Scoring Oral Test Responses by Computer

Computer Scoring of Spoken Responses

Jared C. Bernstein

Scoring Oral Test Responses by Computer
[or Computer Scoring of Spoken Responses]

Jared C. Bernstein
Stanford.University
jared.bernstein@stanford.edu

Scoring oral test responses by computer is the estimation of spoken language ability or its component skills by the operation of a computer on one or more spoken responses that are presented within an oral language test. This estimation process will be referred to here as *computer scoring of spoken responses*, or CSSR. As shown in Figure 1, automatic scoring is one component in a system for computer-based testing (CBT) that may present, record, and distribute spoken response material. CSSR refers to the automatic analysis and scoring of the responses that have typically been collected through such a CBT system.

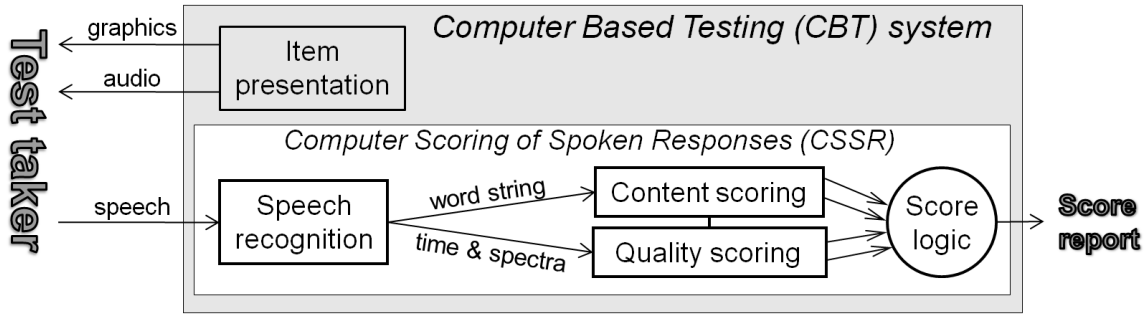

Figure 1. Schematic view of CSSR within a CBT system.

Figure 1 also shows the basic processing in CSSR within a CBT system. A speech signal is picked up and recorded, then processed by a speech recognition system, the output of which is sent to scoring modules, which return component scores that are combined according to a score-integration logic, then displayed in a score report. As of 2011, automatic oral response scoring may focus on the content of the speech, which may include its turn structure, pragmatic force, linguistic form and lexical content. Scoring may also focus on qualities of the speech itself such as its fluency and pronunciation, or it may combine content and quality aspects into a more general estimate of speaking ability.
Current CSSR systems estimate speaking ability by combining measures of linguistic and lexical structures with measures of fluency and pronunciation, returning scores with high consistency and acceptable accuracy, although the construct validity of some tasks and scoring methods has been questioned.

[A] Why use CSSR?

Human scoring of speaking has been shown to be valid for some purposes, but reliable human scoring is difficult to establish and maintain, even over a coarse 6-level scale and even when well trained raters use carefully constructed scoring rubrics. Score consistency is especially difficult across different language environments, across items, and over time, in part because there is a tendency for human raters in any judgment task to drift or shift to fit the range and distribution of exemplars in a sample (Parducci, 1995, chap. 6). Counteracting natural human tendencies that cause scale-shift within and across human judges requires effective training and equating procedures that can add cost to human-scored speaking tests.

Automatic scoring of spoken performances addresses some of the problems of human scoring. For example, permanent equilibration is possible for a standard distribution of skill levels across samples of speakers of different first languages. Automatic scoring can be calibrated once and reproduced on as many machines as needed to score millions of exams, and then score new material the same way at any later time. These advantages may also carry with them lower costs to the test taker and much shorter scoring delays.

[A] History of CSSR.

Computer scoring of spoken responses applies techniques from several more basic (and overlapping) fields including automatic speech recognition (ASR), spoken language processing, computational linguistics, and statistical pattern recognition (Jurafsky & Martin, 2009;). ASR forms the core of a CSSR system. The earliest ASR systems were implemented in analog hardware in the 1950s, and the first digital implementations of ASR appeared in the late 1960s and were first commercially viable in the 1990s in telephony. Eskenazi (2009) reviews the development of educational applications of ASR as it started appearing in the 1990s. Aist (1999) and Ehsani & Knodt (1998) present even more time depth, reviewing visual feedback systems from the 1970s and 1980s that were used for language learning. Nickerson et al. (1975) present a real-time feedback system for use in deaf education that ran in real-time on hybrid analog/digital processors and displays.

