EVALUATING DICTATION TASK MEASURES FOR THE STUDY OF SPEECH PERCEPTION

Emily Felker, Mirjam Ernestus, Mirjam Broersma

Centre for Language Studies, Radboud University, Nijmegen, The Netherlands
e.felker@let.ru.nl, m.ernestus@let.ru.nl, m.broersma@let.ru.nl

ABSTRACT

This paper shows that the dictation task, a well-known testing instrument in language education, has untapped potential as a research tool for studying speech perception. We describe how transcriptions can be scored on measures of lexical, orthographic, phonological, and semantic similarity to target phrases to provide comprehensive information about accuracy at different processing levels. The former three measures are automatically extractable, increasing objectivity, and the middle two are gradient, providing finer-grained information than traditionally used. We evaluate the measures in an English dictation task featuring phonetically reduced continuous speech. Whereas the lexical and orthographic measures emphasize listeners' word identification difficulties, the phonological measure demonstrates that listeners can often still recover phonological features, and the semantic measure captures their ability to get the gist of the utterances. Correlational analyses and a discussion of practical and theoretical considerations show that combining multiple measures improves the dictation task's utility as a research tool.

Keywords: speech perception, non-native listening, dictation task, reduced pronunciation variants

1. INTRODUCTION

One of the most straightforward ways to test how accurately listeners can decode the acoustic speech signal into linguistic units, such as words, is to have them transcribe a stretch of speech. In the field of applied linguistics, this method is known as the dictation task, and we argue in this paper that the dictation task has untapped potential as a phonetics research tool for the study of speech perception.

In second language (L2) learning and teaching, the dictation task is widely used both as a pedagogical tool and as a testing instrument for listening skills [15, 16, 19, 23]. The dictation task is particularly relevant for training and evaluating perceptual processing abilities, such as phoneme recognition and lexical segmentation [7, 22]. Despite the ubiquity of the dictation task in language education, however, it has seen relatively little use in the field of phonetics, even though written transcriptions of speech are often used in the context of speech intelligibility research [12].

An important reason why dictation is underutilized in phonetics research may be that detailed scoring measures have yet to be developed. In applied linguistics, transcriptions in dictation tasks are usually scored for word- or phrase-level accuracy, with potential latitude given by human raters for misspellings [2, 21]. The percent of words correctly identified is also a typical scoring measure in the field of speech intelligibility testing [12]. However, examining only the proportion of words accurately transcribed does not differentiate completely wrong and more nearly right answers. Consider the utterance "my Friday night" spoken with the consonants not clearly articulated, which one listener transcribes as "my friend and I" and another as "my family" in the experiment we report. Both answers match the target phrase in exactly one word, but the former is a better phonological match. Binary measures like word error rate ignore finer distinctions between answers at the phonological level, such as how well listeners can recover the target words' phonetic features.

We propose that considerable information about listeners' perceptual abilities can be gained by scoring transcriptions with a broader range of measures that capture accuracy at different processing levels. Moreover, using automatically calculated measures increases scoring objectivity. Finally, complementing word-, letter-, and phoneme-based measures with a semantic accuracy measure provides insight into the communicative consequences of perceptual errors.

This paper demonstrates how a dictation task with more precise measures can be used to study speech perception. Specifically, we present four measures—lexical error rate, orthographic edit distance, phonological edit distance, and semantic error rate—and evaluate their usefulness when applied to a dictation study investigating how non-native listeners perceive casual speech with severe speech reductions.

Speech reductions, in which segments and even syllables are weakly articulated or altogether missing, are a hallmark of the casual speech register [6, 10]. While native (L1) listeners can easily process reduced words presented in context, e.g., [3, 9, 11], reductions often cause comprehension problems for non-native listeners, who tend to have less exposure to these pronunciation variants [1, 5].

We tested Dutch non-native and American English native listeners on a fill-in-the-blank dictation task with American English target phrases containing massive phonetic reductions, presented in sentential contexts. To evaluate the four dictation measures, we analyze how well they distinguish the listener groups, how performance differs across the measures, how the measures correlate with the non-natives' language proficiency and usage, and how the measures correlate with each other. Following these analyses, we discuss the measures' utility based on practical and theoretical considerations.

