Interspeech 2022
18-22 September 2022, Incheon, Korea

Comparison of 5 methods for the evaluation of intelligibility in mild to moderate French dysarthric speech

Fougeron C.[1], Audibert N. [1], Kodrasi I. [2], Janbakhshi P. [2,7], Pernon M. [1, 3], Lévêque N. [1,4], Borel S. [4,5], Laganaro M.[6], Bourlard H. [2,7], Assal F. [8]

[1]LPP, UMR7018, CNRS/U. Sorbonne-Nouvelle, Paris, France
[2]IDIAP, Martigny, Switzerland
[3] CRMR Wilson, Hôpital Fondation A. de Rothschild, Paris, France
[4]APHP, Pitié-Salpétrière, Paris, France
[5] Institut du Cerveau, Sorbonne Université, Paris, France
[6] Université de Genève, FPSE, Switzerland
[7] EPFL, Lausanne, Switzerland
[8] HUG, Geneva, Switzerland
{first.lastnames}@sorbonne-nouvelle.fr, @idiap.ch

Abstract

Altered quality of the phonetic-acoustic information in the speech signal in the case of motor speech disorders may reduce its intelligibility. Monitoring intelligibility is part of the standard clinical assessment of patients. It is also a valuable tool to index the evolution of the speech disorder. However, measuring intelligibility raises methodological debates concerning: the type of linguistic material on which the assessment is based (non-words, words, continuous speech), the evaluation protocol and type of scores (scale-based rating, transcription or recognition tests), and the advantages and disadvantages of listener vs. automatic-based approaches (subjective vs. objective, expertise level, types of models used). In this paper, the intelligibility of the speech of 32 French patients presenting mild to moderate dysarthria and 17 elderly speakers is assessed with five different methods: impressionistic clinician judgment on continuous speech, number of words recognized in an interactive face-to-face setting and in an on-line testing of the same material by 75 judges, automatic feature-based and automatic speech recognition-based methods (both on short sentences). The implications of the different methods for clinical practice are discussed.

Index Terms: intelligibility, assessment method, reliability, clinical application.

1. Introduction

Intelligibility is the degree to which a speech signal can be deciphered for the intended message to be recovered (e.g. [1]). In the context of speech disorders, measuring intelligibility is part of the assessment of the speaker's impairment profile, since intelligibility is one of the dimensions which can be altered. Measuring intelligibility is also a way to assess the impact of the speech disorder on the patient's quality of life, as an index of communication impairment and a way to guide the patient's management. Many studies have been devoted to the comparison of different methods to assess speech intelligibility [1, 2, 22]. Each of those has its advantages and disadvantages, which typically arise due to the choices adopted in the method design. In clinical practice, a global intelligibility score rated on a simple scale is often part of the standard exam of the patient. This gross score usually encompasses intelligibility, comprehensibility and a grading of the severity of the speech disorders. For speech therapy assessments, more standardized procedures are often used. They are most often based on a transcription or recognition test in which the number of correctly recognized items gives an estimate of the patient's intelligibility. The nature of the items used gives rise to different definitions of 'intelligibility'. For instance, the recognition of non-words produced by the patients relies on a pure acoustic-phonetic decoding of the speech signal. The recognition of isolated words relies on the decoding of the signal but also on the use of top-down information linked to lexical frequency and lexical competition. For the recognition of words in continuous speech, contextual information also participates in the recovery of the message. Hence, the definition of intelligibility varies from a pure acoustic-phonetic decoding to a more functional notion, and to contextual intelligibility and comprehensibility [3].

For these different methods, intelligibility scores are usually obtained from human responses, either from a single judge in a clinical setting for instance or by a jury in an experimental set-up. In both cases, this evaluation is highly subjective. With the progress of speech technologies and the crucial need of objective, time- and cost-efficient methods for the evaluation of intelligibility, several automatic measures have been proposed. These methods would be invaluable to augment the clinician's assessment for on-line monitoring of the evolution of a disorder or of the efficiency of a treatment. In the past decade, several automatic methods have been proposed which can be broadly be categorized into feature-based and automatic speech recognition (ASR)-based measures (see [4]). Feature-based measures typically refer to the blind assessment of speech intelligibility by extracting acoustic features that are potentially indexing altered speech dimensions. Using feature selection and regression training, an intelligibility measure is then derived ([5, 6, 7], see also [8] for a different approach based on i-vectors). In ASR-based measures, ASR systems are trained on large databases of healthy speech signals and applied to the patient's signal. The word error rate is then used to derive the patient's intelligibility [9, 10]. Although there has been significant progress in developing automatic intelligibility measures, current measures still face major challenges such as lack of applicability to several types of impairments, the requirement for a large amount of training data and for phonetically balanced speech material between speakers [4].