[A] Methods.

Although CSSR is often applied in assessment tasks that require both listening and speaking, we limit the scope here to estimating speaking ability only, disregarding the mode of item presentation. Accurate CSSR depends on many factors, but important elements include the size of the unit to score, the predictability of the spoken material, and the number of independent measurements combined in estimating the score. Some applications of CSSR, e.g. error detection in an interactive pronunciation tutoring system, are inherently more difficult than others. Recognition of the correct answer read aloud from a multiple choice list, for example, is an easy task for ASR. As is common in engineering, the nature of the application guides the selection of method.

Pronunciation tutors are considered important applications for foreign language instruction because spoken interaction with expert pronunciation feedback may not be easily available to learners. A learner of English or an instructional designer may expect that the assessment component of the machine should be able to 'hear' a sentence, then accurately detect the mispronounced segments (phones) and provide tutor-style guidance to encourage correct production. This would rely on quality scoring, as shown in Figure 1, applied to a sequence of recognized segments, each of which is a short event with 5 – 25 acoustic observations. In the error detection scenario, a speech recognizer produces its best estimate of the spoken words and submits a time-aligned transcription of the phones in the spoken response to the quality scoring module. So, for example, in a 1 second utterance of "Sweep the kitchen", the first subprocess inside the ASR module analyzes the speech signal into a sequence of 100 spectra, and assigns the first 9 spectra to [s] and the next 7 spectra to [w] and the next 14 spectra to [i], and so forth. Thus, each phone in each word has an associated sequence of spectra, from which the pronunciation tutor is expected to make a judgment whether or not these spectra should count as a good exemplar of the recognized phone.

The most common and basic approach to scoring pronunciation quality is to combine a spectral likelihood ratio (sometimes used as a confidence measure) with the duration of the phone (which is just the number of spectra assigned to that phone from the sequence of spectra in an utterance). The likelihood ratio highlights the relative compatibility of the observed spectra with the expected acoustic properties of the recognized phone that has been aligned with these spectra. The calculated likelihood ratio is of the form:

given a series of observed spectra and a recognized phone,

$$\text{pronunciation score} = \frac{\text{probability of these spectra given the recognized phone}}{\text{average probability of these spectra given any other phone}}$$

This ratio effectively normalizes the score for channel distortion and voice quality, as both the upper and lower term are affected by these factors. Similar methods are described by Franco et al. (2010) as applied to Spanish spoken by English speakers; Strik et al. (2009) describe and compare related methods for error detection in a two-way distinction between [k] and [x] in non-native Dutch.

Again, an important consideration is the total duration of the sample to be scored; and if one is picking out problem segments in a single utterance, the nature of the segment matters too. Franco et al. (2010) found that for most segment types in Spanish, human-human agreement on the "nativeness" of a single production of a phone yields kappa values below 0.3. That is, in many cases native listeners do not agree about the acceptability of particular instances of phones, although they may well have high confidence in their judgments.

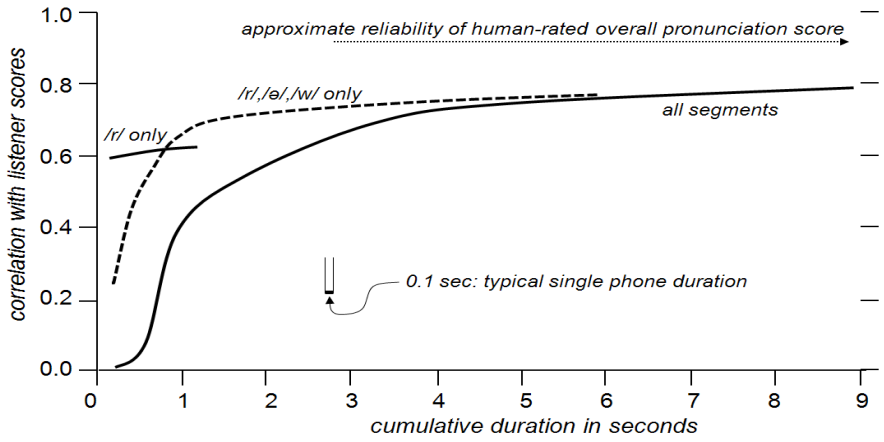

Figure 2. For three classes of phones, correlation between listeners' overall human rated pronunciation scores and corresponding CSSR score as a function of the cumulative signal duration that CSSR operates on. N = 50 test takers.

