JSLHR

Research Note

Automating Error Frequency Analysis via the Phonemic Edit Distance Ratio

Michael Smith,[a] Kevin T. Cunningham,[a] and Katarina L. Haley[a]

Purpose: Many communication disorders result in speech sound errors that listeners perceive as phonemic errors. Unfortunately, manual methods for calculating phonemic error frequency are prohibitively time consuming to use in large-scale research and busy clinical settings. The purpose of this study was to validate an automated analysis based on a string metric—the unweighted Levenshtein edit distance—to express phonemic error frequency after left hemisphere stroke.
Method: Audio-recorded speech samples from 65 speakers who repeated single words after a clinician were transcribed phonetically. By comparing these transcriptions to the target, we calculated the percent segments with a combination of phonemic substitutions, additions, and omissions and derived the phonemic edit distance ratio, which theoretically corresponds to percent segments with these phonemic errors.
Results: Convergent validity between the manually calculated error frequency and the automated edit distance ratio was excellent, as demonstrated by nearly perfect correlations and negligible mean differences. The results were replicated across 2 speech samples and 2 computation applications.
Conclusions: The phonemic edit distance ratio is well suited to estimate phonemic error rate and proves itself for immediate application to research and clinical practice. It can be calculated from any paired strings of transcription symbols and requires no additional steps, algorithms, or judgment concerning alignment between target and production. We recommend it as a valid, simple, and efficient substitute for manual calculation of error frequency.

In the context of communication disorders that affect speech production, sound error frequency is often linked to the severity of the condition. Error frequency measures have clinical value across a wide variety of patient populations, including children with phonological disorders, childhood apraxia of speech (AOS), craniofacial disorders, and a host of acquired and developmental neurogenic communication disorders. In our laboratory, we use error frequency as a severity index for acquired AOS and aphasia with phonemic paraphasia (APP). Disorder severity can vary dramatically, particularly for AOS, and sound error frequency therefore assumes a broad range of values (Haley, Jacks, Richardson, & Wambaugh, 2017). Whereas subjective impressions provide preliminary severity estimates, precise measurements are preferable for comparison and documentation purposes alike.

Phonemic error frequency in AOS and APP can be estimated through a variety of strategies that index the difference between a series of targets and a series of productions. As a basic clinical estimation, accuracy may be expressed as the simple percentage of words in a reasonably challenging speech sample that a speaker produces without phonemic or phonetic errors (Duffy et al., 2017; Haley, Jacks, de Riesthal, Abou-Khalil, & Roth, 2012). Whereas the metric excels in providing a quick snapshot of the severity of the speech impairment, it falls short in its lack of resolution.

Precision increases with whole-word phonetic transcription, but once transcribed, the summary is not straightforward. Phoneme-level analysis has been used to derive metrics such as percent consonants or phonemes correct (Shriberg, Austin, Lewis, McSweeny, & Wilson, 1997; Shriberg & Kwiatkowski, 1982) and the proportion of omission, addition, and substitution errors that a person makes when speaking (Bislick, McNeil, Spencer, Yorkston, & Kendall, 2017; Haley, Bays, & Ohde, 2001; Haley et al., 2012; Miller, 1995; Odell, McNeil, Rosenbek, & Hunter, 1990, 1991; Scholl, McCabe, Heard, & Ballard, 2018). These manual phonemic error counts are the unfortunate antithesis of word accuracy in their inefficiency—they are time consuming, attention demanding, and potentially vulnerable to coding error. Additionally, they depend on coding decisions that require consideration of the phonetic

[a]Division of Speech and Hearing Sciences, Department of Allied Health Sciences, University of North Carolina, Chapel Hill
Correspondence to Katarina L. Haley: Katarina_Haley@med.unc.edu

Editor-in-Chief: Julie Liss
Editor: Stephanie Borrie
Received October 23, 2018
Revision received January 21, 2019
Accepted April 3, 2019
https://doi.org/10.1044/2019_JSLHR-S-18-0423

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

context and the phonetic relationships between alternative segments. This complexity increases the risk of error and reduces utility for settings where analysis time is limited. In our case, inefficient quantification is prohibitive in large-scale studies that involve hundreds of speech samples. Application is equally challenging in clinical settings, where productivity requirements limit time to process assessment results.

