DYSARTHRIC SPEECH INTELLIGIBILITY EVALUATION – A COMPARISON BETWEEN HUMAN AGENTS AND ASR SYSTEMS

Bernadett Dam[1,2,3,4], Árpád Berta[5], László Tóth[5], Lívia Ivaskó[1,3,6]
[1]*Department of General Linguistics, University of Szeged, Hungary*
[2]*Doctoral School of Linguistics, University of Szeged, Hungary*
[3]*Developmental and Neuropragmatic Research Group, University of Szeged, Hungary*
[4]*Institute of Applied Pedagogy, University of Szeged, Hungary*
[5]*Institute of Informatics, University of Szeged, Hungary*
[6]*Department of Oto-Rhino-Laryngology and Head-Neck Surgery, University of Szeged, Hungary*

Abstract

Motor speech disorders can often lead to reduced intelligibility, communication difficulties, ultimately affecting the person's quality of life. In recent years, several initiatives have been launched to improve the processing of disordered speech [1]. However, these initiatives differ in their methodology and their degree of success.

This study focuses on the intelligibility of Hungarian dysarthric speakers, comparing four perceptual metrics: speech intelligibility, speech severity, speech naturalness, and listener effort. We invited trained and untrained listeners in speech analysis to rate the level of intelligibility of dysarthric speech samples. We compared their ratings with the error rates of automatic speech recognition (ASR) systems. Previously [5], we discussed automatic speech evaluation methods using four ASR systems.

We selected 15 dysarthric speech samples from the Hungarian Dysarthria Database [6]. Our dataset contained recordings from 7 mild, 5 moderate, and 3 severe dysarthric speakers, and 5 healthy controls [7]. Speech samples were evaluated based on the four metrics using a visual analog scale. The judgements of human listeners were compared with the error rates of ASR systems.

Findings showed great variation in the evaluation of severity, the degree of severity correlates with intelligibility in both listener groups.

Keywords: *dysarthria, ASR, intelligibility, speech evaluation, listener effort*

1. Introduction

Dysarthria is a medical umbrella term that covers neurologic-related speech disorders that affect several aspects of speech (e.g., articulation, phonation, respiration, resonance, loudness), resulting in impaired speech intelligibility [2, 8].

We are working on a collaborative project called the 'Dysarthria' project at the University of Szeged with two main goals in mind. First, we conduct detailed analysis of dysarthric speech while considering etiological aspects [2, 3]. Secondly, we aim to find solutions which intend to improve the intelligibility of dysarthric speech using automatic procedures [4, 5].

This study focuses on the intelligibility of Hungarian dysarthric speakers, comparing two groups of human

Corresponding author: *Bernadett Dam, dam.bernadett@szte.hu.*

listeners (trained vs. untrained) with automatic speech recognition (ASR) systems.

The experiment builds on the results of our previous work [5], where we tested automatic speech evaluation methods using four ASR systems.

Hypothesis:

Based on the results of previous outcome measurements found in the literature [9] we supposed that there would be a difference between the evaluation of intelligibility of dysarthric speech across human agents (trained and untrained) and the ASR systems depending on the severity of dysarthria.

2. Methods

For the current study, we selected 15 dysarthric speech samples from the Hungarian Dysarthria Database [6]. Our dataset contained recordings from 7 mild, 5 moderate, and 3 severe dysarthric speakers, and 5 healthy controls [7]. Speech samples were evaluated by 25 listeners: 19 trained and 6 untrained participants. Trained listeners included people who had a formal education in speech and language therapy, untrained listeners included native Hungarian speakers with no previous experience with speech disorders.

The study was conducted using PsychoPy [10]. Stimuli were presented on a laptop, participants were asked to rate speech samples based on the four metrics using a visual analog scale (see Fig.1.). Their ratings were stored as numeral values between 0 and 1.

The judgements of human listeners were compared with the error rates of ASR systems. We compare recognition results from OpenAI's Whisper (Large-V3) [11] and Meta's Massively Multilingual Speech (MMS) [12]. We also chose to work with models developed for Hungarian speech recognition, namely, BEAST2 [13] and FastConformerHU, developed by the company SpeechTex [14]. For a detailed account of automatic evaluation methods used for this study, see [5].