2. MEASURES

This section describes in detail the measures that we propose and evaluate. All measures yield scores between zero and one, with zero indicating a perfect match between a transcription and target phrase. For the first three measures, which are calculated programmatically, transcriptions are pre-processed to remove capitalization, punctuation, and extra spaces.

2.1. Lexical error rate

The traditional dictation scoring method (as described in, e.g., [2, 8]) involves calculating the lexical error rate, which is simply the proportion of words in the target phrase that are absent in the participant's transcription. For example, for the target phrase "She wants to be a police officer," the transcription "She is a police officer" receives a score of 0.43 (3/7 of target words missing). To avoid reliance on human judgments about the source or severity of spelling errors, words must be spelled correctly to count.

2.2. Orthographic edit distance

The orthographic edit distance is a measure of how accurately listeners perceived the sounds of the target phrase, using letters as a proxy for sounds. Compared to the lexical error rate, it gives more credit to imperfect transcriptions containing similar sets of letters in similar orders to those of the target phrases.

We implement the orthographic edit distance between the transcribed and target phrases as the two strings' Levenshtein distance: the minimum number of single-character edits, namely, insertions, deletions, or substitutions, required to transform one into the other [14]. For instance, to transform the transcription "my fright night" into the target phrase "my Friday night" requires minimally three substitutions: replacing the last three characters of "fright." To normalize the edit distance to lie between zero and one, we divide it by the number of characters in the longer phrase, as this length represents the maximum possible distance between two items.

2.3. Phonological edit distance

The phonological edit distance, based on methods used to phonetically measure dialect distance [18], provides a closer estimate of how well participants were able to recover the phonemes, and even the specific phonological features, of the target phrase. It is based on the same principle as the orthographic edit distance, but it uses phonemes rather than letters and captures the insight that some phonemes are more similar to each other than others. Thus, replacing a /t/ with a /d/ incurs less penalty than replacing it with /n/ because fewer features change.

To calculate the phonological edit distance, the target phrase and transcribed phrases are first converted from Latin letters to IPA characters using a word-to-phoneme dictionary, such as the CMU Pronouncing Dictionary for English [3]. Words not in the dictionary, such as misspellings or uncommon names, are converted to IPA characters using a grapheme-to-phoneme engine, such as g2p_en [20].

Once the IPA transcription of the target phrase and participant transcription are obtained, the phonological edit distance is calculated using the weighted feature edit distance of the PanPhon library [17], which represents every IPA segment as a vector of phonological features and weights the costs of feature edits differently depending on their class and subjective variability. To normalize the phonological edit distance to lie between zero and one, we then divide it by the weighted feature edit distance between an empty string and the longer of the two strings, as this represents the maximum possible weighted feature edit distance between them.

2.4. Semantic error rate

The semantic error rate gauges how well a transcription conveys the broad meaning of a target phrase. The target phrase is broken down into its key conceptual elements, defined by the phrase's open-class lemmas and personal pronouns. For example, for the target phrase "since I stopped going to the gym," the key elements are *I, stop, go,* and *gym.* We score the participant transcriptions manually by calculating the proportion of key concepts from the target phrase that are missing from the transcribed phrase, interpreting any spelling errors generously. For a noun-phrase concept to count as present, it must fill the correct thematic role in the sentence, and for a verbal concept to count, the verb's polarity (positive/negative), but not tense or aspect, has to match that of the target phrase. Thus, for the example given above, the transcription "since I'm going to the gym" receives a score of 0.25 (1/4 key concepts [*stop*] missing), and "since I went to Germany" scores 0.50 (2/4 key concepts [*stop, gym*] missing).

3. METHODS

To evaluate the four dictation measures, we implemented them in a dictation task with reduced speech given to non-native and native listeners.

3.1. Participants

The participants were 116 native Dutch speakers (mean age = 21.7 years, SD = 2.8) with advanced L2 English proficiency and 25 native American English speakers (mean age = 24.1 years, SD = 2.7).

