Utterance Verification-Based Dysarthric Speech Intelligibility Assessment Using Phonetic Posterior Features

Julian Fritsch and Mathew Magimai-Doss

Abstract—In the literature, the task of dysarthric speech intelligibility assessment has been approached through development of different low-level feature representations, subspace modeling, phone confidence estimation or measurement of automatic speech recognition system accuracy. This paper proposes a novel approach where the intelligibility is estimated as the percentage of correct words uttered by a speaker with dysarthria by matching and verifying utterances of the speaker with dysarthria against control speakers’ utterances in phone posterior feature space and broad phonetic posterior feature space. Experimental validation of the proposed approach on the UA-Speech database, with posterior feature estimators trained on the data from auxiliary domain and language, obtained a best Pearson’s correlation coefficient ($r$) of 0.950 and Spearman’s correlation coefficient ($\rho$) of 0.957. Furthermore, replacing control speakers’ speech with speech synthesized by a neural text-to-speech system obtained a best $r$ of 0.931 and $\rho$ of 0.961.

Index Terms—Dysarthric speech, objective intelligibility assessment, posterior features, utterance verification.

I. INTRODUCTION

DYSARTHRIA is a motor speech disorder resulting from damage to either or both the central and peripheral nervous systems [1], [2]. Such a damage can affect the speech production at various levels such as respiration, phonation, resonance, articulation, speaking rate, and prosody, leading to reduction in speech intelligibility. Assessment of speech intelligibility helps in characterizing the level of severity and in guiding speech therapy, treatment and intervention [1]. Currently, dysarthric speech intelligibility assessment is carried out through subjective listening tests, which is costly (in terms of both time and money); is susceptible to listener biases; and can be irreproducible. Objective speech intelligibility assessment is a potential alternative.

Previous work on objective dysarthric speech intelligibility assessment can be broadly grouped as:

i) assessment without explicit use of linguistic information: Legendre et al. proposed prediction of intelligibility using amplitude modulation spectra [3]. In [4], Falk et al. investigated modeling of short- and long-term temporal dynamics information. In [5], inspired from the notion that intelligibility can be expressed as a linear combination of perceptual dimensions phonation, nasality, articulation and prosody [6], a signal processing-based composite measure was proposed. Janbakshi et al. proposed the P-ESTOI measure [7], which builds upon the speech intelligibility measures STOI [8] (short-time objective intelligibility) and extended-STOI [9]. Different subspace-based methods such as iVector-based [10], use of spectral subspaces extracted through principal component analysis or approximate joint diagonalization [11] have also been proposed. The subspace methods assess intelligibility by measuring the deviation or distance between the control speech and dysarthric speech in the trained subspace.

ii) assessment based on explicit use of linguistic information: Kim et al. [12] proposed an approach where automatic speech recognition (ASR) with a confusion network is used to obtain “phone-to canonical-phone” mappings. These mappings are summarized in per-speaker histograms for a defined set of words and are then used to estimate an intelligibility score for each speaker. Middag et al. [13] proposed an approach where the dysarthric speech is aligned using an ASR system to obtain phone probabilities or phonological feature probabilities based confidences. These confidences are then accumulated over a specified groups of phones for each speaker to estimate intelligibility score. Finally, ASR system accuracy based intelligibility assessment has been also investigated [10], [14].

In recent years, phone posterior feature based speech assessment approaches have emerged, where sequences of phone posterior probabilities obtained from reference speech and test speech are matched for (a) speech codec and transmitted speech intelligibility assessment [15], (b) synthesized speech intelligibility assessment [15], and (c) degree of nativeness assessment [16]. Inspired by these works, the present paper develops an objective dysarthric speech intelligibility assessment approach. In this approach, the speech intelligibility of speakers with dysarthria is measured as percentage correct words spoken from a given set of words. The correctness of each word spoken is determined by verifying the utterance of a speaker with dysarthria against a set of control speakers’ utterances by matching the
Algorithm 1: Objective Intelligibility Score Estimation.