10.21437/Interspeech.2022-10590

The goal of the present study is to compare the assessment of the intelligibility of 49 French speakers, presenting mild to moderate dysarthria or no dysarthria, with five different methods, 3 listener-based methods and 2 automatic methods.

2. Method

A population of 49 French male and female speakers was selected for this investigation. They all participated in the evaluation of the MonPaGe protocol [11, 12], which is designed for a quantitative assessment of the speech of patients presenting speech motor disorders along several dimensions, with intelligibility being one of them. The population is described in Table 1. It includes 32 speakers presenting mild to moderate dysarthria and 17 elderly speakers with no attested dysarthria. Dysarthria associated with four etiologies are included in the pathological groups, with various severity levels ranging from mild to moderate dysarthria. The severity level was indexed by the Perceptual Scores of the BECD [13] rated by one expert clinician on a 20 point scale (0 = normal).

2.1. Listener-based evaluation methods

MonPaGe 'face2face' intelligibility rating: The "face2face" evaluation of intelligibility corresponds to the standard intelligibility testing of the MonPaGe protocol. Taking place at the beginning of the session, a short intelligibility test is administered in the form of an interactive task between the experimenter and the participant in a face-to-face setting. The participant is asked to instruct the experimenter to place some test-words on a 5x5 grid of shapes and colors. The participant, but not the experimenter, sees the test-word and its associated location on the computer screen. The experimenter has to write the test-word that they heard on the corresponding colored shape on a paper grid. The final intelligibility score of the patient is computed based on the number of correctly understood test words.

For each session/speaker, a randomization procedure included in the MonPaGe software allows for the random extraction of 15 target words and 15 locations (colored shapes) on the grid. Test-words are drawn from a database of 437 picturable French words, where each word has 1 to 6 competitors within the database and possibly more in the French lexicon. Competitors are phonologically similar words, organized in 5 subsets of contrasts: place of articulation, voice, manner, nasality/cluster and vowel. Three words are randomly selected from these 5 subsets in order to have 15 test-words selected for each session, which are randomly assigned to a location and presented one by one to the patient on the computer screen. Real French words were chosen over pseudo-words in order to facilitate the testing of cognitively impaired patients and to allow the presentation of the test items in a written and picture form. The speaker is instructed to always give the directive to the experimenter with the same, pre-learned, carrier sentence: "Place the word [target_word] on the [color] [shape]" (e.g. 'Place the word dog on the red circle'). This sentence allows for the presentation of the target in a continuous speech flow but with a control of undesired contextual influence or predictability (e.g. the word is not preceded by an article). The interactive set-up allows to test intelligibility in a communicative situation instead of read speech. Experimenters are instructed to always write something on the grid in order to not discourage unintelligible patients, or to induce artificial 'extra' hyperarticulation. No more than two responses are allowed in case of doubt (e.g. 'pale'/'male'). Ratings presented here include sessions of 9 different experimenters, each having assessed 1 to 11 speakers (each speaker being assessed by only one experimenter). Errors in the location on the grid are not considered. A rating of 1 is given to single correct response (i.e. accurate identification of the test-word), 0 to incorrect responses, and 0.5 for one correct response when two responses were provided. In the following comparison, we use the intelligibility scores per speaker/participant computed as the average of the scores for the 15 test words. It should be noted that the speakers produced 15 stimuli each, except for 5 speakers with only 14 stimuli.

Multi-judge audio-only word transcription: The "multi-judge" evaluation corresponds to the assessment of the intelligibility of the sentences recorded during the MonPaGe evaluation by a pool of 75 judges. The test was administrated online and judges were instructed to listen (using headphones and in a quiet room) to the test-words in the carrier sentences and to transcribe orthographically the test-word they understood (e.g. <dog> in 'Place the dog on the red circle'). The 15 stimuli of each speaker were presented in succession and could be listened to only once. As for the face2face test, the judges were instructed to always give a response with a maximum of two possible words. Due to the large number of stimuli and to reduce the test to 30 minutes, a full cross design across judges was used only for a subset of the data. This way, 4 dysarthric speakers (extracted from each group and with similar severity) were rated by the 75 judges, while the remaining speakers were split in 5 groups of 9 speakers each which were rated by 15 judges. This resulted in a total of 14430 ratings with each judge rating only 13 speakers (9 shared with the group and 4 shared with all raters). The order of presentation of the speakers was randomized for each judge.

