INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Assessing Intelligibility in Non-native Speech: Comparing Measures Obtained at Different Levels

Xing Wei[1], Roeland van Hout[2], Catia Cucchiarini[1], Daniëlle Reuvekamp[3], Helmer Strik[1,2,3,4]

[1]Centre for Language and Speech Technology (CLST), Radboud University, The Netherlands
[2]Centre for Language Studies (CLS), Radboud University, The Netherlands
[3]Linguistics, Radboud University, The Netherlands
[4]Donders Institute for Brain, Cognition and Behaviour, Radboud University, The Netherlands

xing.wei@ru.nl, roeland.vanhout@ru.nl, catia.cucchiarini@ru.nl,
danielle.reuvekamp@cito.nl, helmer.strik@ru.nl

Abstract

Speech intelligibility (SI) is essential in communication and second language learning. In this study, non-native SI was measured through Visual Analogue Scale (VAS) scores and Orthographic Transcriptions (OTs) of read aloud sentences. Seven measures automatically derived from the OTs at word and subword levels were studied. The reliability of the intelligibility measures and the correlations between VAS scores and OT-based measures were also explored. Despite the different speaker language backgrounds, the recruited raters exhibited high scoring reliability. The correlations between VAS scores and OT-based measures were weak, corroborating previous assumptions that they refer to two related but distinct notions, comprehensibility (VAS) and intelligibility (OT). OT-based measures are reliable and valid indicators of SI. The results are discussed in relation to previous studies and avenues for future research are proposed.

Index Terms: speech intelligibility, non-native speech

1. Introduction

The growing application of Automatic Speech Recognition (ASR) based technology in the field of second language (L2) learning, can potentially alleviate the shortage of qualified teachers. Although L2 learners can practice their pronunciation with the help of such applications, it remains a big challenge to establish to what extent their speech is intelligible or comprehensible. In L2 learning, speech intelligibility (SI) generally refers to the extent to which listeners can actually understand L2 sentences, while comprehensibility refers to the difficulty or ease listeners experience in understanding utterances. [1]. In speech pathology, on the other hand, slightly different definitions are employed [2]. Both intelligibility and comprehensibility are affected by linguistic background, proficiency levels, speech rate, and several other factors [3]. In L2 learning research researchers suggest that comprehensibility should be measured by collecting scalar judgments, e.g., through a Likert scale [4], or the Visual Analogue Scale (VAS) [5]. Intelligibility, on the other hand, should be measured through manual orthographic transcriptions (OTs) of speech, as mentioned in [6]. However, various methods are employed to measure L2 intelligibility [7], and in speech pathology both metrics are employed to measure intelligibility [8]. For manual transcriptions, various speech materials can be employed, including whole sentences, isolated words or pseudowords, as well as Semantically Unpredictable Sentences (SUS) as is often done in speech pathology research. All these types of speech materials have their own advantages and disadvantages. In the case of isolated words, the influence of the context on the listeners can be minimised. With the help of a speech recognition-based pronunciation trainer, in [9], the intelligibility of English segments produced in isolated words by Chinese speakers was significantly improved even without teacher interaction. On the other hand, an obvious flaw of isolated words is their unnatural context, which makes it unclear how some specific phonemes in isolated words relate to SI in a more natural context. For these reasons, it seems preferable to transcribe whole sentences rather than isolated words. Sentences may be more intelligible because they contain context, this is a more natural condition than isolated, decontextualized words. In [8], researchers designed three experiments with different speech materials including sentences and isolated words. The results showed that for all measures the intelligibility of sentences was noticeably higher than that of isolated words.

The above-mentioned methods rely on subjective human judgments, which are costly, time-consuming, and difficult to implement on a large scale. In addition, the assessment may be affected by the type of measurement or the listeners. For instance, the measurements collected from experienced listeners were less varied than those by inexperienced listeners [10]. Therefore, researchers explored some relatively objective approaches, such as the usability of speech technology to evaluate intelligibility, like ASR-based algorithms or ASR-free features, to predict SI scores [11-12]. In [13], a Bidirectional Long Short-Term Memory (BLSTM) and Multilayer Perceptron (MLP) - Linear Regression (LR) jointed model was proposed to automatically grade non-native speech.