3.2. Materials

The dictation task comprised eight fragments of spontaneous English speech produced by a female American from Arizona in an informal dialogue. Each fragment was one or two sentences long and contained highly reduced productions. For each fragment, a critical sequence of consecutive words was selected to be the fill-in-the-blank target phrase for participants to transcribe. The target phrases and their broad phonetic transcriptions, illustrating massive reductions, are listed in Table 1.

Table 1: Dictation task target phrases.

Target Phrase	Transcription of Phrase as Spoken
I didn't really know that, but I need to take it to graduate	aɪ ɪn ɹɪli noʊ ðæːt bət aɪ niə teɪkɪtə gɹædʒuɛt
since I stopped going to the gym	saɪ stɑp gowɪnə ʤɪm
She wants to be a police officer	ʃɑns i pəlːis ɔvəsəɹ
I was thinking of just applying to jobs in San Diego	aɪz θɪŋə ʤɪst əplaɪnə ʤɑbz ɪn sæn dieɪgoʊ
My Friday night	mʌ fɪɛ̃
she's gonna let me know for sure today	ʃiz gənə lɛt mi noʊ fʊɹ ʃʊɹ tədeɪ
'cause that way we can be together	ksæ weɪ i knː bi dəgɛðəɹ
I told him that I was thinking about going to	aɪ toʊld ɪm ðæt aɪz θɪŋmə goʊnə

3.3. Procedures

The dictation task was presented in the form of an online, self-paced Qualtrics survey with one audio fragment per page, which could be replayed as often as desired. On each page, a partial transcription of the recording was provided, and the participants' task was to listen to the recording and to type in the missing words in the blank.

After the dictation task, all Dutch participants completed a language background questionnaire, and a subset (n = 45) took the LexTale [13], a measure of their English vocabulary knowledge.

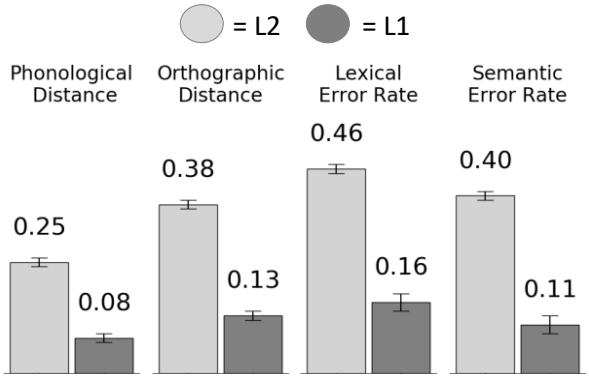
3.4. Data pre-processing

To make the transcriptions comparable to each other and to the target phrases for automatic scoring, we processed the data so that for each contraction in the target phrases, all versions of that contraction in the transcriptions were converted to the same form (e.g., "because", "'cause", and "cuz" were all mapped onto "'cause."). As the Dutch listeners often wrote compound nouns as one word (e.g., "policeofficer" for "police officer"), we separated these forms into two words to avoid penalizing this error pattern relating to orthography rather than speech perception.

4. RESULTS

The four dictation measures clearly distinguish the transcriptions of non-native and native listeners. As shown in Figure 1, the Dutch listeners performed significantly worse than the American listeners on all measures (phonological distance (t(415.13) = 16.58), orthographic distance (t(343.27) = 17.41), lexical error rate (t(329.99) = 16.53), and semantic error rate (t(297.60) = 12.73); all p's < 0.001).

Figure 1: Mean dictation scores for the set of transcriptions made by Dutch (L2) listeners and American (L1) listeners, with bar height representing the amount of error and error bars representing the standard error of the mean.



The four measures also show that participants' answers incorporate more phonological and semantic information than lexical error rate alone might suggest. Transcriptions were most different from the target phrases in lexical error rate, which was higher than orthographic distance and semantic error rate (t(140) = 21.22 and t(140) = 13.25 respectively, both p's < 0.001). Transcriptions were closest to the target phrases in phonological distance, as this score was lower than the orthographic distance, semantic error rate, and lexical error rate (t(140) = 35.95, t(140) =

17.02, and $t(140) = 34.08$, respectively, all p's < 0.001). The scores for the measures of orthographic distance and semantic error rate were equivalent ($t(140) = 1.80$, $p = 0.07$).