Set $w=1$, $\text{#CorrectWords}=0$.
while $w \leq W$ do
  Set $k=1$, vote count $v_{w}=0$.
  while $k \leq K$ do
    Match $Z^{w}$ and $Y_{k}^{w}$ to obtain a score $L^{w}$,
    Perform hypothesis testing based on $L^{w}$ to verify whether $Z^{w}$ and $Y_{k}^{w}$ are the same word or not;
    if same word then
      $v_{w} = v_{w} + 1$;
    end
  end
  if $v_{w} \geq K/2$ then
    $\text{#CorrectWords} = \text{#CorrectWords} + 1$;
  end
end
Result: $\text{IntScore} = \frac{\text{#CorrectWords}}{W} \times 100\%$

respective posterior feature sequences, and taking a majority voting. We validate the proposed approach on the UA-Speech corpus.

The remainder of the paper is organized as follows. Section II presents the proposed approach. Section III presents the experimental setup. Section IV presents the results and analysis. Finally, we conclude the paper in Section V.

II. PROPOSED APPROACH

In a clinical setting, dysarthric speech intelligibility can be assessed through an isolated word pronunciation test, where a speaker with dysarthria pronounces a set of isolated words, and the speech intelligibility is measured as percentage of correctly identified words by human listeners [1], [17], [18]. The proposed approach goes along that direction, where percentage correct words spoken by a speaker with dysarthria is estimated to assess speech intelligibility.

Let $w \in \{1, \ldots, W\}$ denote a word index $w$ from a set of words containing $W$ words. Let $k \in \{1, \ldots, K\}$ denote a control speaker index $k$ from the set of $K$ control speakers. Let $Z^{w}$ denote the speech produced for word $w$ by the speaker with dysarthria. Let $Y_{k}^{w}$ denote the speech produced for word $w$ by the control speaker $k$. Based on this information, Algorithm 1 presents the proposed objective intelligibility score $\text{IntScore}$ estimation method.

In the remainder of the section, we first present how $Z^{w}$ and $Y_{k}^{w}$ are matched to obtain match score $L^{w}$ and then present how hypothesis testing is performed to decide whether $Z^{w}$ and $Y_{k}^{w}$ are the same word or not.

A. POSTERIOR-FEATURE BASED MATCHING OF $Z^{w}$ AND $Y_{k}^{w}$

The match between $Z^{w}$ and $Y_{k}^{w}$ is obtained by matching posterior feature vector sequences $(z_{1}^{w}, \ldots, z_{n_{1}}^{w}, \ldots, z_{N}^{w})$ and $(y_{1}^{w}, \ldots, y_{m_{1}}^{w}, \ldots, y_{M_{k}}^{w})$, where $N$ denotes the number of frames in an utterance of a speaker with dysarthria of word $w$, $M_{k}$ denotes the number of frames in the $k^{th}$ control speaker’s utterance of word $w$, and posterior feature vectors $z_{n}^{w} = [z_{n,1}^{w}, \ldots, z_{n,m}^{w}, \ldots, z_{n,D}^{w}]^{T}$ and $y_{m}^{w} = [y_{m,1}^{w}, \ldots, y_{m,m}^{w}, \ldots, y_{m,D}^{w}]^{T}$ are $D$ dimensional phones or broad phonetic classes posterior probabilities estimated using neural network(s) (see Section III-B), $\forall n \in \{1, \ldots, N\}$ and $\forall m \in \{1, \ldots, M_{k}\}$.

The match between the two posterior feature sequences is obtained using dynamic time warping [19]. The dynamic programming recursion is as:

$$L^{w}(m, n) = l(y_{m}^{w}, z_{n}^{w}) + \min[L^{w}(m-1, n), \ L^{w}(m, n-1)],$$

where, $l(y_{m}^{w}, z_{n}^{w})$ is the local match score computed as symmetric Kullback-Leibler divergence between $y_{m}^{w}$ and $z_{n}^{w}$,

$$l(y_{m}^{w}, z_{n}^{w}) = \frac{1}{2} \left[ \sum_{d=1}^{D} y_{m,d}^{w} \log \frac{y_{m,d}^{w}}{z_{n,d}^{w}} + \sum_{d=1}^{D} z_{n,d}^{w} \log \frac{z_{n,d}^{w}}{y_{m,d}^{w}} \right],$$

and $L^{w}(m, n)$ is the accumulated match score at $(m, n)$. The dynamic programming results in a global match score $L^{w}(M_{k}, N)$, which is then normalized by the path length.