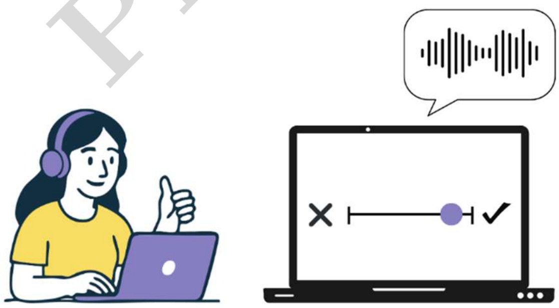

Figure 1. Study design

3. Results

The following section discusses the results of the four perceptual measures. A correlation matrix showing Spearman correlations between these measures can be seen in Fig. 5.

3.1. Intelligibility

The main effect of the listener group was not significant: $F(1, 24.6) = 1.56$, $p = 0.223$ (see Table 1.). That is, there was no difference between intelligibility evaluation of trained and untrained listeners. Participants evaluated four types of dysarthria depending on its severity. Regarding the intelligibility of dysarthric speech, severity of dysarthria had significant impact on the evaluation given by listeners: $F(3, 16.4) = 19.19$ $p < 0.001$.

Table 1. The effect of listener group and dysarthria severity on intelligibility ratings

Fixed Effect Omnibus tests

	F	Num df	Den df	p
Listeners	1.56	1	24.6	0.223
Dys_severity	19.19	3	16.4	<.001
Listeners ✻ Dys_severity	2.53	3	468.1	0.057

Note. Satterthwaite method for degrees of freedom

Table 2. ASR error rates (%) for the categories defined by speech-language therapists. [5]

	healthy		mild		moderate		severe	
ASR engine	**WER**	**CER**	**WER**	**CER**	**WER**	**CER**	**WER**	**CER**
Whisper-large-v3	19.3	11.1	70.8	41.5	85.1	48.1	100.8	73.7
MMS-1B-FL102	31.5	8.9	89.4	62.4	94.1	74.0	98.9	78.5
BEAST2	14.3	3.8	62.8	36.6	76.0	49.7	88.3	65.9
FastConformerHU	6.5	2.1	45.2	20.7	55.8	26.4	82.7	51.5

3.2. Speech severity

The main effect of the listener group was significant: $F(1, 24.8) = 5.42$, $p = 0.028$ (see Table 4.). There was a difference between the judgements of trained and untrained listeners. Their evaluation was different depending on the severity of dysarthria: $F(3, 16.8) = 44.02$, $p < 0.001$. However, their judgements were not the same in each subgroup (see also Fig. 2.).

In the case of mild dysarthria, there was a greater difference between the evaluation of trained and untrained listeners: $M_{trained} = 0.38$, $M_{untrained} = 0.29$. In the case of moderate dysarthria, trained listeners gave a significantly higher severity score than untrained listeners: $M_{trained} = 0.65$, $M_{untrained} = 0.51$, $p = 0.045$. In the case of severe dysarthria, there was also a difference between the evaluation of trained and untrained listeners: $M_{trained} = 0.88$, $M_{untrained} = 0.78$.