The mentioned approaches may provide objective results comparable to those obtained with human ratings, which could alleviate the need to rely on human efforts. A limitation of these methods is the lack of adequate speech resources that are well transcribed. For this reason, researchers in the field of speech pathology have proposed a semi-automatic approach to evaluating the intelligibility of disordered speech [14]. In this study, the intelligibility of dysarthric speakers measured by OTs was evaluated at sentence, word, and subword levels. The word and subword level ratings were automatically derived by forced alignment and conversion methods. The results proved that the introduced approach was feasible and reliable while providing a more detailed and informative measurement of intelligibility.

In the field of second language learning, there is no consensus on which method should be used to measure SI [15]. In [8],

10.21437/Interspeech.2023-1635

researchers compared five different methods to evaluate the intelligibility of six English distinct varieties. The listeners were required to assess the speech by responding to true or false statements, scalar ratings, perception of nonsense and filtered sentences, and transcription of speech. The results showed that all the methods were effective for intelligibility measurement, but the correlations between the methods were not strong.

The studies discussed above were mainly in the field of pathological speech or L2 English, while relatively few studies have studied SI measurement for other languages. In the present research, an online listening experiment was conducted in order to investigate the SI of non-native speech as measured by VAS scores and measures automatically derived from OTs. The following research questions were addressed: 1) to what extent can VAS scores and OT-based measures provide a reliable basis for the assessment of non-native SI? 2) How strong are the correlations between VAS scores and OT-based measures?

2. Method

2.1. Speech material

The material was selected from a corpus named JASMIN [16], which includes both native and non-native speech from various groups: native primary school children, native secondary school students, non-native children, non-native adults, and native senior citizens. For each group around ninety-five hours of speech recordings were collected, half of which are read speech. Approximately two-thirds of the speakers in this corpus were recruited in the Netherlands and the rest in Flanders.

For non-native read speech, the corpus includes recordings from forty-six speakers. The reading material consists of phonetically rich sentences, stories, and general texts. Given that the raters had to both score and transcribe the recordings sentence by sentence, the listening materials could not be too long to prevent the raters from making 'mistakes' due to lack of memory instead of lack of intelligibility of the speakers. In order to give away as little context as possible to the raters, phonetically rich utterances were selected from the non-native speech part of the corpus. Unlike stories and general texts, the phonetically rich utterances are independent of each other.

In order to control the duration of the experiment and avoid deviations caused by fatigue, the same five sentences from nine non-native speakers were selected (see Table 1). Their recordings lasted on average 7 seconds. Additionally, loudness normalisation was applied for all the recordings to make sure these would be perceived as equally loud by the raters and for a more pleasant listening experience.

Table 1: *Sentences selected from the nine speakers.*

1	ik wou al om half drie hier zijn om alles in de etalage te zetten
2	de voetballer belooft zijn contractuele verplichtingen na te komen
3	de juffrouw rust een middagje uit en doet een dutje
4	de chauffeur tracht met wilde bewegingen de kuilen in de weg te omzeilen
5	de huiseigenaar kwam aan de deur om de huur op te halen

2.2. Rating procedure

Twenty-one expert raters were recruited through a call in a Facebook group that was not accessible to everyone. None of them was familiar with the materials used in the present research. Fourteen raters were certified L2 teachers, but all twenty-one had at least one year of experience in language teaching. Three are men and eighteen are women. On average, the age of the raters and years of experience as an L2 teacher are around 47 and 9.8, respectively.

The listening experiment was conducted using the program Radboud Online Linguistic Experiment Generator (ROLEG), which is a software application from Radboud University that can be used to create behavioural experiments and surveys. Before the listening experiment, the raters could familiarise themselves with the task by listening to extra recordings and receiving extensive instructions on the orthographic transcription and the VAS assessments.

The entire experiment, aside from intermission, lasted approximately 40 to 50 minutes and consisted of two parts with an option to pause in between. In the first part, the first three utterances in Table 1 were scored and transcribed by the raters. The last two sentences were played in the second part. In both parts, the recordings were played randomly. This way the raters had the option to pause and get a new impulse by hearing new sentences in the second part of the experiment.