Figure 2 shows how machine scores improve with the increasing duration of the stretch of spoken material that is scored. The data shown is for a sample of 50 adult learners of English with a variety of first languages. Three phone-classes are shown: the short solid line shows all and only /r/ segments; the dashed line shows all /r/, /ə/, and /w/ segments, and the long solid line show all segments taken together. For each class, for a set of 50 test-takers speaking in response to an English speaking test, the line shows the value of the correlation between a machine pronunciation score and a reliable composite human score of overall pronunciation for the speakers based on many recordings of each speaker rated by several judges.

Taking 0.1 seconds as a typical phonetic segment length, it can be seen from Figure 2 that a machine can produce speaker pronunciation estimates from two or three tokens of /r/ that correlate with a speaker's overall human-rated pronunciation score with a value of about 0.6, while if all segments are considered, it would likely take about 24 phones to produce a similar correlation with human ratings. In summary, for single pronounced tokens of a phone in context, accurate automatic detection and diagnosis of errors is a challenging technical task that attempts to match a human judgment that Franco et al. (2010) have shown to exhibit generally low agreement between pairs of human raters.

Summative assessment of speaking ability is a more general task that typically includes pronunciation scoring as a component, but can return more accurate and reliable scores because it may reasonably be based on one or more minutes of speech and use several independent sources of information from a set of spoken responses.

Referring back to Figure 1, a summative assessment can be extracted from a set of spoken responses by combining estimates from several different kinds of evidence that can be combined into an overall estimate. Because cultural expectations permit an assessment to last more than an hour and to elicit many short samples and/or several long samples of speech, capturing 5 or 10 minutes of spoken material for computer scoring is culturally acceptable. The "all segments" line in Figure 2

suggests that even with only 2 or 3 minutes of speech, an overall pronunciation score can be accurately estimated.

However, there may be four or more scores coming from the content and quality analyzers, as schematized in Figure 1, which can be combined to produce a summative score that correlates closely with human summative judgments. For example, summing just two automatically extracted measures of linguistic content (syntactic structure and vocabulary) and two automatic measures of production quality (fluency and pronunciation) from 3-4 minutes of speech produces very high correlations (r = .97) with sums of corresponding human ratings of the same spoken materials (Bernstein & Cheng, 2007). In this implementation, the individual automatic sub-scores correlate with the human scores with values of {0.93, 0.94, 0.89, 0.89}. In part, because the cross-correlation of the automatic sub-scores averages about 0.72, the machine-human correlation of the summed scores reaches 0.97 for a human summed score with a reliability of 0.98.

Note that in 3 or 4 minutes of spoken material responding to 20 – 60 spoken prompts, a test-taker may produce 2000 or more phones, 40 or 100 phrases or clauses that perhaps include 40 or as many as 500 words. Given a relatively accurate scoring of these semi-independent content and quality aspects of the person's speech, an accurate overall score can be reported.

[A] Explanation

Although CSSR techniques will produce only a rough estimate of a person's oral proficiency from that person's production of one unit of spoken language, when many of these noisy estimates are summed, the result asymptotes toward the human summative rating. CSSR works best on constrained tasks for two reasons. First, the component content and quality estimates are constrained by the accuracy of the underlying speech recognition technology, and predictability of the spoken material is a main determinant of ASR accuracy. Second, when tasks elicit relatively predictable spoken responses, test developers can build much tighter performance models from samples of speakers of various levels. That is, within a constrained context, the variance of many acoustic, lexical, and rhythmic units is reduced, especially in high-proficiency speech, and the variance of these units then can be used as a yardstick in scoring other spoken responses.