Fifty-nine female and sixteen male judges (19 to 53 years old), all native French speakers, participated in the experiments. Their familiarity with speech disorders varied, but no difference was found between familiarity levels ($\chi^2(2)=2.01$, $p=.37$), therefore the 75 raters were grouped for further analysis.

Table 1: *Speakers' distribution by sex, age, etiologies and severity assessed by BECD Perceptual Score (PS) on a 0-20 scale, 0=normal. Age and PS specified as mean <min-max>.*

Population	N	age	PS
Friedreich Ataxia	4 f, 4 m	39.5 <29-50>	11.3 <7-16>
Parkinson disease	2 f, 6 m	59.6 <49-70>	6.8 <3-12>
Amyotrophic Lateral Sclerosis	2 f, 6 m	55 <45-61>	7.8 <5-10>
Wilson disease	8 m	34.9 <25-49>	9 <6-12>
Healthy elderly	10 f, 7 m	81.8 <77-88>	1.76 <0-5>

Expert global rating on continuous speech: The 'expert' assessment is a scoring of the participants' intelligibility provided by an experienced speech pathologist on recordings of the speakers' continuous speech. This material, recorded for each participant during the MonPaGe assessment, includes the reading of a short text and more spontaneous production in a picture description task. Intelligibility was scored together with other speech dimensions (not presented here) on a 4-point scale.

2.2. Machine-based evaluation methods

Feature-based measures: It has been shown that impaired speech intelligibility arises due to several impaired dimensions

of speech, such as long-term temporal dynamics, prosody, and voice quality [14]. To characterize these impaired speech dimensions, several features are used, i.e., low-to-high modulation energy ratio (LHMR), voiced percentage, range and kurtosis of the fundamental frequency (f0), as well as the mean and range of jitter and shimmer [14, 15]. These 8 features were extracted from the complete signal (carrier sentence + target word) available for each speaker.

ASR-based measures: To consider articulatory impairments in addition to the phonation impairments reflected in the previously described features, an ASR system based on a DNN-HMM acoustic model with fMLLR-adapted features is trained using the Kaldi toolkit [16]. The system is trained on the SpeechDat corpus and on our own recordings of telephone speech. It should be noted that in this paper, we are not interested in improving the absolute performance of the ASR system, but rather on the relative performance differences of the ASR system for the different speakers. Hence, no acoustic or language model adaptation of the ASR system to the MonPaGe speech material was done. Two ASR-based measures are considered in this work, i.e., the word error rate (WER) computed on the complete signal (carrier sentence + target word) and the target word accuracy (TWA) (i.e., the recognition accuracy of the target words only).

Table 2: *Intelligibility in % (mean <min-max>) in different listener-based methods, per sub-population.*

Sub-populations	I.face2face	I.multi-judge	I.expert
Friedreich Ataxia	85 <67-100>	84 <60-98>	53 <0-75>
Parkinson disease	99 <93-100>	93 <80-99>	88 <50-100>
Amyotrophic Lateral Sclerosis	94 <87-100>	90 <83-97>	84 <50-100>
Wilson disease	88 <60-100>	83 <60-93>	72 <50-100>
Healthy elderly spk	98 <93-100>	93 <68-100>	100
all speakers	*94 <60-100>*	*90 <60-100>*	*83 <0-100>*
non-fully intelligible speakers	*90 <60-93>*	*85 <60-100>*	*74 <0-100>*

Composite measure via regularized linear regression: In addition to the individual feature-based and ASR-based measures, we also consider a composite measure computed as a linear combination of all measures. Given the small amount of available data and speakers, we use regularized linear regression on a 5-fold cross-validation framework to find the optimal weights of the individual measures for the composite measure. The individual measures are normalized in each training fold and the regularization parameter in each fold is optimized on the training set of the fold. The performance is then assessed as the average performance on the validation set across all folds.

3. Results

Preliminary analysis across etiologic groups revealed that speakers rated as 100% intelligible in one of the five methods are found in all sub-groups. This confirms that not every dysarthric patient has an intelligibility impairment. Least intelligible speakers did not show scores lower than 60% in the listener-based word recognition tasks as shown in Table 2. Therefore, even for the more severe cases in this pool of mild to moderate dysarthria, sufficient information in their speech signals enable listeners to correctly identify more than half of the test-words. Interestingly, non-fully intelligible patients were also found in the healthy elderly speakers' group. In order to compare further the methods and see how ratings can be linked to this distribution between fully intelligible and less intelligible speakers, the population was split into 2 groups according to a cut-off score of 94%, as determined by the "multi-judge" rates. This threshold was determined with regard to the cut-off between healthy and dysarthric speakers. 30 speakers (including 4 healthy speakers and speakers of the 4 dysarthria groups) constitute this non-fully intelligible group.

