Expert and crowdsourced annotation of pronunciation errors for automatic scoring systems

Anastassia Loukina, Melissa Lopez, Keelan Evanini, David Suendermann-Oeft, Klaus Zechner

Educational Testing Service, USA

{aloukina, mlopez002, kevanini, suendermann-oeft, kzechner}@ets.org

Abstract

This paper evaluates and compares different approaches to collecting judgments about pronunciation accuracy of non-native speech. We compare the common approach, which requires expert linguists to provide a detailed phonetic transcription of non-native English speech, with word-level judgments collected from multiple naïve listeners using a crowdsourcing platform. In both cases we found low agreement between annotators on what words should be marked as errors. We compare the error detection task to a simple transcription task in which the annotators were asked to transcribe the same fragments using standard English spelling. We argue that the transcription task is a simpler and more practical way of collecting annotations which also leads to more valid data for training an automatic scoring system.

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

1. Introduction

Automatic scoring systems for spoken responses such as the ones described in [1, 2, 3, 4] provide a fast and efficient way to evaluate language proficiency of non-native speakers and give an immediate feedback. Such systems rely on machine learning algorithms to compute a proficiency score based on a set of features extracted from each response. These features usually cover various aspects of proficiency from general fluency of speech to content coverage.

One aspect of automatic speech assessment is automatic evaluation of pronunciation accuracy. In addition to holistic measures of pronunciation accuracy such as the ones described in [5, 6], it is also desirable to include features that can identify specific mistakes to provide feedback or additional measures of proficiency. Like many other classification tasks, training a system for automatic identification of pronunciation errors requires a labeled corpus of such errors. In this paper we consider different approaches to collecting such a corpus and consider whether crowdsourcing can be used to replace experts for this task.

One of the challenges with pronunciation error annotation is the definition of what constitutes a pronunciation error. Non-native speech usually shows multiple deviations from any single native variety. Much previous research on pronunciation learning has focused on systemic errors which presumably arise from mismatches between native and non-native phonological systems. These can include the use of different variants of the same phoneme, such as the use of a retroflex consonant instead of a dental one, consistent substitution between two phonemes, or changes to phonotactics such as insertions or deletions in consonant clusters. In addition to such segmental errors, non-native speech also usually shows deviations in prosody and rhythm.

A common approach to annotation in this case would be to ask the expert raters to provide a phonetic transcription or nativeness judgment for each phone in the utterance (cf. [7, 8, 9, 10]). This approach works well when the annotators are required to annotate all instances of a small number of phones (as was done for example by [11], who asked the annotators to transcribe all instances of [θ]). However, detailed phonetic annotation of all phones in the utterance is a difficult and subjective task even for trained phoneticians (cf. also [12]).

The subjectivity of the error annotation task is evidenced by low inter-rater agreement reported in previous studies. For example, [8] reported 80.2% agreement on localization of errors (phone level) on Spanish data. For English, [13] reported 67% agreement on localization of errors (phone level). Almost no studies report Cohen's kappa as a measure of agreement, but the numbers reported in [10] suggest κ=0.29 for Dutch data. Finally, [9] report ICC between 0.28 and 0.56.

Furthermore, manual expert annotation of pronunciation errors is a costly and time-consuming task. Since a large amount of labeled data is usually required to train an automatic system, only one or two judgments are collected for each word. Without several judgments, it is impossible to distinguish between clear and ambiguous cases when training the system and this in turn may have a negative effect on performance. For example, [14] showed that in the case of semantic annotations, using ambiguous cases in classification leads to lower performance of the classifier even for easy cases. It also has negative consequences for system evaluation since equal weight is given to misclassified cases which are unambiguous for human raters and the cases which cause substantial disagreement even among expert judges.

Finally, most language assessments covered by automatic scoring systems focus on assessing the communicative abilities of the learner; in terms of pronunciation, this corresponds to intelligibility. For example, the scoring rubrics for TOEFLiBT, an academic English proficiency test, focus on intelligibility and general accuracy of the speaker and allow small variations in pronunciation which do not affect intelligibility [15]. Previous research has shown that not all errors have an equally detrimental effect on successful communication. For example, consistent deviations in pronunciation such as transfer errors generally may have less of an effect on intelligibility than, for example, errors in prosody [16]. This is consistent with findings from speech perception which showed that listeners can very quickly accommodate to consistent changes in pronunciation [17]. Detailed phonetic annotation may therefore result in the identification of too many false positives by flagging errors which may not necessarily affect comprehensibility.