In the work presented by Bernstein & Cheng (2007), the performances scored are produced in real time – without any preparation time or note taking. Evidence from Hulstejn (2007) suggests that apparent automaticity is a key element in judgments of general speaking ability. If so, the unprepared nature of a task may bring out aspects of automaticity in performance that are particularly salient to listeners and interlocutors in real conversation.

[A] Challenges

As of 2011, CSSR techniques yield less reliable scores for spoken-answer tasks that elicit widely variable or unpredictable responses. This is partly because speech recognition relies strongly on 'top-down' processing, which is typically implemented in ASR and CSSR systems as a statistical language model, or LM. LMs are conceptually similar to expectation grammars in the applied linguistics literature (Oller, 1971), however LMs usually model word-word dependencies, thus mixing lexical and syntactic expectation. LMs, therefore, are quite topic sensitive, as human listening is under unfavorable noise conditions.

An example to clarify: a language model based on N-grams uses the previous words to predict the next word, based on probability estimates derived from transcribed speech in response to the same assessment prompt. Thus, given "raise income tax" a LM might strongly predict "rates" to be the next word because "raise income tax rates" is a commonly occurring 4-gram (4-word sequence) in response to a given prompt. The same kind of LM prediction could be made for a 3-gram like "income tax rates", or a word pair like "tax rates", and the unigram probability for the word "rates" (without any information about the preceding words) is helpful recognizing speech.

Although ASR takes advantage of some of the top-down expectation that human listeners use, many features of ASR behavior are not similar to human performance. ASR systems do not produce errors in patterns similar to human listeners' patterns of errors, and systems need to be designed with this difference in mind. One notable difference is that human listeners are quite good at isolating a single unknown word in a sequence, while ASR technology does not return reliable "confidence" scores for single words. For example, a reasonably common human exchange might be:

Speaker 1: *The [unintelligible] one didn't show up.*
Speaker 2: *The **which** one didn't show up?*

The listener (speaker 2) was sure of all the words but one and then asked for that one. Similar to pronunciation scoring, ASR lexical decision is not very reliable on a single word, the processes gain accuracy over sequences of several words. In the ASR literature, this is called the "Out-of-Vocabulary" (OOV) problem, and no general solution is yet available that enables automatic dialogue turns like Speaker 2's question.

[A] Future Directions

There may be broadly two directions of future enhancement for CSSR: one is analysis for sharper diagnostic reporting; and the other is re-integration of speaking skills with cognitive and social skills to reconstitute the traditional communicative spoken language construct from more elemental constructs. First, and most simply, many of the tasks that elicit spoken responses use spoken prompts, so one can presume that response performances are partly dependent on listening as well as speaking. It would be good to test speaking with tasks that elicit the kind of spontaneous speech found in natural dialog without conflating the speaking scores with listening skills. Secondly, there are many perceptible qualities of the speech

itself that are not reliably scoreable from unconstrained samples. For example, indexical properties like friendliness, or affective properties like apparent mood. Lastly, it would be useful if articulation and prosody in speech could be analyzed such that the most effective route toward phonological improvement was known for a given speaker. The reintegration of CSSR with those cognitive and social aspects of live communication will depend on ongoing development of technologies that extract features of declarative and social meaning that are still nascent in applied linguistics and in spoken language engineering.

[Cross-references]

CALL, Spoken language proficiency, fluency, expectation grammar, pronunciation, listening, computer scoring of essays, speech recognition

[References]

Gregory Aist (1999) Speech recognition in computer assisted language learning. In K. C. Cameron (ed.), Computer Assisted Language Learning (CALL): Media, Design, and Applications. Lisse: Swets & Zeitlinger.

Jared Bernstein, Alistair van Moere & Jian Cheng (2010) "Validating Automated Speaking Tests" *Language Testing*. 27(3) pp. 355-377.

Jared Bernstein & Jian Cheng (2007) "Logic, Operation and Validation of a Spoken English Test," a chapter in V.M. Holland & F.P. Fisher, (Eds.) *Speech Technologies for Language Learning*. NY: Routledge, 174-194.