3.1. Comparison between listener-based methods

Before turning to the comparison with other methods, the reliability of the "multi-judge" method was tested. Moderate inter-judge agreement was found, with an ICC of .69 for the five groups of 15 listeners, and an ICC of 0.64 for the ratings of the 75 listeners on the shared subset. As found in other studies (e.g. [17]), inter-rater agreement was found to decrease with the severity of the dysarthria (r=.54).

The three listener-based methods are first compared with pair-wise Spearman's rank correlations to evaluate the strength of the association between the scores obtained in the different methods. Over the entire population, scores in the "face2face" and "multi-judge" methods are strongly correlated (ρ=.72), while the relationship between the scores of the "expert" method and that of the "face2face" (ρ=.63) or the "multi-judge" (ρ=.58) is a bit lower. When computed on the non-fully intelligible set of speakers, correlations decrease slightly between the "face2face" scores and that of the "multijudge" (ρ=.64) or of the "expert" (ρ=.58) methods. It was expected that correlation will decrease according to the severity prevalence in the speaker's distribution: different rankings are more likely to occur for speakers that are non-fully intelligible. There is however a drastic reduction of the correlation between the "expert" and the "multi-judge" (ρ=.37) scores for these 30 non-fully intelligible speakers. The relationship between ratings based on a global estimate of the patient's intelligibility by an expert clinician and the scores based on a word transcription task depend on the severity of intelligibility impairments. This is also shown by the fact that "expert" scores highly depend on the dysarthria severity of the patient (ρ=-.87), while this relationship is not that strong for the other two methods (ρ=-.58 for face2face*severity, ρ=-.54 for inter-judge*severity).

Discrepancies between methods are also found in the average intelligibility level for the different population as shown in Table 2. Across all speakers, intelligibility scores in the "face2face" setting are higher, as well as for most of the sub-populations and for the non-fully intelligible group. The "multi-judge" method yields slightly lower scores than the "face2face" method, while the scores obtained with the "expert" method are on average 10% lower. For the ataxic group, the scores drop by 30%. Judgment with the "expert" method thus appears to be more severe than with the other two-listener based methods, especially for more severely impaired dysarthria group (i.e. the Friedreich Ataxia group).

Finally, differences between methods also have an impact on the number of speakers considered fully intelligible. Out of 49 speakers, 25 speakers were rated as 100% intelligible with the "face2face" method, 27 with the "expert" ratings, but only 3 with the "multi-judge" method.

3.2. Automatic- vs. listener-based methods

To investigate the applicability of automatic measures in indexing the subjective intelligibility derived from different methods, correlation between automatic methods (feature-based, ASR-based, and composite measures) and the three listener-based methods are used. Table 3 presents the correlation values, with the highest (absolute) correlation coefficient in each category and for each listener-based method presented in bold. Most correlation values presented in Table 3 are statistically significant. Overall, feature-based measures yield better, although moderate, correlations ($\rho \approx .50$) with the "face2face" and "expert" intelligibility ratings, whereas ASR-based measures yield better correlations ($\rho \approx .68$) with the "multi-judge" intelligibility ratings.

Table 3: *Spearman's correlations between automatic measures and the three listener-based methods.*

Measure	face2face	multi-judge	expert
LHMR	-0.35	-0.22	**-0.51**
voice percentage	-0.15	-0.3	0.15
f0 range	-0.29	-0.29	-0.17
f0 kurtosis	0.31	0.29	0.41
jitter mean	**0.51**	0.26	0.41
jitter range	-0.46	**-0.38**	-0.47
shimmer mean	0.48	0.16	0.44
shimmer range	-0.39	-0.13	-0.38
WER	-0.39	**-0.68**	**-0.57**
TWA	**0.41**	0.67	0.42
composite	**0.70**	**0.70**	**0.74**