To summarize, a traditional expert annotation of pronunci-

ation errors performed as a phone-by-phone detailed phonetic transcription is not only labor- and time-intensive, but it is also likely to provide unreliable data because of the highly subjective nature of the task and a high number of false positives.

One way to address this issue is to ask the annotators to focus only on errors that are likely to affect intelligibility. For example, this was done by [18]. This may make the task more manageable and reduce the number of false positives, but it also adds another layer of subjectivity to an already subjective task. Crowdsourcing is another common way to quickly collect multiple judgments, and it has been successfully used for collecting annotations which are inherently subjective. For example, [19] showed that using multiple judgments for grammar errors leads to better performance of a grammar error detection system. Crowdsourced annotations of pronunciation errors have been previously collected by [20], who used three annotators to annotate the CU-CHLOE corpus of read sentences and paragraphs. This study reported pairwise agreement between Turkers with κ varying from 0.3 to about 0.6, or an aggregated κ of 0.51. [21] used crowdsourcing to collect word-level judgments for the same corpus and reported aggregated kappas of 0.37 to 0.43. Both these studies used read speech obtained from speakers with the same native language. They also used inter-rater agreement as the only evaluation for the annotation.

In this paper we compare several approaches to annotating pronunciation errors: expert annotation, which asks annotators to focus on errors which may affect the speaker's intelligibility, and crowdsourcing using two different tasks, error annotation and transcription. Since our goal is to compare two approaches to annotation rather than the performance of naïve and expert annotators, we collected each set of annotations under the setup most common or suitable for each group.

We evaluate these annotations on three dimensions: first of all, which approach shows the best inter-annotator agreement in terms of localization and number of errors. Second, we evaluate the validity of annotations obtained using these methods. We consider both (1) the agreement between the annotation results and the proficiency scores assigned by expert raters, and (2) to which extent the annotation guidelines are aligned with the goals of automated scoring system. Finally, we consider which approach is most robust to external factors such as annotator diligence.

Unlike the previous studies which used corpora of read speech, we use unscripted speech. Unscripted speech has better ecological validity for spoken language proficiency assessment, but it also provides unique challenges for the task of pronunciation error annotation, since the lexical content varies widely among different utterances.

2. Data and methodology

2.1. Corpus of non-native speech

The study is based on a corpus of non-native English speech which contains 143 responses to a test of English language proficiency collected from 140 non-native speakers of seven different native languages.

All but one speaker responded to one of four test items (one speaker responded to all four 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 one 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 scoring rubrics for English proficiency tests (cf. [15]) but focused on pronunciation and general fluency. 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

We first obtained orthographic transcriptions for all 143 responses. We then used the Penn Phonetics Lab Forced Aligner [22] to align the transcriptions with the recording and identify locations of word boundaries, phoneme boundaries, and silences in each response.

2.2.1. Crowdsourced annotation

The crowdsourced annotations were collected using Amazon Mechanical Turk. The responses were split into shorter fragments based on pauses identified by the forced alignment, automatically detected clause boundaries [23] and punctuation from the orthographic transcriptions. The average length of each fragment was 8.3 words, and there were on average 12.1 fragments per response. The final set consisted of 1,767 fragments; these were presented to the annotators in randomized order.

We used the Amazon Mechanical Turk crowdsourcing platform to collect multiple judgments about pronunciation errors 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 error detection task was modeled after the task in [20]. We provided the transcription for each fragment and asked the annotators to play the audio and mark the words that they considered to be "noticeably mispronounced". We provided various examples of mispronounced words. We also asked the annotators 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. We also asked the annotators to rate the audio quality of each recording as 0 ('OK'), 1 ('Somewhat Poor'), and 2 ('Poor').

For the transcription task, the annotators were asked to play the audio and transcribe the words that they heard using standard English spelling. This task was posted first, before the error detection task, to make sure the annotators were not familiar with the content of the fragment when completing their transcriptions (since several annotators participated in both tasks). All crowdsourced transcriptions were checked for spelling errors.

We limited the annotators to those with addresses in the United States. In addition, we created a short qualification test which included a sample transcription and error detection task and collected demographic information about the annotators. After collecting all responses, we applied several statistical analyses to identify and exclude the annotators whose responses were significantly different from the rest of the group. Finally, we 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 from different areas of the United States. Of these, 56 reported North American English as their native language. One annotator reported that they were a native speaker of Singaporean English.