Although the raters could replay the recordings multiple times, they were asked to repeat a recording no more than once (i.e., listen maximally twice). The raters first made the OTs and then assigned the VAS scores without OTs prompts. The OTs were without punctuation and could be grammatically or semantically incorrect and even include nonsense words since the raters were asked to transcribe precisely what they heard. One rater missed three recordings belonging to two speakers. Consequently, the total number of VAS scores and orthographic transcriptions is 942 (i.e., 9 speakers × 5 sentences × 21 raters − 3 missed sentences).

2.3. Intelligibility measures

2.3.1. Sentence level measure

For VAS rating, a continuous bar was shown on the screen, where the left end denotes completely incomprehensible speech (score = 0) and the right end means perfectly intelligible speech (score = 100). Raters were encouraged to use the entire scale.

2.3.2. Word level measure

Orthographic transcriptions were compared with the reference transcriptions (the prompts in Table 1) after removing punctuation marks, and then Levenshtein distance was applied to calculate word accuracy (W_{Acc}) as shown:

$$Acc = 100 \times (N_{total} - N_d - N_s)/N_{total} \quad (1)$$

where N_{total}, N_d, and N_s represents the number of total words, deletion, and substitution, respectively.

2.3.3. Subword level measures

Grapheme strings were automatically converted to phoneme strings through an online Grapheme to Phoneme (G2P) tool (https://webservices.cls.ru.nl/g2pservice), and all subsequent subword level analyses were carried out for both the grapheme and the phoneme strings. The transcriptions of the spoken

sentences were first aligned with the reference transcriptions by means of the 'ADAPT' algorithm [17]. For graphemes and phonemes, accuracy (G_{Acc} *and* P_{Acc}) was then calculated using equation (1), while distance (*Dist*) and the number of changes (*Ch*) using equations (2) and (3), respectively.

$$Dist = N_d \times C_d + N_i \times C_i + N_s \times C_s \quad (2)$$

$$Ch = N_d + N_i + N_s \quad (3)$$

where N_i represents the number of insertions. C_d, C_i, and C_s denote the cost for deletion, insertion and substitution. For grapheme, C_d and C_i *equalled* 1, and C_s was 2. As for phoneme, C_d and C_i were 3, and C_s was calculated by employing metrics with articulatory features. These different weights were employed in different versions of the 'ADAPT' algorithm that was used for alignment.

3. Results

3.1. Reliability of intelligibility measures

Boxplots of VAS scores and W_{Acc} are shown in Figure 1. The variance in W_{Acc} is less than in VAS, which may have consequences for their correlation with the OT measures.

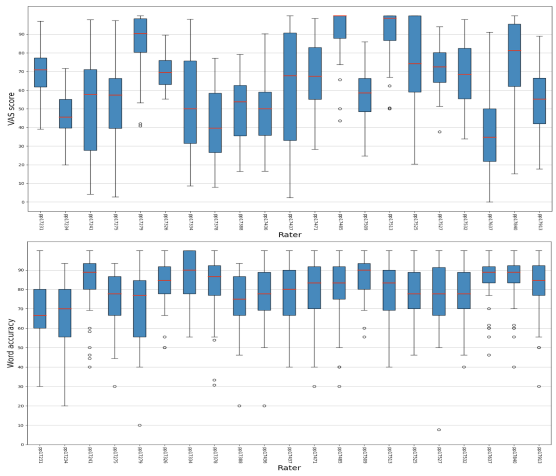

Figure 1: *Boxplots of the VAS scores (upper) and* W_{Acc} *(lower) per rater. The horizontal axes denote the twenty-one raters.*

In order to study the reliability, the Intraclass Correlation Coefficient (ICC) per sentence for VAS scores and the three *Acc* measures derived from the OTs was calculated, see Table 2. The ICC results were computed through SPSS [18], and the parameter for model and type is 'Two-Way random' and 'consistency', respectively. It is apparent that the raters have a higher ICC for sentences 2, 3, and 5, with the lowest ICCs for sentence 4. The majority of the coefficients are higher than 0.9 which reveals that the consistency in ratings among the raters is high for VAS and the three accuracy measures.

