener-derived WER scores to 0.75.

Figure 5 shows a scatter of machine predicted intelligibility scores vs. intelligibility estimates from an analysis of the listener responses. The plot is for a set-aside test set (one third of the data).

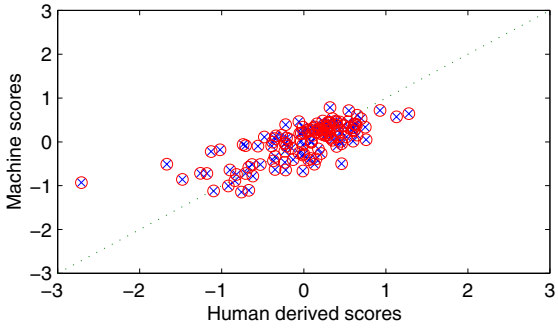

Figure 5: *Machine-predicted WER vs. Listener-based WER; N = 153; r = 0.75*

4. DISCUSSION

This work describes the basis of a computer-based procedure that automatically estimates a nonnative candidate's position on an intelligibility scale based on a set of acoustic base measures on segments, syllables, words, and phrases, given no context, known material and a particular listener population. In order to go further to test the hypothesis embodied in equations (1) and (2), one needs to justify an appropriate entropy scale on the sentence materials and make finer predictions of intelligibility.

5. CONCLUSION

Initial work on predicting intelligibility (calculated as WER) has shown positive results. There is a relatively smooth, if noisy, relation between SET-10 pronunciation scores and WER, although preliminary experiments show promising improvement in the predictive relationship.

REFERENCES

[1] Bachman, L. (1990). Fundamental considerations in language testing. Cambridge: Cambridge University Press.

[2] Bernstein, J., Lipson, M., Halleck, G., & Martinez, J. (1999). Comparison of oral interviews and automatic tests of spoken language. LTRC-99, July, Tsukuba, Japan.

[3] Read, J. (2000). Assessing vocabulary. Cambridge: Cambridge University Press.

[4] Miller, G. & Isard, S. (1963). Some perceptual consequences of linguistic rules." JVLVB (2) pp. 217-228.

[5] Fourcin A., Harland, G. , Barry, W. & Hazan, V. (1989). Speech Input and Output Assessment. NY: J. Wiley.

ISBN 1-876346-48-5

ISBN 1-876346-48-5