One of the main challenges with manual error coding is that it is not always clear how to properly segment utterances and align equivalent segment strings with one another. This problem is particularly evident for speakers who make numerous errors (Haley, Cunningham, Eaton, & Jacks, 2018; Haley, Jacks, & Cunningham, 2013). To demonstrate segmentation challenges, consider the following example (see Table 1). The target is the word *octopus* (/ɑktəpʊs/). The speaker produces /təbɛn/, and Coder A decides that the production has no phonetic relationship to the target, thereby coding all of its segments as substitutions and the two missing phonemes as omissions, resulting in seven total phonemic errors. Coder B, on the other hand, recognizes the common syllable /tə/, as well as the shared characteristics of /pʊs/ and /bɛn/. In these syllables, the first sounds are bilabial stops, the second sounds are produced with approximately the same tongue height, and the third sounds are alveolar continuants. As a result, Coder B links this production to the target and consequently considers only the final three segments to be substitutions, reducing the total number of errors to five. These instances in which some correspondence can be made between target and production have the potential to create unnecessary discrepancy and introduce ample opportunity for human judgment differences and errors. Whereas algorithms are available to achieve rule-based phoneme alignment based on maximal phonetic similarity (Covington, 1996), a simpler solution is preferable for the basic purpose of estimating phonemic error frequency. Ideally, phonetically trained clinicians and coders should be able to allocate their effort to the transcription process itself and rely on automated processes for segment alignment and error quantification.

Fundamentally, phonetic transcriptions are character strings, and error coding requires that these strings be compared to a corresponding character string for the target word. The edit distance string metric is based on an algorithm that quantifies the difference between two such strings. In its simplest form, it is known as the Levenshtein distance (Levenshtein, 1966). This algorithm generates a sum of the fewest number of operations that can transform one string into another. The operations are defined as omissions, additions, and substitutions—the same phonemic error categories we track manually when we calculate phonemic error frequency.

Because the edit distance is suitable for strings of any form, such as text, numbers, shapes, and item clusters, it has been implemented through many different algorithms and employed extensively in the natural sciences. Applications have also been productive in the area of speech, but primarily within computational linguistics (Gooskens & Heeringa, 2004; Yang & Castro, 2008) and automatic speech recognition (Kruskal & Liberman, 1999; Schluter, Nussbaum-Thom, & Ney, 2011). To answer questions about phonetic variation and similarity, custom weights or costs have usually been assigned to the operations based on phonetic features and prevalence in a referenced speech corpus. With weights appropriate to the sample and purpose, the magnitude of similarity between alternative productions can potentially be expressed with good precision. Despite the demonstrated power of edit distance metrics in other fields, comparatively few applications have been used to analyze speech disorders. A recent study used a dynamic cost model from computational linguistics to answer a question about speech development in a communication disorder (Faes, Gillis, & Gillis, 2016). The research team used a weighted Levenshtein distance to quantify similarity between words produced by children with cochlear implants and a reference standard and then compared the magnitude of this similarity to the corresponding value for age-matched, normally hearing peers to identify differences in the developmental trajectory.

For the purpose of calculating phonemic error frequency, the simpler unweighted Levenshtein edit distance appears to be a more practical choice. The minimum number of addition, omission, and substitution operations that separates two strings of phonemes should, in concept, express frequency of phonemic errors. With this algorithm, alignment of transcribed phonemes would occur as a natural consequence of simply selecting the smallest number of operations for the transformation. Applied to the example in Table 1, where differences in observer judgment affected the manual alignment and error count, the Levenshtein edit distance matches the character strings "t" and "ə" between target and production, simply because this alignment reduces the

Table 1. Illustration of alignment difficulties that arise with manual coding of phonemic error types.