Table 2: *ICC for the intelligibility measures per sentence.*

Sentence	VAS	W_{Acc}	G_{Acc}	P_{Acc}
1	0.912	0.885	0.807	0.813
2	0.953	0.941	0.971	0.965
3	0.964	0.972	0.960	0.937
4	0.917	0.939	0.778	0.763
5	0.958	0.934	0.977	0.966

3.2. Statistical analysis of intelligibility measures

The mean, standard deviation (SD), median, and variance of all intelligibility measures are given in Table 3. For the sentence level, there are VAS scores, while all the *Acc* measures relate to lower levels (W, G, P). It can be observed in Table 3 that the mean values of G_{Acc} and P_{Acc} are similar, and both are larger than W_{Acc}, while the magnitude of the VAS scores is much lower. In addition, the number of changes at the grapheme level is larger than that at the phoneme level. Since all the Acc measures were derived from OTs, the correlations between W_{Acc} and VAS scores were further explored in 3.3. A closer inspection of the scattergram in Figure 2 indicates that the same W_{Acc} corresponds to various VAS scores, especially when W_{Acc} scores are between 70 and 90.

Table 3: *Mean, standard deviation, median, and variance of measures at different levels.*

Level	Measure	Mean	SD	Median	Variance
Sentence	VAS	62.99	24.28	64.06	589.74
Word	W_{Acc}	79.05	15.92	83.33	253.36
Grapheme	G_{Acc}	92.62	8.09	95.0	65.48
	G_{Dist}	8.53	7.13	7.0	50.84
	G_{Ch}	6.90	6.34	5.0	40.26
Phoneme	P_{Acc}	92.37	7.73	94.23	59.82
	P_{Dist}	17.52	15.99	14.0	255.69
	P_{Ch}	5.96	5.45	5.0	29.71

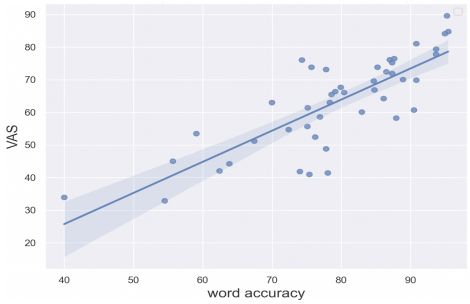

Figure 2: *Scattergram of* W_{Acc} *versus VAS at the sentence level. The shaded zone denotes the 95% confidence interval.*

In order to explore which specific words were relatively more challenging in terms of intelligibility, the OTs at the grapheme level were further analyzed by computing the correctness of the transcription of the content words. Table 4 presents the top 10 highest and lowest correctly transcribed content words. The top 10 highest correct words consist mainly of short and high-frequency verbs, while the top 10 lowest correct words are mainly long words or short words with the rounded front phonemes /y/ ('uu') or /œy/ ('ui'). For instance, the word 'huur' was transcribed in various ways, such as 'hu' (final r deletion),

'guur' (fricative instead of the semivowel), and 'heur' (a lower rounded vowel).

Table 4: *The top 10 highest and lowest corrected transcribed words with percentage.*

Highest correctness		**Lowest correctness**	
Word	**Percentage**	**Word**	**Percentage**
zijn	96.8%	contractuele	19.1%
kwam	92.0%	huiseigenaar	28.3%
halen	89.8%	kuilen	37.4%
zetten	87.8%	uit	43.9%
komen	87.2%	omzeilen	46.0%
wou	86.8%	bewegingen	49.2%
etalage	84.7%	middagje	52.4%
weg	84.0%	huur	57.2%
alles	83.1%	verplichtingen	58.2%
hier	81.5%	voetballer	59.6%

3.3. Correlations between intelligibility measures

The correlation coefficients among the explored intelligibility measures in Table 5 reveal that W_{Acc} is strongly correlated with the *Acc* measures at the subword level, while this is not the case for the VAS scores. The highest correlation obtained, as expected, was between G_{Acc} and P_{Acc} (= 0.919). It is important to note that the correlation coefficient between the VAS and W_{Acc} is weak (= 0.325; see Figure 2), and even lower with the two *Acc* measures at the subword level (see Table 5).