After collecting the annotations, we first identified words from the original reference transcription that were marked as "transcription error" by the majority of annotators (at least 3) during the error detection task (explained earlier in this section). There were 195 (1.6%) such words, which were excluded from further analysis. We also excluded 15 fragments (0.85%) which had an average audio quality rating below the 'somewhat poor' threshold. The final corpus used for the analysis presented in this paper thus consisted of 1,752 fragments extracted from 143 responses which included 14,374 words.

2.2.2. Expert annotation

We used a subset of 75 responses to two items to collect expert annotations of pronunciation errors following the standard approach used in previous studies for annotating pronunciation errors [7, 8, 9, 10]. The annotators could listen to the whole recording or selected parts of the recording multiple times. They also had access to the spectrogram and the waveform.

When developing the guidelines for the annotators, we adopted a version of the approach previously used by [18], who asked raters to identify "the most serious errors to be corrected in the subjects' speech", letting the annotators make their own judgment about what errors should fall under this category.

Twelve responses (about 15%) were selected for double annotation to test inter-annotator agreement. The remaining files were split between two annotators using stratified sampling so that each annotator was assigned an equal number of responses from speakers with different native languages. The files selected for double annotation were interspersed with other responses and the annotators were not aware which responses were selected for double annotation.

3. Results

3.1. Inter-annotator agreement

Table 1 shows inter-annotator agreement for all tasks and annotators. For error detection task we used the binary label provided by each annotator for each word ('correct/error'). For transcription task we first aligned the reference transcription and the transcription provided by the annotator and then assigned each word in the reference transcription a binary label depending on whether it was correctly recognized in the supplied transcription.

The agreement on localization of errors was computed as Cohen's κ for expert annotation (2 annotators) or Fleiss's κ for crowdsourced annotation (5 annotators). The agreement on the number of errors was computed as a correlation between relative number of errors corrected in each response by each of the annotators. For the expert annotation, we simply used the correlation between the two annotators. For the crowdsourced annotation, we computed pairwise correlations between all 10 annotator pairs and used the median value.

Finally, for each set, we also computed the average percentage of words in each response that were flagged as mispronounced in the error detection task or transcribed incorrectly in the transcription task.

The inter-annotator agreement was generally lower for crowdsourced annotations than for expert annotations. Furthermore, the agreement for the error detection task was lower than for the transcription task. At the same time we found that crowdsourced annotations showed higher agreement for the number of errors in each response, with a particularly high agreement for the transcription task.

Table 1: *Inter-annotator agreement for different sets of annotations for the error detection (ED) and transcription (TR) tasks. The table shows the number of words (N_w) and the number of responses (N_r) in each set, the agreement on localization of errors (κ) and the number of errors (r) for both tasks, and average percentage of errors in each response ($\%_{err}$) (see main text for further detail).*

Task	Annotation	N_w	N_r	κ	r	$\%_{err}$
ED	Crowd	14,374	143	0.297	0.71	12%
ED	Expert	1,443	12	0.492	0.53	29%
TR	Crowd	14,374	143	0.429	0.82	27%

3.2. Annotation validity

To evaluate the validity of our annotations, we computed correlations between the number of words marked as mispronounced in each response and the proficiency score assigned by the expert raters. Our expectation was that responses with lower proficiency scores should also contain more pronunciation errors.

For expert annotations, we computed the correlation between the proficiency score and the percentage of words corrected by each annotator. For responses annotated by both annotators we used the mean value. For crowdsourced annotations, we assigned each word a pronunciation error probability score (P_{pron}) based on how many annotators out of 5 marked that word as an error. We also computed the transcription error probability score (P_{tr}) based on how many annotators failed to correctly transcribe this word. We then computed the average $\bar{P}_{pron}$ and $\bar{P}_{tr}$ for each response and correlated these with the expert proficiency rating for the response. The correlations are shown in Table 2.

Table 2: *Correlation (Spearman's ρ) between $\bar{P}_{pron}$/$\bar{P}_{tr}$ and the proficiency score assigned to each response. N_r indicates the total number of responses. For crowdsourced results the numbers in brackets indicate the values for the same 75 responses as annotated by the experts. All correlations are significant at α = 0.0001.*

Task	Annotation	N_r	ρ
ED ($\bar{P}_{pron}$)	Crowd	143 (75)	-0.7 (-0.72)
ED ($\bar{P}_{pron}$)	Expert	75	-0.48
TR ($\bar{P}_{tr}$)	Crowd	143 (75)	-0.56 (-0.58)

3.3. Comparison between expert and crowdsourced annotations

We compared the agreement between expert and crowdsourced annotations for 75 responses (5,155 words) for which we had both annotations. We used the 'majority' rule to classify all words in the crowdsourced annotations as 'correct' or 'error' and compared these labels with those assigned by expert annotators. For the error detection task, the results was Cohen's κ

= 0.33 for the agreement between crowdsourced labels and the first annotator (N = 2,595) and κ = 0.27 for the second annotator (N = 2,560). For the transcription task, the results was Cohen's κ = 0.28 for both annotators.