Source	Transcription							Substitution errors	Addition errors	Deletion errors	Total errors	Phonemic errors (%)
Target: "octopus"	ɑ	k	t	ə	p	ʊ	s					
Coder 1 (A)	t	ə	b	ɛ	n	—	—	5	0	2	7	100.0
Coder 2 (B)	—	—	t	ə	b	ɛ	n	3	0	2	5	71.4

Note. Em dashes indicate no phoneme was transcribed in this location. Phonemic errors (%) are the ratio of coded errors to the number of phonemes in the target word.

number of necessary operations to five (three substitutions and two deletions), which is the lower limit for converting the seven-character target string /ɑktəpʊs/ to the produced five-character string /təbɛn/. The unweighted Levenshtein edit distance is not influenced by similarity in the final syllable, yet in our sample, the score is the same as the manual count for Coder B, who aligned syllables based on subjective judgment of phonetic similarity. We hypothesized that this would be the case sufficiently often to validate the unweighted Levenshtein edit distance as a reasonable solution in a large sample of speakers with AOS and APP. To obtain a metric that is comparable to the proportion of target phonemes produced in error, we express the ratio of the unweighted Levenshtein distance to the number of segments in the target word. We call this metric the *phonemic edit distance ratio*. The phonemic edit distance ratio for the single word in Table 1 is 5 divided by 7, which, when multiplied by 100, is functionally expressed as a phonemic error rate of 71.4%.

The purpose of this study was to determine how well the unweighted phonemic edit distance ratio, calculated automatically and without rules for segment alignment other than the inherent consequence of minimizing the number of string operations, matches manually calculated phonemic error rates based on subjective alignment decisions made by phonetically trained coders. We applied the Levenshtein edit distance algorithm using two different resources (an online calculator and a software package) to evaluate feasibility in both clinical and research environments.

Method

Speech Samples

The analyses were conducted on a set of transcriptions from two recent investigations conducted in our laboratory. Together, they featured a total of 65 participants with aphasia and impaired speech sound production. The first sample (A) consisted of 24 speakers diagnosed with AOS (Haley, Smith, & Wambaugh, 2019). These participants repeated 27 words with varied length and phonetic complexity during a motor speech evaluation (Duffy, 2013). Words ranged from one to four syllables. The second sample (B) consisted of 41 speakers with conduction aphasia from the online AphasiaBank database (Haley, Harmon, et al., 2017; MacWhinney & Fromm, 2016). These speakers named 15 pictures while completing the short form of the Boston Naming Test (Kaplan, Goodglass, & Weintraub, 2001).

Phonetic Transcription

The first author transcribed both speech samples. At the onset of the project, he had completed two university courses in narrow phonetic transcription under the direction of the third author and accumulated over 500 hr of narrow phonetic transcription experience with stroke survivors specifically. Sample A featured a total of 648 words (27 words × 24 speakers), and Sample B featured 615 words (15 words × 41 speakers).

The transcriptions were completed using Klattese, in which every American English phoneme is provided an easily accessible keyboard correspondent—for example, "@" for /æ/, "J" for /d͡ʒ/, and "Y" for /a͡ɪ/ (Vitevitch & Luce, 2004). Though any set of characters can be used to derive the edit distance, Klattese is convenient because it is easy to learn by extrapolating from familiar International Phonetic Alphabet symbols and can be typed quickly on a standard English language keyboard. Sample A was transcribed from audio recordings using the Praat software through which waveform and spectrographic analyses were displayed (Boersma & Weenink, 2011). The transcriptions generated text files that were converted into spreadsheets. Sample B was transcribed from a video directly into a spreadsheet, using a spreadsheet template. With both methods, normal allophonic variation was purposely transcribed identically for the target and production, thereby ensuring they were not counted as errors.

