Pronunciation accuracy and intelligibility of non-native speech

Anastassia Loukina, Melissa Lopez, Keelan Evanini,
David Suendermann-Oeft, Alexei V. Ivanov, Klaus Zechner

Educational Testing Service, USA
{aloukina, mlopez002, kevanini, suendermann-oeft, aivanou, kzechner}@ets.org

Abstract

This paper investigates the connection between intelligibility and pronunciation accuracy. We compare which words in non-native English speech are likely to be misrecognized and which words are likely to be marked as pronunciation errors. We found that only 16% of the variability in word-level intelligibility can be explained by the presence of obvious mispronunciations. In some cases, a word remained recognizable or could be identified from the context despite obvious pronunciation errors. In many other cases, the annotators were unable to identify the word when listening to the audio but did not perceive it as mispronounced when presented with its transcription. At the same time, we see high agreement when the results are aggregated across all words from the same speaker.

Index Terms: pronunciation error detection, annotation, crowd-sourcing, educational applications, second language acquisition, tutoring systems

1. Introduction

Non-native speech is commonly characterized on three dimensions: perceived *accentedness*, *comprehensibility* (perceived ease of understanding) and *intelligibility* (the listener's ability to report the words pronounced by the speaker) [1, 2].

While there is a close connection between accentedness and comprehensibility [2], speech perceived as heavily accented may remain highly intelligible [3, 4, 1].

In this paper we use a large corpus of non-native English unscripted speech and multiple judgments obtained via crowdsourcing to explore the connection between word-level intelligibility and pronunciation accuracy of non-native speech.

Previous research has shown that not all pronunciation errors contribute equally to the intelligibility of non-native speech. For example, consistent deviations in pronunciation such as transfer errors generally may have less effect on intelligibility than, for example, errors in prosody [5]. This is consistent with findings from speech perception which showed that listeners can quickly accommodate to consistent differences in pronunciation even across multiple speakers [6, 7]. In general, [8] reported a correlation of r = -0.67 between strength of accent and intelligibility and this finding agrees with other studies [3, 9, 10].

Bent and Bradlow [10] further explored word-level intelligibility by comparing keyword recognition rates with detailed phonetic transcriptions provided by two expert phoneticians for 16 sentences read by 15 speakers of Mandarin Chinese. They aggregated segmental accuracy scores across different positions and found that errors in certain positions, such as strong syllables, word-initial and syllable-initial positions, are more detrimental for overall intelligibility than others.

The multiple factors that contribute to intelligibility are not limited to the acoustic properties of non-native speech. A listener's prior experience, expectations and attitudes also play an important role, and there is substantial variation in assessment of non-native speech [11, 3, 12]. These results are consistent with previous research on transcription accuracy which has demonstrated that transcribers can achieve agreement as high as 95-98% (or word error rate (WER) of 2-5%) on transcriptions of native speech [13, 14]. However, the WER between transcribers increases to 15-20% for non-native speech [15, 16]. Previous studies showed that identification of pronunciation errors in non-native speech is a similarly subjective task [17, 18, 19].

Given such variability between listeners, reliable results can only be obtained by using multiple judgments. Several studies have demonstrated that crowdsourcing platforms like Amazon's Mechanical Turk can produce high-quality data efficiently and inexpensively [20, 21, 22]. Crowdsourcing has been used previously for annotating pronunciation errors by [24] and [25].

In this paper, we further explore the connection between perceived intelligibility and pronunciation accuracy at the word level. We follow the approach previously used by [10] but collect multiple judgments of pronunciation accuracy and intelligibility for each word. This approach allows us to assess not only the connection between overall speaker intelligibility and segmental accuracy but also to which extent intelligibility of each particular word is affected by pronunciation accuracy of that word.

Finally, while most previous studies relied on read speech from around 10-20 speakers we use a large corpus of unscripted speech.

2. Data and methodology

2.1. Corpus of L2 speech

The study is based on a corpus of L2 speech which contains responses to a test of English language proficiency collected from 143 non-native speakers with different native languages (14,718 words in total).

Each speaker responded to one of four test items. Two of the items required test takers to listen to an audio file and respond to a prompt about the conversation or lecture they heard. For the other two items, the test takers were required to read a short passage and listen to an audio file and then integrate information from both sources in their responses to a prompt. The speakers were given 1 minute to record their responses.