On the one hand, the stronger association between the feature-based measures and the "face2face" and "expert" ratings may reflect the influence of the various impaired speech dimensions quantified by these features on the listener's perception in these two settings. More specifically, the highest correlation (in the order of 0.50) with the "face2face" and "expert" ratings is achieved using the mean of jitter and LHMR, showing that voice instability and impaired temporal dynamics may contribute to the judgment of the listener when interacting with the speaker (as in the "face2face" method) or of the clinician assessing a larger amount of continuous speech (as in the "expert" method). On the other hand, the performance of the ASR system is found to be a reliable indicator of the word recognition score of an 'average' listener who has only access to the audio signal (as in the "multi-judge" setting). Correlations of the "multi-judge" ratings with both TWA and WER are strong, and, as expected, no statistically significant differences between TWA and WER are found. Although the correlation values of individual measures presented here might not be impressively high, it should be noted that these measures are being used on a small amount of speech data, and on speakers with mild to moderate speech impairment. Interestingly, while different individual measures provide a reasonable correlation with different listener-based ratings, Table 3 shows that by combining different measures into a composite one via regularized linear regression yields a significant performance improvement.

4. Discussion

The five methods compared in this paper present different ways of measuring how the transmission of oral messages may be compromised by an altered quality of the phonetic-acoustic signal. Over the entire population tested, the methods present comparable results, with most correlations around 0.6/0.7. Nonetheless, discrepancies between methods are found, especially for the more severe intelligibility impairments, and they can be related to the distinctive aspects of these methods.

In the "face2face" setting, intelligibility may have been improved by the use of cues other than acoustic cues, which are not present in the "multi-judge" or "expert" settings. Facial expressions, hand movements and other para-linguistic supplementation strategies help the decoding of the speaker's message (see [18, 19] for the benefit of visual cues) and yield better recognition scores. In this respect, this method provides a more ecologically valid index of functional intelligibility. Results of the "feature-based" methods and the strong correlations between the composite measure and the listener-based methods also support the fact that acoustic dimensions other than the ones directly linked to phonetic contrasts (e.g. voice quality is not contrastive in French) may improve intelligibility. Impaired intelligibility results from a combination of altered speech dimensions [1, 3, 15, 20, 21] and the performances of composite measures relying on several dimensions are particularly promising for the development of a reliable automatic intelligibility assessment measure. In the future, we would like to investigate if the performance of automatic measures can be further improved by incorporating additional (possibly visual) features, by incorporating an ASR system adapted to the speech task, by incorporating an ASR system that puts more weight on the recognition of acoustic features while restricting the amount of linguistic information, or by simply considering more data for each speaker.

Other factors may explain differences between methods. In the "face2face" method, the experimenters are exposed to the speaker's speech during the set-up of the assessment. In the "expert" method, the clinician also relies on a larger quantity of speech material. In the "multi-judge" and automatic methods, intelligibility scores are based on more limited amount of speech data. Perseveration and fatigue effects are known to affect listeners' ratings in experimental set-ups like the one used in the "multi-judge" method [1, 2, 17, 20]. These effects, together with listener-specific top-down effects as well as uncontrolled listening conditions, may explain the moderate inter-judge agreement found in the "multijudge" setting as well as discrepancies between the listener-based methods (especially for the less intelligible speakers). However, over the subjective nature of the judgments, it also reflects the fact that the decoding of the speech signal in everyday life and in clinical settings is a matter of both the speaker' and of the listener' performances. Automatic assessment of intelligibility has the great potential of complementing listener-based assessment of the functional intelligibility of the speaker with objective measures. While such automatic measures have been typically used on a larger amount and more diverse speech data, in this paper we show their applicability to a smaller amount of data, such as the one available with the MonPaGe protocol.

Acknowledgements

This study was supported by the Swiss FNS Grants N. CRSII5_173711 & 202228 and the ANR-10-LABX-0083. The authors would like to thank Srikanth Madikeri for his assistance with the ASR system.

References

[1] R. D. Kent, G. Weismer, J. F. Kent, and J. C. Rosenbek, "Toward phonetic intelligibility testing in dysarthria," *J Speech Hear Disord*, vol. 54, no. 4, pp. 482–499, Nov. 1989.

[2] 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.

[3] M. S. De Bodt, M. E. Hernández-Díaz Huici, and P. H. Van De Heyning, "Intelligibility as a linear combination of dimensions in dysarthric speech," *Journal of Communication Disorders*, vol. 35, no. 3, pp. 283–292, 2002.

[4] P. Janbakhshi, I. Kodrasi, H. Bourlard, "Pathological speech intelligibility assessment based on the short-time objective intelligibility measure," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019

[5] M. S. Paja and T. H. Falk, "Automated Dysarthria Severity Classification for Improved Objective Intelligibility Assessment of Spastic Dysarthric Speech," in *Proc. 13th Annual Conference of the International Speech Communication Association*, Oregon, USA, 2012, pp. 62–65.