Error Frequency Coding

First, the transcriber computed phonemic accuracy manually for both speech samples. This was done by counting the frequency of omissions, additions, and substitutions for each word and dividing by the total number of phonemes in the targets to express the percentage of phonemes with phonemic errors. Subjective judgment was inherent to this process in terms of how to align the target and production character strings based on phonetic similarity. Next, the research team computed the phonemic edit distance ratio, based on the standard unweighted Levenshtein distance. To increase external validity, the algorithm was applied to the Klattese transcription strings via two methods. For Sample A, we used a strategy that is suitable for a large-scale data analysis, as would be the case in research settings. We used R Version 3.5.1 and a custom code that we ran entirely within the native R suite (R Core Team, 2018). This code is provided as an R package (Cunningham, 2018). With an input file containing the transcription of the target word and the transcription of the participant's attempt on the same row, the script calculated the edit distance ratio for each speaker using the native R Levenshtein function. The sum of the edit distances for the words produced by each speaker was divided by the sum of the total number of target phonemes to yield the phonemic edit distance ratio.

Sample B was analyzed with a strategy we anticipated would be useful in clinical settings where results for a single or small number of clients are needed quickly. The transcription strings were copied and pasted into an online edit distance calculator (Holsinger, 2017), again with the Klattese transcriptions of the target and production on the same row.

Interobserver Agreement

To estimate interobserver reliability of the phonetic transcription, a second transcriber coded Sample B in its entirety. This second transcriber had completed the same

two transcription courses as the primary transcriber and accumulated over 150 hr of narrow phonetic transcription for speech samples produced by speakers with left hemisphere stroke. Independent phonetic transcriptions were generated. Interobserver reliability for the manual coding, expressed as a Pearson correlation, was .940 for the manual calculation of combined substitution, addition, and omission errors. The mean speaker-level difference between the two transcribers was 5.3 percentage points, and the maximal difference was 18.4 percentage points. For the online calculator, interobserver reliability, expressed as the Pearson correlation, was .938; the mean speaker-level difference between observers was 5.6 percentage points; and the maximal difference was 18.9 percentage points.

Results

Convergent validity between the manual error coding and the phonemic edit distance ratio was evaluated separately for Speech Samples A and B. The mean difference between the two metrics was less than 2 percentage points for both samples, and the Pearson correlation was almost perfect at .993 and .994 (see Table 2).

Discussion

The results demonstrate that the phonemic edit distance ratio is a valid estimate of phonemic error frequency and, as such, suitable for application to clinical documentation and research. We conclude that, at least for speakers with AOS and APP, a simple unweighted Levenshtein edit distance summarizes the frequency of phonemic errors for target–production word pairs the same way phonetically trained listeners do, with the added advantage of standardized rules. Applications are likely straightforward to other communication disorders for tasks where there is an agreed-upon target transcription, and it is otherwise feasible to derive error frequency.

The automated processing does introduce a need to anticipate systematic decisions up front, rather than addressing them as integral components to the manual coding process. For example, it is important to submit appropriate transcription sequences for analysis. In the case of AOS and APP, revisions, repetitions, and abandoned attempts sometimes precede the full word production. If these are transcribed and included in the analysis, the phonemic edit distance ratio will consider them segment additions and generate an exaggerated estimate of error frequency. Moreover, because all phonemes are treated as individual characters without consideration of feature similarity, clinicians may wish to follow our example and account for normal allophonic variation simply by refraining from transcribing them as different from target phonemes.

Table 2. Relationship between phoneme error frequency (in %) and phonemic error distance expressed as percentage of target phonemes for the two speech samples.

Measure	Sample A	Sample B
Pearson correlation	.993	.994
Mean difference (%)	1.9	0.17
Maximal difference (%)	6.9	3.6

There are of course other reasons to conduct phonemic analyses of speech disorders than to estimate overall error frequency, and for these purposes, a weighted method may be advantageous. Using our target population as an example, more precise estimates of phonetic similarity could help differentiate diagnostically between speech profiles where the errors are phonetically similar to targets and profiles where they are more distant (Canter, Trost, & Burns, 1985). A reasonable prediction is also that degree of phonetic similarity predicts responsiveness to treatment. There are likely numerous customized applications through which differently weighted edit distance operations can help answer questions that are specific to clinical populations.