For each of the four measures, an overall dictation score was calculated for each participant by averaging across the eight items. Table 2 presents correlations between the Dutch listeners' four overall dictation scores and their self-rated English language proficiency in speaking, listening, reading, and writing; their average weekly hours of English listening and speaking; and their LexTale scores.

Table 2: Correlations between Dutch listeners' dictation scores on the four measures and language background questionnaire variables (* $p < 0.05$, ** $p < 0.0018$, the Bonferroni-corrected alpha).

		PHON	ORTH	LEX	SEM
Self-Rated Proficiency	Speaking	-.19*	-.23*	-.30**	-.27*
	Listening	-.24*	-.25*	-.31**	-.32**
	Writing	-.19*	-.23*	-.27*	-.23*
	Reading	-.24*	-.29**	-.32**	-.30**
Weekly Hours	Speaking	-.03	-.09	-.07	-.11
	Listening	-.12	-.19*	-.22*	-.29**
	LexTale	-.36*	-.35*	-.44*	-.40*

As shown in Table 3, the four measures have medium to high correlations with each other. The orthographic distance correlates highly with both the lexical error rate and phonological distance; this follows from the fact that they all depend on the specific letter sequences in the transcription for their calculation. As to be expected, the lowest correlation is between the semantic and phonological measures.

Table 3: Matrix of Pearson correlation coefficients for the transcriptions' scores on the four measures.

	PHON	ORTH	LEX	SEM
PHON	1.00	.90	.77	.67
ORTH	.90	1.00	.92	.79
LEX	.77	.92	1.00	.87
SEM	.67	.79	.87	1.00

5. DISCUSSION

This paper demonstrated how four measures targeting accuracy at different levels, with different degrees of granularity involved in their calculation, can be used to score dictation data, thereby increasing the amount of information that dictation tasks can yield for speech perception research.

From a practical standpoint, the easiest measures to implement are lexical error rate and orthographic edit distance as they are both calculated automatically and do not need an external data source. Phonological edit distance, while automatically calculated, requires a dictionary for converting words or graphemes to phonemes, which may be hard to find for some languages. Semantic error rate, relying on a human rater, is more time-consuming, subjective, and error-prone. It could conceivably be automated with the right language model, but the time investment may be prohibitively high except for very large data sets.

Given the four measures' high intercorrelations, using a subset of them can still be informative. For instance, the lexical and orthographic measures, both based on the degree of matching between the letter sequences in the transcription and target phrase, provide almost the same information except that the former is binary (a word matches exactly or not at all) while the latter is gradient (similarly spelled words are less penalized). Thus, unless spelling accuracy is of additional theoretical interest, the orthographic edit distance could be used by itself as it already provides a very good estimate of word recognition ability.

Combining the phonological edit distance and the semantic error rate, which themselves have a lower intercorrelation, sheds light on different aspects of performance: how accurately phonological features were recovered from the acoustic signal and how well the meaning of the utterances was comprehended. As listeners may employ different transcription strategies, prioritizing either bottom-up or top-down information, using both measures paints a more complete picture of their abilities.

Using writing as a proxy for speech perception comes with some caveats. For non-native listeners, whose sound-to-orthography mappings can differ from those of native listeners, dictation performance may be less informative about their actual sound representations. Also, since listeners tend to write real words even when they are not a perfect match for the perceived input, errors in letter sequences unrelated to the sounds actually perceived can arise. Still, the phonological distance measure allows the dictation task to evaluate phoneme perception, even for English with its notoriously irregular spelling system.

Overall, the combination of lexical, orthographic, phonological, and semantic similarity measures provides richer information than the traditional word error rate about what linguistic units listeners recover from the speech input. While we have shown how these measures can be used to analyze transcriptions of reduced speech, they are also suitable for any research on speech perception in difficult conditions, whether these involve properties of the speech itself, background noise, or listener characteristics.