We then compared $\bar{P}_{pron}$ and $\bar{P}_{tr}$ introduced in the previous section with relative number of errors corrected by expert annotators for each response. We found that $\bar{P}_{pron}$ was strongly correlated with the number of corrections made by expert annotators (r = 0.71, $p < 0.00001$). The relationship between the number of expert corrections and $\bar{P}_{tr}$ was somewhat weaker: r = 0.5, $p < 0.00001$.

3.4. The effect of external factors on the results of crowdsourced annotation

We further explored the extent to which the crowdsourced annotations could have been influenced by external factors such as reported audio quality and the number of times the annotator played each fragment.

3.4.1. Audio quality

The quality of audio varied between fragments and therefore as described in section 2.2.1 we asked the annotators to mark the quality of the recordings.

First of all, we found that even though we used the same fragments for both tasks, the annotators gave lower quality judgments for the transcription task, where they only had access to the recording, than for the error detection task, where they could also see the reference transcription. For the error annotation task, 83% of all fragments were marked as "good quality" by all annotators. For the transcription task this was the case for only 40% of all fragments.

We then aggregated the average audio quality judgment for each task and average percentage of errors and misrecognized words in each fragment. For the transcription task, fragments with lower quality judgments also had a higher percentage of transcription errors (r = -0.75, $p < 0.0001$). This appeared to be due both to quality of the recording and possibly lower pronunciation accuracy since the annotators also tended to flag more errors in fragments marked as low quality during transcription task (r = 0.34, $p < 0.0001$). The correlation between the percentage of errors and the quality judgments obtained during error detection task was much lower: r = 0.16 ($p < 0.0001$).

3.4.2. Number of times played

The annotators had to play each fragment at least once before submitting their annotation, but there was no limit on the total number playbacks. We tracked how many times each annotator played each fragment to investigate whether this can provide additional evidence about the diligence of the annotators or the difficulty of the task.

We found that some annotators played fragments more times than others, with the average number of playbacks for each annotator varying between one and eight. For the error detection task, annotators who listened to each fragment more often also tended to mark a higher percentage of words as mispronounced: partial correlation after controlling for proficiency score r = 0.39 (p = 0.01). There was no such correlation for the transcription task, i.e., the average percentage of transcription errors for each annotator was not correlated with the average number of times they played each fragment.

3.5. Discussion

In this study, we compared different approaches to manually labeling mispronounced words for training automatic speech scoring systems. We compared the common approach, which requires expert linguists to provide a phonetic transcription of non-native speech, with judgments collected from multiple naïve listeners using the Amazon Mechanical Turk crowdsourcing platform.

In both cases, we found low agreement on what words should be marked as errors, which is not surprising given the subjectivity of the task and the fact that non-native speech usually shows multiple deviations from any single native variety. We note, however, that our results for inter-annotator agreement between expert annotators compare favorably with inter-annotator agreement reported in previous studies.

We found that the agreement between naïve annotators was higher for the transcription task, which is an easier and more intuitive task than annotation of pronunciation errors. The results of this task were also less influenced by external factors. For example, we found that the annotators who played the fragments multiple times were also more likely to mark more errors. There were no such inter-annotator differences for the transcription task.

We saw high agreement on the relative number of errors in a given response both within each group of annotators and between the two groups, especially for crowdsourced annotations. Furthermore, the correlation between this number and the proficiency score assigned by expert rater was higher for crowdsourced annotations than for expert annotations. Expert annotators listened to the whole responses and therefore had greater opportunity to accommodate to the speaker's accent. They also listened to multiple responses to the same prompt and therefore developed certain expectations about what key terms are likely to be mispronounced and paid greater attention to these words (cf. also [24] on the effect of word identity).

Finally, the two tasks, error detection and transcription, measure related but different aspects of non-native speech: while the first task evaluates pronunciation accuracy, the second task evaluates intelligibility. For assessments which focus on communicative competency of language learners, the transcription task provides annotations which are more aligned with the goals of the assessment (cf. also [25] for further discussion).