Edit distance weights can also be used to modify string operations based on sequential properties. We considered some of the more common edit distance metrics that use such adjustments but rejected them for the purpose of expressing phonemic error frequency. For example, the Demerau–Levenshtein distance differs from the standard Levenshtein formula by counting transposed characters as a single edit versus two edits (Damerau, 1964). While a common occurrence with typos, the transpositions are not a feature of particular interest after left hemisphere stroke. The Jaro edit distance (Jaro, 1989) treats characters as common if they occur within half the length of the longest string. In this context of disordered speech, these edits would be treated as errors and thus included in the analysis. Other modifications to the sequential relationship among operations may be suitable for purposes that are specific to communication disorders. For our target population of AOS and APP, a better understanding of phonemic perseverations comes to mind as a potentially productive pursuit.

In conclusion, we demonstrated that the phonemic edit distance ratio offers feasibility of measurement precision for large-scale research and clinical settings. The automated calculation of error frequency allows clinicians and researchers to allocate their full attention and time to the skilled auditory analysis that is the necessary foundation for any error analysis. It also minimizes the risk of human error during data processing. When numerous transcription analyses are to be performed, we recommend running the analyses in batches via the provided R package or similar software. The code also generates a breakdown for basic phonemic error categories (omission, addition, substitution), which is often useful to better understand performance profiles. However, for clinical purposes, an online calculator (Holsinger, 2017, or equivalent) is adequate and likely preferable for its ease of use and efficiency in small samples. In either case, the analysis yields a fine-grained, reliable, and valid metric of overall error frequency that should be useful for both research and clinical applications.

Acknowledgments

This research was supported by the American Speech-Language-Hearing Association Students Preparing for Academic Research Careers awarded to the first author. We are grateful to Ian Kim who shared his programming expertise and to Taylor Petroski for careful perceptual analysis.

References

Bislick, L., McNeil, M., Spencer, K. A., Yorkston, K., & Kendall, D. L. (2017). The nature of error consistency in individuals with acquired apraxia of speech and aphasia. *American Journal of Speech-Language Pathology, 26*(2S), 611–630.

Boersma, P., & Weenink, D. (2011). *PRAAT: Doing phonetics by computer (Version 5.2.45)* [Computer program]. Retrieved from http://www.fon.hum.uva.nl/praat/

Canter, G. J., Trost, J. E., & Burns, M. S. (1985). Contrasting speech patterns in apraxia of speech and phonemic paraphasia. *Brain and Language, 24,* 204–222.

Covington, M. A. (1996). An algorithm to align words for historical comparison. *Computational Linguistics, 22*(4), 481–496.

Cunningham, K. T. (2018). editRatio. *R Package* (Version 0.1.0). Retrieved from https://github.com/unccard/editRatio

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the Association for Computational Linguistics, 7*(3), 171–176.

Duffy, J. R. (2013). *Motor speech disorders: Substrates, differential diagnosis, and management* (3rd ed.). St. Louis, MO: Mosby.

Duffy, J. R., Hanley, H., Utianski, R., Clark, H., Strand, E., Josephs, K. A., & Whitwell, J. L. (2017). Temporal acoustic measures distinguish primary progressive apraxia of speech from primary progressive aphasia. *Brain and Language, 168,* 84–94.

Faes, J., Gillis, J., & Gillis, S. (2016). Phonemic accuracy development in children with cochlear implants up to five years of age by using Levenshtein distance. *Journal of Communication Disorders, 59,* 40–58.

Gooskens, C., & Heeringa, W. (2004). Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Language Variation and Change, 16*(3), 189–207.

Haley, K. L., Bays, G. L., & Ohde, R. N. (2001). Phonetic properties of aphasic–apraxic speech: A modified narrow transcription analysis. *Aphasiology, 15*(12), 1125–1142.

Haley, K. L., Cunningham, K. T., Eaton, C. T., & Jacks, A. (2018). Error consistency in acquired apraxia of speech with aphasia: Effects of the analysis unit. *Journal of Speech, Language, and Hearing Research, 61*(2), 210–226.