B. UTTERANCE VERIFICATION BASED ON $L^{w}(M_{k}, N)$

It can be argued that when the dysarthric speech is unintelligible, the uttered word tends to map to a word other than the target word. As a result, the listeners are not able to identify the target word. This could be formulated as an utterance verification problem, i.e. testing the hypothesis whether the speech utterances $Y_{k}^{w}$ and $Z^{w}$ correspond to the same word or not. A similar understanding has been recently applied to assess intelligibility of text-to-speech synthesis systems [20]. In the literature, it is well known that comparison of probability distributions using KL-divergence and other measures such as Bhattacharya distance is equivalent to hypothesis testing and yields an estimate of log-likelihood ratio [21], [22]. The global match score $L^{w}(M_{k}, N)$ is a sum of KL-divergence between phone or broad phonetic class posterior probability distributions on the best matching path normalized by the path length. So, $L^{w}(M_{k}, N)$ can be interpreted as an estimate of log-likelihood ratio of the test utterance being same as the reference utterance, through which utterance verification can be carried out. In order to do that, we need to apply a threshold on $L^{w}(M_{k}, N)$. As illustrated in Figure 1, the threshold is determined in the following manner:

1) Creating same word utterance pairs from the control speakers data, matching them and obtaining a distribution of global match score for the same word hypothesis;
2) Creating different word utterance pairs from the control speakers data, matching them and obtaining a distribution of global match score for NOT the same word hypothesis; and

3) determining the threshold at the intersection of the two distributions, referred to as \( T_{\text{inter}} \) or at the center of the two means of the histogram, referred to as \( T_{\text{cen}} \).

### III. EXPERIMENTAL SETUP

This section presents the experimental setup. In our experiments, we have used different off-the-shelf neural networks for posterior feature estimation and to synthesize control speech. Due to space limitations, their description in Sections III-B and III-C is kept short, the reader is referred to the supplementary material.

#### A. UA-Speech Database

We validated the proposed approach on the UA-Speech database [23]. The database consists of 15 English speakers with cerebral palsy (11 males, 4 females) and 13 healthy speakers (9 males, 4 females). Each impaired and control speaker has uttered 765 isolated words in total: 155 isolated words repeated 3 times and 300 isolated words spoken only once. In the database, each subject’s intelligibility score has been obtained by having five naive listeners (native speakers of American English) transcribe the isolated words and then calculating the average number of correct transcriptions. The subjective intelligibility scores of the patients range from 2% to 95%. Similar to the previous works [7], [11], we use the 5th channel recordings for our experiments. An energy-based voice activity detection using Praat ([24]) was used to extract the speech segments.

#### B. Posterior Feature Estimators

We investigated two different categories of posterior feature spaces: (a) phone posterior space and (b) broad phonetic or articulatory feature (AF) space to understand the posterior feature space that helps in characterizing dysarthric speech intelligibility well. To estimate posterior feature vectors \( x^m \) and \( y^m \) corresponding to phone classes or broad phonetic classes, a posterior feature estimator is needed. As collecting large amounts of data in a domain-dependent manner in a clinical environment is hardly possible, inspired from the previous works on speech intelligibility [15] and degree of non-native assessment [16], we investigated the use of posterior feature estimators trained on auxiliary domain data and auxiliary language.

**Phone space:** We used an off-the-shelf single hidden layer multilayer perceptron trained on 232 hours Switchboard conversational telephone speech to classify 44 context-independent phonemes and silence class, i.e. \( D = 45 \) [25].

**AF space:** There are different ways to represent phonemes as articulatory features such as binary features [26] or multi-valued features [27]. In this work, we conducted studies with binary features and multi-valued AF representations:

- **AF\textsubscript{binary}:** We used Phonet toolkit [28], which consists of 18 recurrent neural network-based binary AF classifiers trained on 17 hours of clean FM podcasts in Mexican Spanish. We extracted 18 AF binary probability vectors and used them as the posterior feature, i.e. \( D = 18 \times 2 \).

b) **AF\textsubscript{multi–manner}:** We used an off-the-shelf CNN-based estimator trained on AMI corpus with raw waveform as input to classify 9 multi-valued manner of articulation AF [29], i.e. \( D = 9 \).