All responses were assigned proficiency scores on a four-point scale ranging from 1 to 4 by expert raters. The scoring guidelines were modeled after the scoring rubrics for English proficiency tests (cf. [23]) but focused on pronunciation only.

The raters were asked to evaluate each speaker's fluency, the overall intelligibility of the response, and the listener effort required to understand the response. Thus score 4 was described as "clear, well-paced speech which may include minor difficulties that do not affect overall intelligibility". Score 1 was described as "choppy and fragmented speech, where consistent difficulties cause considerable listener effort".

2.2. Annotation

2.2.1. Data preprocessing

We first obtained orthographic transcriptions for all 143 responses. We then used The Penn Phonetics Lab Forced Aligner to [26] to align the transcriptions with the recording and identify pauses in each response. We also used a clause boundary detection algorithm [27] to identify clause boundaries in the transcription. We used these clause boundaries along with punctuation from the orthographic transcriptions and pauses identified by the forced alignment to split the recordings into short fragments of 5-13 words in length.

This procedure produced an average of 12.1 fragments per response, and the average length of each fragment was 8.3 words. The final set consisted of 1,767 fragments; these were presented to the annotators in randomized order.

By segmenting the responses into shorter fragments, we limited the cognitive load on the annotators. Randomized presentation was necessary to ensure that the annotators had little opportunity to accommodate to an individual speaker's foreign accent or to make use of the extended context.

We conducted further analyses to make sure that our approach to fragment selection did not lead to differences in fragment length between the fragments extracted from responses assigned different proficiency scores. If, for example, fragments extracted from speech of low-proficiency speakers contained fewer words than the fragments from high-proficiency speakers, this could lead to differences on transcription accuracy due to lack of sufficient context rather than intelligibility of speech (cf. [12]). We found that there was no difference in average number of words in fragments extracted from responses assigned different scores by human raters. As one may expect the number of fragments per response was correlated with the proficiency score since high-proficiency speakers usually produce longer responses (Spearman's $\rho = 0.58$, $p < 0.0001$).

2.2.2. Procedure

We used the Amazon Mechanical Turk crowdsourcing platform to collect multiple judgments about intelligibility and pronunciation accuracy for each word. Our experiment included two task types: a transcription task and an error detection task. We collected 5 judgments for each task for a total of 17,670 judgments.

The first part of our experiment was a transcription task. This task was posted first to make sure the annotators were not familiar with the content of the fragment (since several annotators participated in both tasks). For the transcription task, the annotators were asked to play the audio and transcribe the words that they heard using standard English spelling. We also asked the annotators to rate the audio quality of each recording as 0 ('OK'), 1 ('Somewhat Poor'), and 2 ('Poor').

After the transcription task was complete, we posted an error detection task modeled after the task in [24]. We provided the original expert transcription for each fragment and asked the annotators to play the audio and mark the words that they considered to be mispronounced. We also asked them to mark possible errors in the reference transcription. This was done to distinguish between perceived deviations in pronunciation and potential discrepancies between the transcription and the audio due to inaccurate forced alignment or mistakes in the original transcription. As we had done for the transcription task, we asked the annotators to make an additional judgment about audio quality.

2.2.3. Annotators and quality control

Each fragment was annotated by 5 annotators recruited through Amazon Mechanical Turk. Our aim was to evaluate the intelligibility of non-native speech to an average North American listener; therefore, we selected annotators with addresses in the United States. In addition, we created a short qualification test which included a sample transcription and error detection task. Only workers who successfully passed the qualification test were allowed to take part in the experiment.

After collecting all responses, we identified and excluded the annotators whose responses were significantly different from the rest of the group. This was done by fitting mixed-level logistic regression with annotation as dependent variable and annotator and fragment identity as random factors. Such model allows for random variation between baseline likelihood of errors in each fragment and for each worker. In both cases the baselines are assumed to have normal distribution across workers and fragments. We then identified the workers for whom the baseline probability of error was an outlier defined as more than four standard deviations from the mean value. We found that there were several workers who marked unusually small percentage of errors in comparison to other workers. We further reviewed their annotations to confirm that they indeed showed signs of negligence on behalf of the worker and obtained new annotations as necessary so that the total number of annotators for each fragment was 5. The results presented in this section only include the annotators whose responses were used for the analysis.