Farzad Ehsani and Eva Knodt. Speech technology in computer-aided language learning: Strengths and limitations of a new CALL paradigm. *Language Learning & Technology* 2(1): 54-73. 1998.

Maxine Eskenazi (2009) "An overview of spoken language technology for education" *Speech Communication* v.51 pp. 832-844.

Horacio Franco, Harry Bratt, Romain Rossier, Ramana Rao, Elizabeth Shriberg, Victor Abrash & Kristin Precoda (2010) "EduSpeak®: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications" *Language Testing* 27(3) pp. 401-418.

J. H. Hulstijn **(**2007) Psycholinguistic perspectives on second language acquisition. In Cummins, J. and Davison, C., editors, *The international handbook on English language teaching*. Norwell, MA: Springer, 701-13.

Daniel Jurafsky & James H. Martin (2009) Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition, 2nd ed. Upper Saddle River, N.J.: Pearson Prentice Hall.

John Oller (1971) *Coding information in natural languages*. The Hague, Mouton

Allen Parducci (1995) Happiness, Pleasure, and Judgment: The Contextual Theory and Its Applications. Mahwah, N.J.: Lawrence Erlbaum Associates,

Helmer Strik, Khiet Truong, Febe de Wet, Catia Cucchiarini (2009) Comparing different approaches for automatic pronunciation error detection. *Speech Communication,* 51, pp. 845-852.

C.L.J. de Bot (1980) The role of feedback and feed-forward in the teaching of pronunciation - an overview. *System* 8(1):35-45. 1980.

[Suggested Readings]

J. Bernstein & H. Franco (1996) "Speech Recognition by Computer," in *Principles of Experimental Phonetics*, (N. Lass, ed.), Mosby, St. Louis.

Christopher M. Bishop (2006) Pattern Recognition and Machine Learning Springer, New York.

Steve Young (1996). "Large Vocabulary Continuous Speech Recognition." IEEE Signal Processing Magazine 13(5): 45-57.

Grant Henning (1983) Oral Proficiency Testing: Comparative validities of interview, imitation, and completion methods. Language Learning 33(3) pp. 315–332.

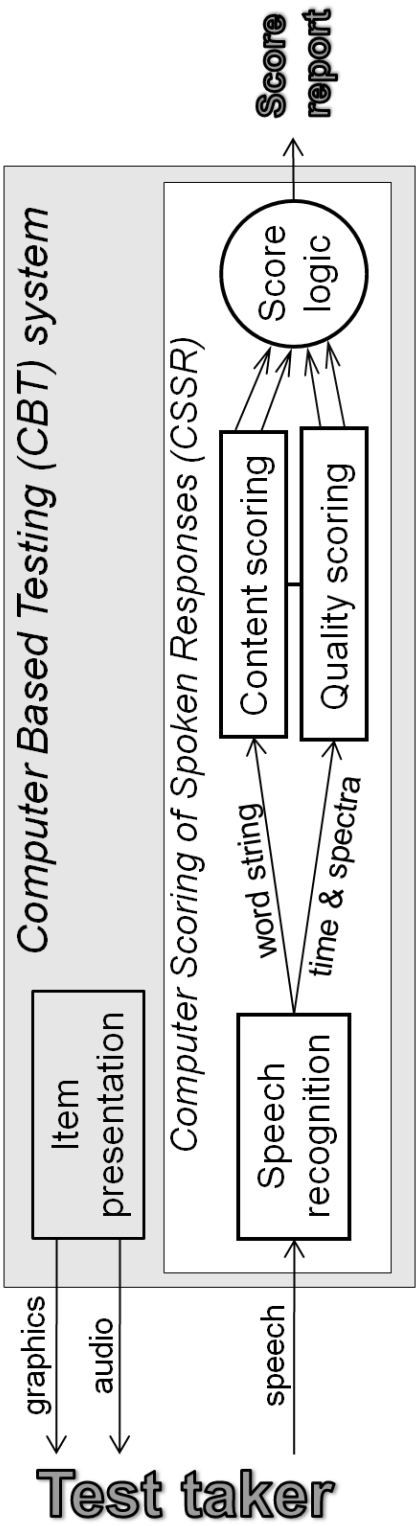

Figure 1. Schematic view of CSSR within a CBT system.