[6] R. Hummel, W. Y. Chan, and T. H. Falk, "Spectral features for automatic blind intelligibility estimation of spastic dysarthric speech," in *Proc. 12th Annual Conference of the International Speech Communication Association*, Florence, Italy, 2011, pp. 3017–3020.

[7] T. H. Falk, R. Hummel, and W. Y. Chan, "Quantifying perturbations in temporal dynamics for automated assessment of spastic dysarthric speech intelligibility," in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, Prague, Czech Republic, 2011, pp. 4480–4483.

[8] I. Laaridh, C. Fredouille, A. Ghio, M. Lalain, and V. Woisard, "Automatic Evaluation of Speech Intelligibility Based on i-vectors in the Context of Head and Neck Cancers," in *Interspeech*, Hyderabad, India, 2018, pp. 2943–2947.

[9] M. Schuster, E. Nöth, T. Haderlein, S. Steidl, A. Batliner, and F. Rosanowski, "Can you understand him? Let's look at his word accuracy-automatic evaluation of tracheoesophageal speech," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Philadelphia, PA, USA, 2005.

[10] A. Maier, T. Haderlein, F. Stelzle, E. Nöth, E. Nkenke, F. Rosanowski, A. Schützenberger, and M. Schuster. "Automatic Speech Recognition Systems for the Evaluation of Voice and Speech Disorders in Head and Neck Cancer," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2009, Dec. 2009.

[11] C. Fougeron, V. Delvaux, M. Pernon, N. Léveque, S. Borel, P. Pellet, O. Bagou, R. Trouville, L. Ménard, S. Catalano, U. Lopez, T. Kocjancic-Antolik, and M. Laganaro. "MonPaGe : un protocole informatisé d'évaluation de la parole pathologique en langue française,". In *Actes du colloque UNADREO Orthophonie et technologies innovantes (Joyeux N. & Topouzkhanian S., eds)*, 2016.

[12] C. Fougeron, V. Delvaux, L. Ménard, and M. Laganaro, "The MonPaGe_HA database for the documentation of spoken French throughout adulthood," in *Proc. 11th International Conference on Language Resources and Evaluation*, Miyazaki, Japan, 2018.

[13] P. Auzou, and V. Rolland-Monnoury. *Batterie d'Evaluation Clinique de la Dysarthrie*. Isbergues : Ortho Edition, 2006.

[14] T. H. Falk, W. Y. Chan, and F. Shein, "Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility," *Speech Communication*, vol. 54, no. 5, pp. 622–631, Jun. 2012.

[15] C. Fang, H. Li, L. Ma, and M. Zhang, "Intelligibility Evaluation of Pathological Speech through Multigranularity Feature Extraction and Optimization," *Computational and Mathematical Methods in Medicine*, vol. 2017, Jan. 2017.

[16] D. Povey, A. Ghoshal, G. Boulianne, N. Goel, M. Hannemann, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.

[17] W. Ziegler and A. Zierdt, "Telediagnostic assessment of intelligibility in dysarthria: a pilot investigation of MVP-online," *J Commun Disord*, vol. 41, no. 6, pp. 553–577, Dec. 2008.

[18] L. Hunter, T. Pring, and S. Martin, "The use of strategies to increase speech intelligibility in cerebral palsy: an experimental evaluation," *Br J Disord Commun*, vol. 26, no. 2, pp. 163–174, Aug. 1991.

[19] C. K. Keintz, K. Bunton, and J. D. Hoit, "Influence of Visual Information on the Intelligibility of Dysarthric Speech," *American Journal of Speech-Language Pathology*, vol. 16, no. 3, pp. 222–234, Aug. 2007.

[20] K. C. Hustad, "Estimating the intelligibility of speakers with Dysarthria," *Folia Phoniatrica et Logopaedica*, vol. 58, no. 3, pp. 217–228, Feb. 2006.

[21] F. L. Darley, A. E. Aronson, and J. R. Brown, "Differential Diagnostic patterns of dysarthria," *Journal of Speech, Language, and Hearing Research*, vol. 12, no. 2, pp. 246–269, Jun. 1969.

[22] Lehner, K., & Ziegler, W. (2021). The impact of lexical and articulatory factors in the automatic selection of test materials for a web-based assessment of intelligibility in dysarthria. *Journal of Speech, Language, and Hearing Research*, 64(6S), 2196-2212