In total, there were 57 unique annotators; 47 of these worked on the transcription task and 38 worked on the error identification task. Out of 53 annotators who completed the demographic survey, there were 16 males (30%) and 35 females (70%) (two of the annotators did not provide their gender). Only one of the annotators reported that North American English was not their native language (this annotator reported that Singaporean English was their native language), and they were spread out geographically in the United States (they reported 25 different states as their home states). In terms of exposure to foreign languages, only 6 annotators (11%) reported having lived abroad and 34 (64%) reported having minimal exposure to non-native English speakers in their daily lives through close friends, colleagues, or family.

2.3. Data postprocessing

We first identified and corrected all spelling errors in the crowdsourced transcriptions. We then identified words from the original expert transcription that were marked as "transcription error" by the majority of annotators during the error detection task (see 2.2.2) and evaluated how many of these words were recognized in the transcription task. There were 195 (1.6%) such words and 180 of these words were not recognized by the majority of the workers. These were excluded from further analysis. We also excluded 15 fragments (0.85%) which had an average audio quality rating greater than 1.

Approximately 62% of all words in our corpus were function words (prepositions, pronouns, articles etc.). Previous work in transcription accuracy has shown that short function words are often mistranscribed even in clear native speech. Therefore for this paper we excluded function words from further analysis and focus on content words only, which we will call "keywords" (cf. [8]). The final corpus used for the analysis presented in this paper thus consisted of 1,719 fragments extracted from 143 responses which included 5,423 content words with an average of 3 keywords per fragment.

For each such keyword in the reference transcription, we computed how many annotators recognized that word in their transcription ('intelligibility score' or I_w) and how many annotators did not mark that word as mispronounced ('pronunciation score' or P_w). Both these values were scaled to a 0-1 range.

3. Results

3.1. Word-level results

Around 46% of all keywords were recognized by all annotators in the transcription task, while 8% were not recognized by any of the annotators (see Table 1). There were 354 fragments (20%) where all annotators recognized all keywords and 37 fragments (2%) where none of the annotators recognized any of the keywords.

For pronunciation scores, 55% of all keywords were not marked as mispronounced by any of the annotators and only 3% of words were marked as mispronounced by all annotators.

Table 1: *Distribution of Intelligibility (I_w) and Pronunciation (P_w) scores. The table shows % words assigned each score.*

	0	0.2	0.4	0.6	0.8	1
I_w	8.0	6.7	8.2	10.5	20.7	45.8
P_w	2.7	4.2	6.5	10.7	20.4	55.5

The correlation between average intelligibility score and average pronunciation score was Spearman's ρ = 0.36 (p <0.0001), or in other words, the word pronunciation score is a very weak predictor of word intelligibility score.

We used the majority rule to classify all words into 'mostly recognized' ($I_w > 0.4$), 'mostly unrecognized' ($I_w \leq 0.4$), 'mostly marked correct' (P_w >0.4) and 'mostly marked error' ($P_w \leq 0.4$). The results of cross-tabulation are shown in Table 2. The majority of words were recognized and classified as correct by the majority of annotators. About half of these words (32% of all words) were recognized correctly by all annotators and not marked as errors by any of the annotators. The second largest category (846 words or 15.6% of all words) were words that were mostly marked as correct but also generally unrecognized. In other words, 18% of all 4,696 words marked as correct were not recognized in the transcription task. Of these, 84 were words that no annotator marked as mispronounced and yet nobody could recognize when transcribing. Of words marked as errors by the majority of annotators, about half of the words were recognized and another half not recognized by the majority of annotators.

Since the data has a hierarchical structure, we used a multilevel linear model to evaluate how much variance in I_w can be explained by the variance in P_w. We fitted the linear model using speaker and word identity as crossed random factors, P_w as the fixed factor, and I_w as the dependent variable (see for example [28] for further discussion on use of mixed-effects modeling with crossed random factors). Likelihood ratio tests showed that P_w had a significant effect on I_w ($p < 0.0001$).