5. ACKNOWLEDGMENTS

This research was supported by an NWO Vidi grant awarded to the third author. We thank Natasha Warner for providing us with the speech materials.

6. REFERENCES

[1] Brand, S., Ernestus, M. 2018. Listeners' processing of a given reduced word pronunciation variant directly reflects their exposure to this variant: evidence from native listeners and learners of French. *Quarterly Journal of Experimental Psychology, 71*, 1240-1259.

[2] Buck, G. 2001. *Assessing Listening*. Cambridge: Cambridge University Press.

[3] CMU Pronouncing Dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict (Version 0.7b).

[4] Ernestus, M., Baayen, H., Schreuder, R. 2002. The recognition of reduced word forms. *Brain and Language*, *81*, 162-173.

[5] Ernestus, M., Dikmans, M., Giezenaar, G. 2017. Advanced second language learners experience difficulties processing reduced word pronunciation variants. *Dutch Journal of Applied Linguistics*, 6(1), 1-20.

[6] Ernestus, M., Warner, N. 2011. An introduction to reduced pronunciation variants. *Journal of Phonetics*, *39*(SI), 253-260.

[7] Field, J. 2003. Promoting perception: lexical segmentation in L2 listening. *ELT Journal*, *57*(4), 325-334.

[8] Irvine, P., Altai, P., Oller Jr., J. R. 1974. Cloze, dictation, and the Test of English as a Foreign Language. *Language Learning*, *24*(2), 245-252.

[9] Janse, E., Ernestus, M. 2011. The roles of bottom-up and top-down information in the recognition of reduced speech: Evidence from listeners with normal and impaired hearing. *Journal of Phonetics*, *39*(3), 330-343.

[10] Johnson, K. 2004. Massive reduction in conversational American English. In *Spontaneous speech: Data and analysis. Proc. 1st Session of the 10th International Symposium*. 29-54. Tokyo, Japan: The National International Institute for Japanese Language.

[11] Kemps, R., Ernestus, M., Schreuder, R., Baayen, H. 2004. Processing reduced word forms: The suffix restoration effect. *Brain and Language*, *90*, 117-127.

[12] Kent, R. D., ed. *Intelligibility in Speech Disorders: Theory, measurement, and management*. 1992. Amsterdam: John Benjamins.

[13] Lemhöfer, K., Broersma, M. 2012. Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior research methods*, *44*(2), 325-343.

[14] Levenshtein, V. I. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics-Doklandy*, *10*(8), 707-710.

[15] Matthews, J., O'Toole, J. M. 2015. Investigating an innovative computer application to improve L2 word recognition from speech. *Computer Assisted Language Learning*, *28*(4), 364-382.

[16] Morris, S. 1983. Dictation—a technique in need of reappraisal. *ELT Journal*, *37*(2), 121-126.

[17] Mortensen, D. R., Littell, P., Bharadwaj, A., Goyal, K., Dyer, C., Levin, L. 2016. PanPhon: A resource for mapping IPA segments to articulatory feature vectors. *Proc. COLING 2016: Technical Papers* Osaka, 3475–3484.

[18] Nerbonne, J., Heeringa, W. 1997. Measuring dialect distance phonetically. In *Computational Phonology: 3rd Meeting of the ACL Special Interest Group in Computational Phonology*.

[19] Oller, J. W., Streiff, V. 1975. Dictation: A test of grammar-based expectancies. *ELT Journal*, *30*(1), 25-36.

[20] Park, K., Kim, J. 2018. g2p_en. https://github.com/Kyubyong/g2p (Version 1.0.0).

[21] Savignon, S. J. 1982. Dictation as a measure of communicative competence in French as a second language. *Language Learning*, *32*(1), 33-47.

[22] Siegel, J., Siegel, A. 2015. Getting to the bottom of L2 listening instruction: Making a case for bottom-up activities. *Studies in Second Language Learning and Teaching*, *5*(4), 637-662.

[23] Stansfield, C. W. 1985. A history of dictation in foreign language teaching and testing. *The Modern Language Journal*, *69*(2), 121-128.