Haley, K. L., Harmon, T., Smith, M. T., Jacks, A., Richardson, J., Dalton, S., & Shafer, J. (2017, November). *Apraxia of speech in conduction aphasia: A clinical reality*. Poster session presented at the meeting of American Speech-Language-Hearing Association Annual Convention, Los Angeles, CA.

Haley, K. L., Jacks, A., & Cunningham, K. T. (2013). Error variability and the differentiation between apraxia of speech and aphasia with phonemic paraphasia. *Journal of Speech, Language, and Hearing Research, 56*(3), 891–905.

Haley, K. L., Jacks, A., de Riesthal, M., Abou-Khalil, R., & Roth, H. L. (2012). Toward a quantitative basis for assessment and diagnosis of apraxia of speech. *Journal of Speech, Language, and Hearing Research, 55*(5), S1502–S1517.

Haley, K. L., Jacks, A., Richardson, J. D., & Wambaugh, J. L. (2017). Perceptually salient sound distortions and apraxia of speech: A performance continuum. *American Journal of Speech-Language Pathology, 26*(2S), 631–640.

Haley, K. L., Smith, M., & Wambaugh, J. L. (2019). Sound distortion errors in aphasia with apraxia of speech. *American Journal of Speech-Language Pathology, 28*(1), 121–135. https://doi.org/10.1044/2018_AJSLP-17-0186

Holsinger, E. (2017). *Edit distance calculator*. Retrieved from http://www.ripelacunae.net/projects/levenshtein

Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association, 84*(406), 414–420.

Kaplan, E. F., Goodglass, H., & Weintraub, S. (2001). *The Boston Naming Test* (2nd ed.). Philadelphia, PA: Lippincott Williams & Wilkins.

Kruskal, J. B., & Liberman, M. (1999). The symmetric time-warping problem: From continuous to discrete. In J. B. Kruskal, & D. Sankoff (Eds.), *Time warps, string edits, and macromolecules: The theory and practice of sequence comparison* (pp. 125–161). Stanford, CA: CSLI Publications.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady, 10*(8), 707–710.

MacWhinney, B., & Fromm, D. (2016). AphasiaBank as BigData. *Seminars in Speech and Language, 37*(1), 10–22.

Miller, N. (1995). Pronunciation errors in acquired speech disorders: The errors of our ways. *International Journal of Language & Communication Disorders, 30*(3), 346–361.

Odell, K., McNeil, M. R., Rosenbek, J. C., & Hunter, L. (1990). Perceptual characteristics of consonant production by apraxic speakers. *Journal of Speech and Hearing Disorders, 55*(2), 345–359.

Odell, K., McNeil, M. R., Rosenbek, J. C., & Hunter, L. (1991). Perceptual characteristics of vowel and prosody production in apraxic, aphasic, and dysarthric speakers. *Journal of Speech and Hearing Research, 34*(1), 67–80.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Schluter, R., Nussbaum-Thom, M., & Ney, H. (2011). On the relationship between Bayes risk and word error rate in ASR. *IEEE Transactions on Audio, Speech, and Language Processing, 19*(5), 1103–1112.

Scholl, D. I., McCabe, P. J., Heard, R., & Ballard, K. J. (2018). Segmental and prosodic variability on repeated polysyllabic word production in acquired apraxia of speech plus aphasia. *Aphasiology, 32*(5), 578–597.

Shriberg, L. D., Austin, D., Lewis, B. A., McSweeny, J. L., & Wilson, D. L. (1997). The percentage of consonants correct (PCC) metric: Extensions and reliability data. *Journal of Speech, Language, and Hearing Research, 40*(4), 708–722.

Shriberg, L. D., & Kwiatkowski, J. (1982). Phonological disorders III: A procedure for assessing severity of involvement. *Journal of Speech and Hearing Disorders, 47*(3), 256–270.

Vitevitch, M. S., & Luce, P. A. (2004). A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, & Computers, 36*(3), 481–487.

Yang, C., & Castro, A. (2008). Representing tone in Levenshtein distance. *International Journal of Humanities and Arts Computing, 2*(1–2), 205–219.