Table 2: *Distribution (count and % of all words) words that were recognized ($I_w > 0.4$) or unrecognized ($I_w \leq 0.4$) and marked as error ($P_w \leq 0.4$) or correct ($P_w > 0.4$) by the majority of annotators*

	Recognized	Unrecognized
Correct	3,850 (71.0%)	846 (15.6%)
Error	330 (6.1%)	397 (7.3%)

To compute the amount of variance in I_w explained by P_w we used the approach described in [29] who give further detail. We first fitted a first order multi-level model which only included the random effects of speaker and word identity. The total variance in I_w for this model would be the sum of variances attributed to speaker and word identity as well as residual (unexplained) variance. We next ascertained how much variance in I_w could be explained by linear effects of the P_w by fitting the full model which included both speaker and word as random factors and P_w as fixed factor. The explained variance was computed as the proportional reduction in total variance between the first-order multi-level model and this full model [31].

The analysis of proportional reduction of variance between the models with and without P_w showed that P_w explained 16% of variance in I_w.

3.2. Speaker-level results

In addition to the word-level results, we computed an average intelligibility score (I_{sp}) and pronunciation score (P_{sp}) for each of the 143 speakers in our corpus by averaging, respectively, the I_w and P_w for all words from the speaker's response. We found that the correlation between these two speaker-level scores was r = 0.61 ($p < 0.00001$). A linear model with I_{sp} as the dependent variable and P_{sp} as the independent variable showed that speaker pronunciation score accounted for about 36% of variance in the speaker intelligibility score ($F_{(1,141)}$ = 81.5, Adj. r^2 = 0.362, $p < 0.0001$).

Finally, we compared the scores obtained via crowdsourcing with proficiency scores assigned by professional raters (see 2.1). Both I_{sp} and P_{sp} were positively correlated with expert scores: Spearman's ρ = 0.59 for I_{sp} and ρ = 0.68 for P_{sp} ($p < 0.00001$ in all cases). Combined into a single linear model, both I_{sp} and P_{sp} had significant contributions to the proficiency score and explained about 50% of variance in these scores ($F_{(2,140)}$ = 65.9, Adj. r^2 = 0.48, $p < 0.0001$).

3.3. Discussion

We analyzed the connection between word intelligibility and perceived pronunciation accuracy and found that at the word level, mispronunciation only predicts a small amount of variance in intelligibility. Our results showed that words that are perceived as mispronounced remain intelligible in about half of all cases.

Furthermore, we found that words that are not perceived as mispronounced when the annotators are provided with a transcription of non-native speech may not be intelligible without a transcription. Further qualitative analysis of such cases showed that these were often words that were unlikely in a given context because of poor lexical choice (for example, 'band' in "play band with their friend"), incorrect grammar ('figure' in "he

figure out the courses") or low frequency words occurring in a broad context ('floral' in "some other shapes like the floral shapes"). These results once again show that the full model of intelligibility should incorporate context-related effects in addition to pronunciation accuracy.

At the same time, we found that the agreement between pronunciation accuracy and intelligibility increases when the results are aggregated across the whole one-minute response. In addition, we found that pronunciation accuracy had a higher effect on the general proficiency score which reflects the rater's judgment of speaker intelligibility and overall fluency. Within the terminology adopted in this paper, the proficiency score is closest to comprehensibility or the ease of understanding as perceived by the rater. Our results thus support previous findings that the connection between 'comprehensibility' and 'strength of accent' is stronger than the connection between these two subjective measures and intelligibility [9, 8]. This finding also has a practical application: it shows that the proficiency score assigned by the raters in the context of language assessment may not be an accurate reflection of objective intelligibility of speaker's speech. In order to provide a comprehensive assessment of all three dimensions of non-native speech, human ratings could be supplemented with an automatic system which evaluates objective intelligibility of the speaker.

While crowdsourcing allowed us to collect multiple judgements from naïve listeners for a large corpus of data, it is also a limitation of this study. The environment in which the annotators listened to the stimuli differed between the listeners and was most likely different from that of a controlled lab experiment in terms of background noise, output quality and possibly the level of attention. At the same time we note that our results at the response level closely replicate the results previously obtained in more traditional laboratory-based experiments.