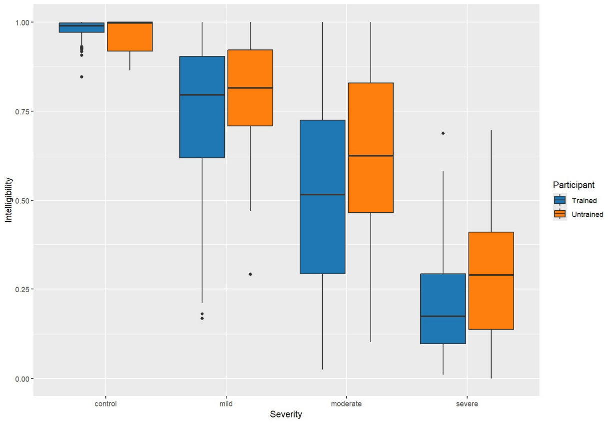

Figure 2. Boxplot showing trained and untrained listener evaluations

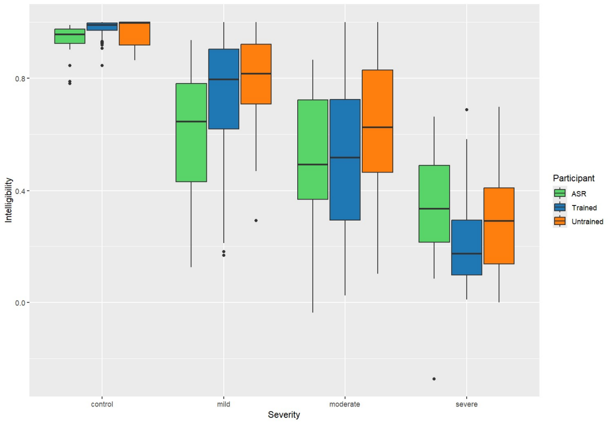

Figure 3. Boxplot showing trained and untrained listener evaluations compared to ASR scores

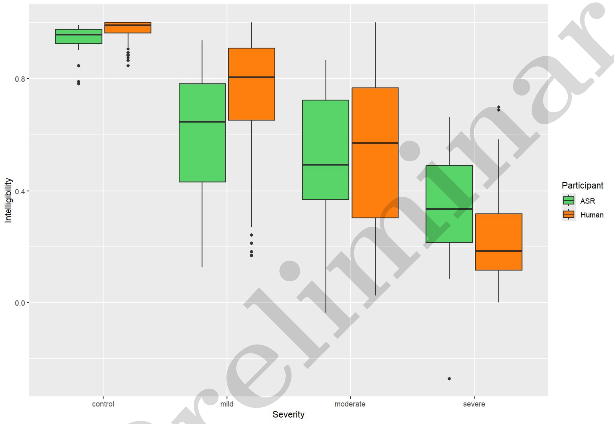

Figure 4. Boxplot comparing human listener evaluations to ASR scores

Table 3. Mean intelligibility scores of ASR systems compared to human listeners

Severity	ASR	Trained	Untrained
Control	0.935	0.977	0.962
Mild	0.596	0.731	0.797
Moderate	0.504	0.520	0.609
Severe	0.325	0.212	0.276

Table 4. The effect of listener group and dysarthria severity on severity ratings

Fixed Effect Omnibus tests

	F	Num df	Den df	p
Listeners	5.42	1	24.8	0.028
Dys_severity	44.02	3	16.8	<.001
Listeners ✻ Dys_severity	5.50	3	468.1	0.001

Note. Satterthwaite method for degrees of freedom

3.3. Speech naturalness

The main effect of the listener group was not significant: $F(1, 24.8) = 0.80$, $p = 0.381$ (see Table 5.). There was no difference between the judgement of speech naturalness given by trained and untrained listeners.

The severity of dysarthria had a significant effect on perceived speech naturalness: $F(3, 18) = 48.56$ $p < 0.001$. This means that listeners evaluated speech naturalness differently depending on the severity of the speech disorder. Differences between the evaluations of trained and untrained listeners were the same in each dysarthria subgroup.

Table 5. The effect of listener group and dysarthria severity on speech naturalness ratings

Fixed Effect Omnibus tests

	F	Num df	Den df	p
Listeners	0.795	1	24.8	0.381
Dys_severity	48.555	3	18.0	<.001
Listeners✻ Dys_severity	1.717	3	468.1	0.163

Note. Satterthwaite method for degrees of freedom

3.4. Listener effort

The main effect of the listener group is close to being significant: $F(1, 24.9) = 4.22$, $p = 0.051$ (see Table 6.). The main effect of dysarthria severity was significant: $F(3, 16.7) = 23.00$, $p < 0.001$. Participants evaluated listener effort differently depending on the severity of the speech disorder. The interaction of listener group and dysarthria severity was also significant: $F(3, 468.1) = 4.04$, $p = 0.007$. This means that differences between the evaluations of trained and untrained listeners were not the same in each dysarthria subgroup.