Table 5: *Correlations between measures at word and subword levels.*

	Grapheme			**Phoneme**		
	Acc	*Dist*	*Ch*	*Acc*	*Dist*	*Ch*
VAS	0.296	-0.302	-0.249	0.299	-0.222	-0.229
$\boldsymbol{W_{Acc}}$	0.751	-0.647	-0.573	0.693	-0.451	-0.463

4. Discussion and conclusions

In the present study, a listening experiment was conducted aimed at a comprehensive assessment of non-native SI. Speech materials from nine non-native speakers selected from a speech corpus were assessed by native speakers with experience in language teaching through VAS scores and OTs.

Table 2 shows that most of the ICC values per sentence are higher than 0.9, indicating a strong reliability of the raters' measurements. Sentence 4 was more difficult to understand, which might be caused by the length and/or the number of long words (sentence 4 contained three long words listed with the lowest correctness in Table 4).

Three accuracy measures were automatically derived from OTs and prompts at the word, grapheme, and phoneme levels, as well as distance scores and the number of changes. The results for all eight measures shown in Table 3 indicate that all VAS mean values are lower than those of W_{Acc}. In addition, Figure 2 shows that the same W_{Acc} may correspond to various VAS scores, while the values of G_{Acc} and P_{Acc} were close and significantly better than the above two SI measures, which is in line with outcomes in [8, 14]. This is understandable since only one grapheme or phoneme transcription error will result in a score of 0 for the whole word. Additionally, one phoneme may relate to more than one grapheme and then its correctness depends on the associated grapheme to a certain degree [14, 19].

As to the comparison between VAS and of W_{Acc} our results are in line with those of previous research employing these measures for dysarthric speech [14], in which VAS scores appeared to be lower than W_{Acc}, indicating that raters tend to judge sentences to be less intelligible than they in fact are. However, a different trend was observed in [8], which may be ascribed to differences between speech materials and human raters. Our outcomes are also more in line with those of research on non-native SI by Kang et al. [15]. In this study, five different methods of intelligibility measurement were investigated and found to be weakly correlated with each other. We also found a low correlation of VAS scores with the *Acc* measures, while the correlations between the three *Acc* measures were high, which is understandable as these were all derived from the OTs.

The above findings show that although VAS scores can be obtained more easily, they show more variation between raters. The advantage of OTs is that they make it possible to analyze intelligibility at the word and subword level, which is important in the field of L2 learning, but also in speech pathology. By calculating the seven measures used in the current paper, much more detailed information at various levels can be obtained, e.g., the easiest and most difficult words shown in Table 4.

To summarize, we can conclude that all the measures investigated were reliable for assessing different aspects of non-native SI. The selection of the most suitable measurement should be made in relation to the intended purpose. VAS scores appear to be more intuitive and would seem to be more appropriate for assessing the perceived level of speech intelligibility, which is often referred to as comprehensibility. OT measures, on the other hand, seem to be more direct measures of SI and seem appropriate to support research on specific pronunciation problems in a substantial way.

Future work could study the effect of specific words and sounds in natural sentences on the assessment of intelligibility. It is also important to investigate the possibility of automatically generating orthographic transcriptions and assessing intelligibility without human-made transcriptions, through ASR. The rapid developments in ASR technology suggest new avenues of research in this direction that are definitely worth exploring. Another option could be to investigate how SI measurement by experts, as in this study, relates to that by native raters without experience in language teaching. This would certainly be interesting in terms of ecological validity, as non-native speakers will not always be evaluated by language teachers, but eventually by the native speakers with which they will engage in conversation in their daily lives.

5. Acknowledgements

The authors would like to acknowledge the financial support from the China Scholarship Council (CSC), and the twenty-one human raters of the online listening experiment.

6. References

[1] S. Kennedy and P. Trofimovich, "Intelligibility, Comprehensibility, and Accentedness of L2 Speech: The Role of Listener Experience and Semantic Context," *Canadian Modern Language Review*, vol. 64, no. 3, pp. 459–489, Mar. 2008.