Finally, in this paper, intelligibility was construed as the ability to provide word-by-word transcription. Although intelligibility is commonly measured based on keyword recognition, it is not clear how such word-based metric relates to the overall communicative effectivess of the speaker and their ability to convey the information [30]. We plan to explore this connection in future work.

4. Conclusion

We investigated the connection between intelligibility and pronunciation accuracy at the word level by comparing which words are likely to be misrecognized and which words are likely to be marked as pronunciation errors. We found that only 16% of the variability in word-level intelligibility can be explained by the presence of obvious mispronunciations. Words perceived as mispronounced remain intelligible in about half of all cases. At the same time the annotators were often unable to identify the word when listening to the audio but did not perceive it as mispronounced when presented with its transcription. When word-level results were aggregated for the whole response, the results were in agreement with previous studies which reported tighter connection between comprehensibility and strength of accent than between either of these measures and intelligibility.

5. Acknowledgements

We would like to thank Nitin Madnani for his help with setting up the crowdsourcing study; Hillary Molloy for help with annotations; Lawrence Davis for providing access to expert scores; Xinhao Wang, Vikram Ramanarayanan, Beata Beigman-Klebanov, and three anonymous Interspeech reviewers for their comments and suggestions.

6. References

[1] M. J. Munro and T. M. Derwing, "Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language Learners," *Language Learning*, vol. 45, no. 1, pp. 73–97, 1995.

[2] T. M. Derwing and M. J. Munro, "Putting accent in its place: Rethinking obstacles to communication," *Language Teaching*, vol. 42, no. 4, pp. 476–490, 2009.

[3] R. Hayes-Harb and J. Watzinger-Tharp, "Accent, Intelligibility, and the Role of the Listener: Perceptions of English-Accented German by Native German Speakers," *Foreign Language Annals*, vol. 45, no. 2, pp. 260–282, 2012.

[4] M. J. Munro, "Foreign accent and speech intelligibility," in *Phonology and second language acquisition*, M. L. Zampini and E. J. G. Hansen, Eds. Amsterdam: John Benjamins, 2008, pp. 193–218.

[5] M. J. Munro and T. M. Derwing, "The foundations of accent and intelligibility in pronunciation research," *Language Teaching*, vol. 44, no. 3, pp. 316–327, 2011.

[6] A. Cutler, "The abstract representations in speech processing." *Quarterly journal of experimental psychology*, vol. 61, no. 11, pp. 1601–1619, 2008.

[7] A. R. Bradlow and T. Bent, "Perceptual adaptation to non-native speech." *Cognition*, vol. 106, no. 2, pp. 707–729, 2008.

[8] E. Atagi and T. Bent, "Perceptual dimensions of nonnative speech," *Proceedings of the XVIIth International Congress of Phonetic Sciences. Hong Kong, China*, pp. 260–263, 2011.

[9] T. M. Derwing, M. J. Munro, and G. Wiebe, "Evidence in Favor of a Broad Framework for Pronunciation Instruction," *Language Learning*, vol. 48, no. 3, pp. 393–410, 1998.

[10] T. Bent, A. R. Bradlow, and B. L. Smith, "Segmental errors in different word positions and their effects on intelligibility of non-native speech," in *Language experience in second language speech learning: in honor of James Emil Flege*, M. J. Munro and O.-S. Bohn, Eds. Amsterdam: John Benjamins, 2007, pp. 331–346.

[11] O. Kang and D. Rubin, "Intra-rater reliability of oral proficiency ratings," *The International Journal of Educational and Psychological Assessment*, vol. 12, pp. 43–61, 2012.

[12] S. Kennedy and P. Trofimovich, "Intelligibility, comprehensibility, and accentedness of L2 speech: The role of listener experience and semantic context," *The Canadian Modern Languages Review/La Revue canadienne des langues vivantes*, vol. 64, no. 3, pp. 459–489, 2008.

[13] N. Deshmukh, R. Duncan, A. Ganapathiraju, and J. Picone, "Benchmarking human performance for continuous speech recognition," *Proceeding of 4th International Conference on Spoken Language Processing. ICSLP '96*, 1996, vol. 4, pp. 2486–2489.