In the case of the control group, the same amount of mental effort was attributed by both trained and untrained listeners: $M_{trained} = 0.02$, $M_{untrained} = 0.03$. Considering mild dysarthria, there was a greater difference between the evaluations of the two listener groups: $M_{trained} = 0.31$, $M_{untrained} = 0.21$. As for moderate and severe dysarthria, trained listeners attributed more mental effort to speech perceiption: $M_{trained} = 0.54$, $M_{untrained} = 0.41$; $M_{trained} = 0.81$, $M_{untrained} = 0.72$.

Table 6. The effect of listener group and dysarthria severity on listener effort ratings

Fixed Effect Omnibus tests

	F	Num df	Den df	p
Listeners	4.22	1	24.9	0.051
Dys_severity	23.00	3	16.7	<.001
Listeners ✻ Dys_severity	4.04	3	468.1	0.007

Note. Satterthwaite method for degrees of freedom

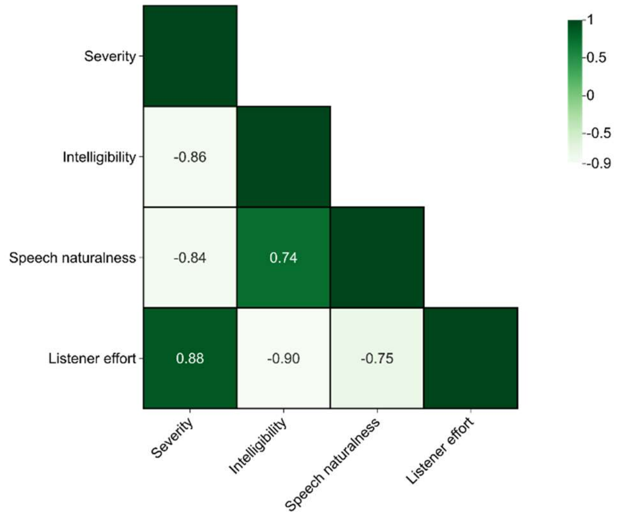

Figure 5. Correlation matrix showing Spearman correlations between the four perceptual measures

4. Discussion

We found a correlation between the severity of dysarthria and the evaluation of intelligibility, severity, speech naturalness, and listener effort in both cases.

The performance of automatic speech recognition (ASR) systems depends on the severity of dysarthria: as the severity of dysarthria increased, a higher character error rate was measured. ASR systems gave lower intelligibility scores than human listeners in all cases, except for severe dysarthria (see Table 3.).

Human agents' evaluation also depends on the severity of dysarthric speech. However, human agents show a great variability in the evaluation of intelligibility in mild and moderate cases (see Fig. 2, 3, 4.). Their performance correlates with the degree of severity.

Trained participants evaluated severity, intelligibility and listener effort in a more rigorous manner than untrained participants. The more severe the speech disorder, the greater the difference between the ratings of trained and untrained participants.

5. Conclusion

There are only a few papers in the literature dealing with the evaluation of intelligibility, severity, speech naturalness, and listener effort in between listener groups in the case of dysarthria. As it is emphasized in the main goal of our project, it has great importance in human interaction to be able to speak in an adequate manner, which can be quite challenging for people with dysarthria. The impact of their motor speech disorder in their everyday language use can cause (severe) problems in their intelligibility. Depending on their communication partner's listener effort, there can be different outcomes for these types of verbal interactions.

One of our main goals in the framework of the 'Dysarthria' project at the University of Szeged is to improve the intelligibility of dysarthric speech using automated tools. However, we face both theoretical and practical problems when talking about intelligibility. That is, we are working with three different viewpoints on intelligibility. First, dysarthric speech is evaluated perceptually by trained clinicians. Second, dysarthric speech should be understood by a wide range of listeners, including untrained listeners. Third, we aim to use automated tools to improve intelligibility. However, these three agents may have a different understanding of intelligibility itself. That is why we found it important to explore the differences between the intelligibility judgements given by these three groups.