#### C. Validation Studies

We obtained the thresholds \( T_{\text{inter}} \) and \( T_{\text{cen}} \) for each of the posterior spaces using all data from the 13 control speakers, as described earlier in Section II-B, and conducted three studies:

1) **all-control:** All control speakers in the UA-Speech database, i.e. \( K = 13 \), are used to obtain the objective score.

2) **single-synthetic-control:** Using a female voice speech synthesized by Tacotron2 [30] (an off-the-shelf neural text-to-speech system) for each of the words in the UA-Speech database as control speech. In this case, \( K = 1 \).

3) **vary-control:** Varying \( K \) from 13 to 1 and randomly selecting \( K \) control speaker(s) to obtain the objective score.

In all the studies, we used Pearson’s correlation coefficient, \( r \), and Spearman’s correlation coefficient, \( \rho \), as the evaluation measures, as done in the previous studies.

### IV. RESULTS AND ANALYSIS

#### all-control: Table I shows the results obtained for the case where all \( K = 13 \) control speakers’ speech is employed for \textit{IntScore} estimation. Under each of the correlation values, a \( p \)-value testing the hypothesis that the two sets of data are uncorrelated is also provided. Besides that, the table also presents the performances using other objective intelligibility assessment approaches proposed and studied on the same UA-Speech database in the literature. A brief overview of these approaches can be found in Section I. It is worth mentioning that the performance for composite measure [5], discriminant analysis [31], temporal dynamics [4] iVectors and word accuracy-based [10] studies are optimistic, as a part of the speaker dysarthria’s data has been used to create the models for intelligibility assessment.

It can be observed that the proposed approach consistently yields high Pearson’s and Spearman’s correlation coefficients for all the posterior feature spaces. Also, all the results are statistically significant. It is interesting to note that the choice

| Posterior feature space | \( T_{\text{cen}} \) | \( T_{\text{inter}} \) | \( r \) \( \rho \) | \( r \) \( \rho \) |
|------------------------|----------------|----------------|-------------|-------------|
| Phone                  | .939          | .939          | .959        | .957        |
| AF\textsubscript{binary} | 5.94e-7    | 2.31e-7      | 5.52e-8    | 2.29e-8    |
| AF\textsubscript{multi–manner} | 1.88e-6       | 1.35e-5      | 1.27e-6    | 1.32e-5    |
| Baseline systems       |               |               |             |             |
| P’ – ESTOI [7]         | .94          | .94         |             |             |
| Composite measure [5]  | .94          | .89         |             |             |
| Discriminant analysis [31] | .92          | -            |             |             |
| Spectral subspace [11] | -.83         | -.88        |             |             |
| Temporal dynamics [4]  | .87          | .85         |             |             |
| iVectors [10]          | .91          | -            |             |             |
| Word accuracy – based [10] | .89          | -            |             |             |
Table II presents the results obtained with the use of synthetic speech as reference. When compared to all-control case, we can observe that both for Phone space and AF multi−manner we obtain comparable \( r \) and \( \rho \), while slightly inferior \( r \) for AF binary. These results are promising. This indicates that in the proposed approach synthetic speech could be used as the control speech.

**Single-synthetic-control**: Table II presents the results obtained with the use of synthetic speech as reference. When compared to all-control case, we can observe that both for Phone space and AF multi−manner we obtain comparable \( r \) and \( \rho \), while slightly inferior \( r \) for AF binary. These results are promising. This indicates that in the proposed approach synthetic speech could be used as the control speech.

**Vary-control**: Fig. 2 presents the results of the study, where the number of control speakers \( K \) is varied from 13 to 1. It can be observed that the performance is pretty stable when \( K \) is reduced, even when selecting one single control speaker for intelligibility assessment, except for AF multi−manner. This indicates that, in the proposed approach, the number of control speakers can be reduced considerably. This observation is also supported by the single-synthetic-control study.