[14] W. D. Raymond, M. Pitt, K. Johnson, E. Hume, M. Makashay, R. Dautricourt, and C. Hilts, "An analysis of transcription consistency in spontaneous speech from the Buckeye corpus," in *Proceedings of the 7th International Conference on Spoken Language Processing ICSLP'02*, 2002, pp. 1125–1128.

[15] K. Zechner, "What did they actually say? Agreement and Disagreement among Transcribers of Non-Native Spontaneous Speech Responses in an English Proficiency Test," in *Proceedings of SLaTE*, 2009, pp. 3–6.

[16] K. Evanini, D. Higgins, and K. Zechner, "Using Amazon Mechanical Turk for transcription of non-native speech," in *CSLDAMT '10 Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010, pp. 53–56.

[17] Bonaventura, D. D. Herron, and W. Menzel, "Phonetic Rules for Diagnosis of Pronunciation Errors." in *KONVENS 2000 / Sprachkommunikation, Vortrge der gemeinsamen Veranstaltung 5. Konferenz zur Verarbeitung natrlicher Sprache (KONVENS), 6. ITG-Fachtagung "Sprachkommunikation", 9. bis 12. Oktober 2000, Technische Universitt Ilmenau*, W. Zühlke and E. G. Schukat-Talamazzini, Eds. VDE Verlag, 2000, pp. 225–230.

[18] T. Cincarek, R. Gruhn, C. Hacker, E. Nöth, and S. Nakamura, "Automatic pronunciation scoring of words and sentences independent from the non-native's first language," *Computer Speech & Language*, vol. 23, no. 1, pp. 65–88, 2009.

[19] X. Yang, A. Loukina, and K. Evanini, "Machine learning approaches to improving pronunciation error detection on an imbalanced corpus," in *Proceedings of IEEE Spoken Language Technology Workshop, South Lake Tahoe*, 2014, pp. 300–305.

[20] R. Snow, B. O. Connor, D. Jurafsky, A. Y. Ng, "Cheap and fast - but is it good? Evaluating non-expert annotiations for natural language tasks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2008, pp. 254–263.

[21] C. Callison-Burch, "Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk," in *EMNLP '09 Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, vol. 1, no. 1. Association for Computational Linguistics, 2009, pp. 286–295.

[22] J. Tetreault, M. Chodorow, and N. Madnani, "Bucking the trend: improved evaluation and annotation practices for ESL error detection systems," *Language Resources and Evaluation*, vol. 48, no. 1, pp. 5–31, 2013.

[23] *TOEFLiBT scoring guides (rubrics) for spoken responses*. [Online] Available at http://www.ets.org/toefl/institutions/scores/guides/

[24] M. A. Peabody, "Methods for pronunciation assessment in computer aided language learning," Unpublished PhD thesis, MIT, 2011.

[25] H. Wang, X. Qian, and H. Meng, "Predicting Gradation of L2 English Mispronunciations using Crowdsourced Ratings and Phonological Rules," *Proceedings of SLaTE 2013, Grenoble, France.*, pp. 127–131, 2013.

[26] J. Yuan and M. Liberman, "Speaker identification on the SCOTUS corpus," *Proceedings of Acoustics*, pp. 5687–5690, 2008.

[27] L. Chen and S.-Y. Yoon, "Detecting structural events for assessing non-native speech," *Proceedings of the 6th workshop on Innovative Use of NLP for Buidling Educational Applications*, pp. 38–45, 2011.

[28] R. Baayen, D. Davidson, and D. Bates, "Mixed-effects modeling with crossed random effects for subjects and items," *Journal of Memory and Language*, vol. 59, no. 4, pp. 390–412, 2008.

[29] A. Loukina, B. Rosner, G. Kochanski, E. Keane, and C. Shih, "What determines duration-based rhythm measures: text or speaker?" *Laboratory Phonology*, vol. 4, no. 2, pp. 339–382, 2013.

[30] B. Bridgeman, D. Powers, E. Stone, and P. Mollaun, "TOEFL iBT speaking test scores as indicators of oral communicative language proficiency," *Language Testing*, vol. 29, no. 1, pp. 91–108, 2011.

[31] T. Snijders and R. Bosker, R. J. "Modeled Variance in Two-Level Models". *Sociological Methods and Research*, vol. 22, no. 3, pp. 342–363.