6. Acknowledgements

The authors of the paper are very grateful for Márk Grastyán for his help with statistical analyses. We also thank the participants of the study for their participation.

References

[1] J. Tröger, F. Dörr, L. Schwed, N. Linz, A. König, T. Thies, M. T. Barbe, J. R. Orozco-Arroyave, and J. Rusz: "An automatic measure for speech intelligibility in dysarthrias-validation across multiple languages and neurological disorders," *Frontiers in Digital Health*, vol. 6, 1440986, 2024. https://doi.org/10.3389/fdgth.2024.1440986

[2] B. Dam and L. Ivaskó: "Distinguishing between dysarthria types based on acoustic parameters," *Beszédtudomány – Speech Science*, vol. 4, no. 1, pp. 118–135., 2024.

[3] L. Ivaskó and B. Dam: "Egy dysarthriás beszédet tartalmazó korpusz építésének dilemmái

[Dilemmas of building a dysarthria corpus]," *Általános Nyelvészeti Tanulmányok*. Accepted.

[4] D. Terbe, L. Tóth, and L. Ivaskó, "Hangkonverzió alkalmazása dysarthriás betegek beszédminőségének javítására [Applying voice conversion to improve the voice quality of dysarthric patients]," in *Proc. of VXIII*th *Conference on Hungarian Computational Linguistics*, (Szeged, Hungary), pp. 161–174, 2022.

[5] Á. Berta, B. Dam, L. Tóth, and L. Ivaskó, "Evaluating and Improving the Intelligibility of Hungarian Dysarthric Speech Using Foundation Models," in *Proc. of XXII*th *Conference on Hungarian Computational Linguistics*, (Szeged, Hungary), pp. 187–200, 2026.

[6] Z. A. Jenei, G. Kiss, and D. Sztachó, "Detection of speech related disorders by pretrained embedding models extracted biomarkers," in *Proc. SPECOM 2022*, (Gurugram, India), pp. 279–289, 2022.

[7] B. Dam, Á. Berta, L. Tóth, and L. Ivaskó, "Ways of speech intelligibility evaluation – a comparison between human agents and ASR systems," presented at *13*th *Novi Sad Workshop on Psycholinguistic, Neurolinguistic and Clinical Linguistic Research*, (Novi Sad, Serbia), 2026.

[8] J. R. Duffy, *Motor speech disorders*. St. Louis, Mo.; London: Elsevier Mosby, 2013.

[9] J. L. Manes et al., "Perceptual Outcome Measures for Supplementing Transcription Intelligibility in Speakers with Parkinson's Disease and Mild Dysarthria," *Folia Phoniatrica et Logopaedica*, pp. 1–14, Feb. 2026, doi: https://doi.org/10.1159/000550174.

[10] J. Peirce et al., "PsychoPy2: Experiments in behavior made easy," *Behavior Research Methods*, vol. 51, no. 1, pp. 195–203, Feb. 2019, doi: https://doi.org/10.3758/s13428-018-01193-y.

[11] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and Ilya Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," Dec. 2022, doi: https://doi.org/10.48550/arxiv.2212.04356.

[12] V. Pratap et al., "Scaling Speech Technology to 1,000+ Languages," arXiv.org, May 22, 2023. https://arxiv.org/abs/2305.13516

[13] M. S. Kádár, G. Dobsinszki, K. Mády, and P. Mihajlik, "„Feeding the BEAST" – A BEA Speech Transcriber továbbfejlesztése és integrálása neurális nyelvmodellel [„Feeding the BEAST" extensions to the BEA speech transcriber, including its integration with a neural language model]," in *Proc. MSZNY*, pp. 135–143, 2023.

[14] P. Mihajlik, É. Székely, P. Barta, S. M. Kádár, G. Dobsinszki, and L. Tóth, "Improved dysarthric speech to text conversion via TTS personalization," in *Proc. EUSIPCO*, pp. 521–525, 2024.