While not directly related to the main purpose of this study, we found that audio quality judgments are dependent on the task. Our results showed that the annotators are more lenient with their quality judgments when the task does not require them to understand and transcribe the speech. At the same time, for the transcription task, their quality judgments were also influenced by the pronunciation accuracy of the speaker, which presumably made the fragments harder to understand.

4. Conclusion

We conclude that the pronunciation accuracy annotations collected using crowdsourcing are more predictive of expert proficiency scores than expert annotations. Furthermore, using a transcription task instead of an error annotation task leads to better inter-annotator agreement between the annotators in terms of both localization and the number of errors. Finally, the transcription task is better aligned with the focus on communication rather than accent reduction, which is common to many proficiency tests.

5. References

[1] D. Higgins, X. Xi, K. Zechner, and D. Williamson, "A three-stage approach to the automated scoring of spontaneous spoken responses," *Computer Speech & Language*, vol. 25, no. 2, pp. 282–306, 2011.

[2] M. Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, vol. 51, no. 10, pp. 832–844, 2009.

[3] J. Bernstein, A. Van Moere, and J. Cheng, "Validating automated speaking tests," *Language Testing*, vol. 27, no. 3, pp. 355–377, 2010.

[4] J. Cheng, Y. Z. D'Antilio, X. Chen, and J. Bernstein, "Automatic Assessment of the Speech of Young English Learners," *Proceedings of the 9th Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 12–21, 2014.

[5] L. Chen, K. Zechner, and X. Xi, "Improved pronunciation features for construct-driven assessment of non-native spontaneous speech," in *Human Language Technologies: The Annual Conference of the North American Chapter of the ACL. 2009, Boulder, Colorado*, pp. 442–449.

[6] L. Chen, K. Evanini, and X. Sun, "Assessment of non-native speech using vowel space characteristics," in *2010 IEEE Spoken Language Technology Workshop*. pp. 139–144.

[7] S. Witt and S. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 2, pp. 95–108, 2000.

[8] H. Franco, H. Bratt, R. Rossier, V. Rao Gadde, E. Shriberg, V. Abrash, and K. Precoda, "EduSpeak®: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications," *Language Testing*, vol. 27, no. 3, pp. 401–418, 2010.

[9] 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.

[10] J. van Doremalen, C. Cucchiarini, H. Strik, and J. V. Doremalen, "Automatic pronunciation error detection in non-native speech: The case of vowel errors in Dutch," *The Journal of the Acoustical Society of America*, vol. 134, no. 2, pp. 1336–1347, 2013.

[11] K. Evanini and B. Huang, "Automatic detection of [th] pronunciation errors for Chinese learners of English," in *Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training, Stockholm.*, 2012, pp. 71–74.

[12] P. Road, K. Nueng, K. Luang, A. Chotimongkol, S. Thatphithakkul, P. Chootrakool, C. Hansakunbuntheung, and C. Wutiwiwatchai, "The design and development of PELECAN: Pronunciation Errors from Learners of English Corpus and Annotation," in *Speech Database and Assessments (Oriental COCOSDA), 2011 International Conference on*, 2011, pp. 36–41.

[13] P. Bonaventura, P. Howarth, and W. Menzel, "Phonetic Annotation of a Non-Native Speech Corpus," in *InSTIL 2000 (Integrating Speech Technology in (Language) Learning), pp. 10-17, Dundee, UK, 29-30 August*, 2000, pp. 225–230.

[14] B. Beigman Klebanov and E. Beigman, "Difficult Cases: From Data to Learning, and Back," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, Maryland, 2014, pp. 390–396.

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

[16] 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, May 2011.

[17] A. Cutler, "The abstract representations in speech processing." *Quarterly journal of experimental psychology*, vol. 61, no. 11, pp. 1601–19, 2008.

[18] A. Neri, C. Cucchiarini, and H. Strik, "Selecting segmental errors in non-native Dutch for optimal pronunciation training," *IRAL - International Review of Applied Linguistics in Language Teaching*, vol. 44, no. 4, pp. 357–404, 2006.

[19] 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.

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

[21] 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.

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

[23] 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 Building Educational Applications*, 2011, pp. 38–45.

[24] 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.

[25] A. Loukina, M. Lopez, K. Evanini, D. Suendermann-Oeft, A. Ivanov, K. Zechner. "Pronunciation accuracy and intelligibility of non-native speech. To appear in *Proceedings of Interspeech 2015*, 2015.