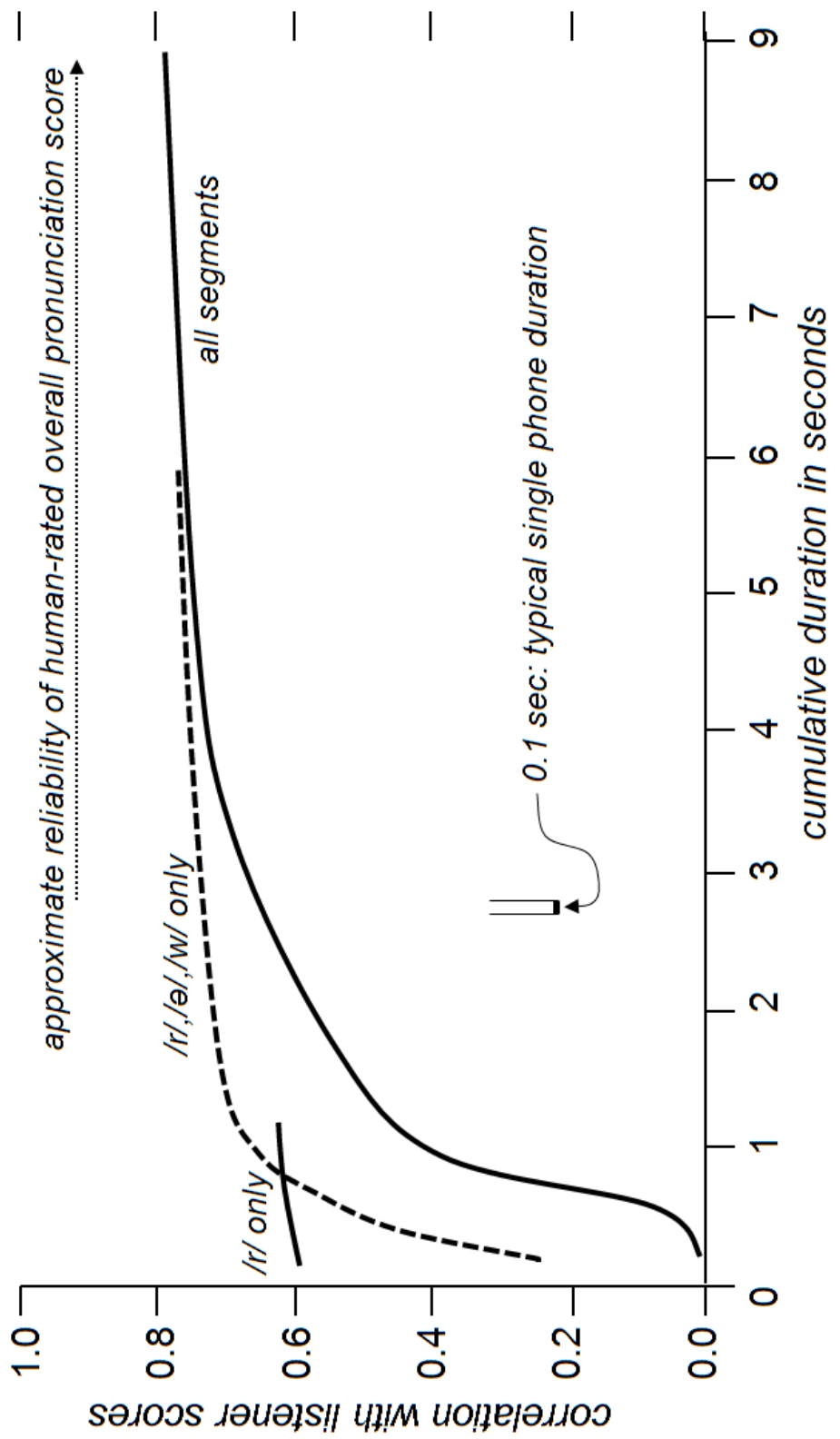

Figure 2. For three classes of phones, correlation between listeners' overall human rated pronunciation scores and corresponding CSSR score as a function of the cumulative signal duration that CSSR operates on. N = 50 test takers.