The proposed approach estimates an intelligibility score \( \text{IntScore} \), i.e. percentage of words correct for each speaker with dysarthria, which can be directly related to the subjective listening score, without any intermediary mapping or regression. Fig. 3 shows the Pearson’s correlation plot overlaid for the different systems, along with root mean square error (RMSE) between listener percentage word accuracy and the \( \text{IntScore} \).
REFERENCES

[1] J. R. Duffy, Motor Speech Disorders: Substrates, Differential Diagnosis, and Management. St. Louis, MO, USA: Elsevier Health Sciences, 2012.

[2] F. Darley, A. Aronson, and J. Brown, Motor Speech Disorders. Philadelphia, PA, USA: Saunders, 1975.

[3] S. Legendre, J. Liss, and A. Lotto, “Discriminating dysarthria type and predicting intelligibility from amplitude modulation spectra,” J. Acoust. Soc. Amer., vol. 125, p. 2530, May 2009.

[4] T. H. Falk, R. Hummel, and W. Chan, “Quantifying perturbations in temporal dynamics for automated assessment of spastic dysarthric speech intelligibility,” in Proc. Int. Conf. Acoust., Speech Signal Process., 2011, pp. 4480–4483.

[5] 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 Commun., vol. 54, no. 5, pp. 622–631, 2012.

[6] M. S. De Bodt, M. E. H.-D. Huici, and P. H. Van De Heyning, “Intelligibility as a linear combination of dimensions in dysarthric speech,” J. Commun. Disorders, vol. 35, no. 3, pp. 283–292, 2002.

[7] P. Janbakhshi, I. Kodrasi, and H. Bourlard, “Pathological speech intelligibility assessment based on the short-time objective intelligibility measure,” in Proc. Int. Conf. Acoust., Speech, Signal Process., 2019, pp. 6405–6409.

[8] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time-frequency weighted noisy speech,” IEEE Trans. Audio, Speech, Lang. Process., vol. 19, no. 7, pp. 2125–2136, Jul. 2011.

[9] J. Jensen and C. H. Taal, “An algorithm for predicting the intelligibility of speech masked by modulated noise masks,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 24, no. 11, pp. 2009–2022, Nov. 2016.

[10] D. Martinez, E. Lleida, P. Green, H. Christensen, A. Ortega, and A. Miguel, “Intelligibility assessment and speech recogniser word accuracy rate prediction for dysphonic speakers in a factor analysis subspace,” ACM Trans. Accessible Comput., vol. 6, no. 3, pp. 1–21, 2015.

[11] P. Janbakhshi, I. Kodrasi, and H. Bourlard, “Spectral subspace analysis for automatic assessment of pathological speech intelligibility,” in Proc. Interspeech, 2019, pp. 3038–3042.

[12] J. M. Kim, Y. Kim, and H. Kim, “Automatic intelligibility assessment of dysarthric speech using phonologically-structured sparse linear model,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 23, no. 4, pp. 694–704, Apr. 2015.

[13] C. Middag, J.-P. Martens, G. Van Nuffelen, and M. De Bodt, “Automated intelligibility assessment of pathological speech using phonological features,” EURASIP J. Adv. Signal Process., vol. 2009, no. 1, pp. 1–9, 2009.

[14] L. Ferrier, H. Shane, H. Ballard, T. Carpenter, and A. Benoit, “Dysarthric speakers’ intelligibility and speech characteristics in relation to computer speech recognition,” Augmentative Altern. Commun., vol. 11, no. 3, pp. 165–175, 1995.

[15] R. Ullmann, M. Magimai-Doss, and H. Bourlard, “Objective speech intelligibility assessment through comparison of phoneme class conditional probability sequences,” in Proc. Int. Conf. Acoust., Speech Signal Process., 2015, pp. 4924–4928.

[16] R. Rasipuram, M. Cernak, A. Nanchen, and M. Magimai-Doss, “Automatic accentedness evaluation of non-native speech using phonetic and sub-phonetic posterior probabilities,” in Proc. Interspeech, 2015, pp. 648–652.