[2] K. M. Yorkston, E. A. Strand, and M. R. T. Kennedy, "Comprehensibility of Dysarthric Speech," *American Journal of Speech-Language Pathology*, vol. 5, no. 1, pp. 55–66, Feb. 1996.

[3] M. J. Munro, T. M. Derwing, and S. L. Morton, “The mutual intelligibility of l2 speech,” *Studies in Second Language Acquisition*, vol. 28, no. 01, Feb. 2006.
[4] K. M. Yorkston and D. R. Beukelman, “A comparison of techniques for measuring intelligibility of dysarthric speech,” *Journal of Communication Disorders*, vol. 11, no. 6, pp. 499–512, Dec. 1978.
[5] C. Finizia, J. Lindström, and H. Dotevall, “Intelligibility and Perceptual Ratings After Treatment for Laryngeal Cancer: Laryngectomy Versus Radiotherapy,” *The Laryngoscope*, vol. 108, no. 1, pp. 138–143, Jan. 1998.
[6] M. J. Munro, and T. M. Derwing. "Intelligibility in research and practice: Teaching priorities." *The handbook of English pronunciation* (2015): 375-396.
[7] R. Thomson, "Measurement of accentedness, intelligibility, and comprehensibility." *Assessment in second language pronunciation*. Routledge, 2017. 11-29.
[8] W. Xue, V. M. Ramos, W. Harmsen, C. Cucchiarini, R. van Hout, and H. Strik, “Towards a Comprehensive Assessment of Speech Intelligibility for Pathological Speech,” *Interspeech 2020*, Oct. 2020.
[9] D. F. Burleson, “Improving Intelligibility of Non-Native Speech with Computer-Assisted Phonological Training,” *IULC Working Papers*, vol. 7, no. 1, 2007.
[10] E. O. Mencke, G. J. Ochsner, and E. W. Testut, “Listener judges and the speech intelligibility of deaf children,” *Journal of Communication Disorders*, vol. 16, no. 3, pp. 175–180, May 1983.
[11] C. Middag, T. Bocklet, J. P. Martens, and E. Nöth, "Combining phonological and acoustic ASR-free features for pathological speech intelligibility assessment," *Interspeech 2011*, pp3005-3008.
[12] C. Spille, S. D. Ewert, B. Kollmeier, and B. T. Meyer, “Predicting speech intelligibility with deep neural networks,” *Computer Speech & Language*, vol. 48, pp. 51–66, Mar. 2018.
[13] Z. Yu *et al*., "Using bidirectional lstm recurrent neural networks to learn high-level abstractions of sequential features for automated scoring of non-native spontaneous speech," *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, AZ, USA, 2015, pp. 338-345.
[14] M. Ganzeboom, M. Bakker, C. Cucchiarini, and H. Strik, “Intelligibility of Disordered Speech: Global and Detailed Scores,” *Interspeech 2016*, Sep. 2016.
[15] O. Kang, R. I. Thomson, and M. Moran, “Empirical Approaches to Measuring the Intelligibility of Different Varieties of English in Predicting Listener Comprehension,” *Language Learning*, vol. 68, no. 1, pp. 115–146, Oct. 2017.
[16] C. Cucchiarini, J. Driesen, H. van Hamme, and E. Sanders, “Recording Speech of Children, Non-Natives and Elderly People for HLT Applications: the JASMIN-CGN Corpus.,” LREC 2008.
[17] B. Elffers, C. van Bael, and H. Strik, “ADAPT Algorithm for Dynamic Alignment of Phonetic Transcriptions,” Technical Report. Department of Language and Speech. Radboud University Nijmegen. The Netherlands, 2005.
[18] IBM SPSS Statistics for Windows, Version 26.0. Armonk, NY: IBM Corp. Released 2019.
[19] D. Abur, N. M. Enos, and C. E. Stepp, “Visual Analog Scale Ratings and Orthographic Transcription Measures of Sentence Intelligibility in Parkinson’s Disease With Variable Listener Exposure,” *American Journal of Speech-Language Pathology*, vol. 28, no. 3, pp. 1222–1232, Aug. 2019.