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 (ρ) of 0.957. Furthermore, replacing control speakers' speech with speech synthesized by a neural text-to-speech system obtained a best r of 0.937 and ρ 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 been also 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 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.

This paper was submitted for review on October 22nd 2020. This work was funded through European Union's Horizon H2020 Marie Skłodowska-Curie Actions Innovative Training Network European Training Network (MSCA-ITN-ETN) project TAPAS under grant agreement No. 766287. J. Fritsch and M. Magimai.-Doss are with the Idiap Research Institute, 1920 Martigny, Switzerland. J. Fritsch is also with the École polytechnique fédérale de Lausanne (EPFL), Switzerland (e-mail: {julian.fritsch, mathew}@idiap.ch). The authors thank Bence Halpern from the NKI–AVL Amsterdam for helping us with synthetic reference generation.

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

Algorithm 1: Objective intelligibility score estimation

```
Set w=1, #CorrectWords=0;
while w ≤ W do
    Set k=1, vote count v_w=0;
    while k ≤ 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 ≥ K/2 then
        #CorrectWords = #CorrectWords + 1;
    end
end
Result: IntScore = (#CorrectWords / W) × 100%
```

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

A. Posterior-feature based matching of $\mathcal{Z}^w$ and $\mathcal{Y}_k^w$

The match between $\mathcal{Z}^w$ and $\mathcal{Y}_k^w$ is obtained by matching posterior feature vector sequences $(\mathbf{z}_1^w, \cdots \mathbf{z}_n^w, \cdots \mathbf{z}_N^w)$ and $(\mathbf{y}_1^w, \cdots \mathbf{y}_m^w, \cdots \mathbf{y}_{M_k}^w)$, 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 kth control speaker's utterance of word w, and posterior feature vectors $\mathbf{z}_n^w = [z_{n,1}^w, \cdots z_{n,d}^w, \cdots, z_{n,D}^w]^{\mathrm{T}}$ and $\mathbf{y}_m^w = [y_{m,1}^w, \cdots y_{m,d}^w, \cdots, y_{m,D}^w]^{\mathrm{T}}$ are D dimensional phones or broad phonetic classes posterior probabilities estimated using neural network(s) (see Section III-B), $\forall n \in \{1, \cdots N\}$ and $\forall m \in \{1, \cdots M_k\}$.

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

$$L^w(m,n) = l(\mathbf{y}_m^w, \mathbf{z}_n^w) + \min[L^w(m-1,n), L^w(m,n-1), L^w(m-1,n-1)] \ , \quad (1)$$

where, $l(\mathbf{y}_m^w, \mathbf{z}_n^w)$ is the local match score computed as symmetric Kullback-Leibler divergence between $\mathbf{y}_m^w$ and $\mathbf{z}_n^w$,

$$l(\mathbf{y}_m^w, \mathbf{z}_n^w) = \frac{1}{2} \cdot [\sum_{d=1}^{D} y_{m,d}^w \log \frac{y_{m,d}^w}{z_{n,d}^w} + \sum_{d=1}^{D} z_{n,d}^w \log \frac{z_{n,d}^w}{y_{m,d}^w}], \quad (2)$$

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 $\mathcal{Y}_k^w$ and $\mathcal{Z}^w$ 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 Thr_{inter} or at the center of the two means of the histogram, referred to as Thr_{cen}.

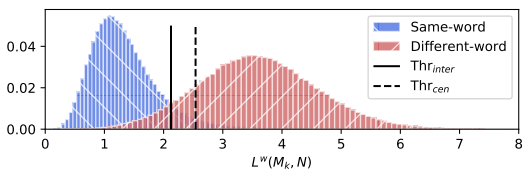

Fig. 1. Distribution of same and different-word pair scores $L^w(M_k, N)$

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 $\mathbf{z}_n^w$ and $\mathbf{y}_m^w$ 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:
(a) $\mathbf{AF}_{binary}$: 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) $\mathbf{AF}_{multi-manner}$: 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 Thr_{inter} and Thr_{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, ρ, 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 $IntScore$ 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 of threshold is not influencing the performance of the proposed approach. Furthermore, the proposed approach consistently performs comparably to or better than the baseline approaches.

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 $\text{AF}_{multi-manner}$ we obtain comparable r and ρ, 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 $\text{AF}_{multi-manner}$. 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 $IntScore$, i.e. percentage of words correct for each speaker

TABLE I
PEARSON'S CORRELATION (r) AND SPEARMAN'S CORRELATION (ρ) BETWEEN SUBJECTIVE AND OBJECTIVE INTELLIGIBILITY FOR **ALL-CONTROL** STUDY. p-VALUES ARE PRESENTED IN ITALICS FONT.

Posterior feature space	Thr_{cen}		Thr_{inter}	
	r	ρ	r	ρ
Phone	.939	.939	**.950**	**.957**
	3.94e-7	*2.31e-7*	*5.52e-8*	*2.29e-8*
$\text{AF}_{\text{binary}}$	.918	.885	.922	.885
	1.88e-6	*1.13e-5*	*1.27e-6*	*1.32e-5*
$\text{AF}_{\text{multi-manner}}$	.922	.910	.917	.894
	1.01e-6	*2.42e-6*	*1.43e-6*	*6.82e-6*
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
PEARSON'S CORRELATION (r) AND SPEARMAN'S CORRELATION (ρ) BETWEEN SUBJECTIVE AND OBJECTIVE INTELLIGIBILITY FOR **SINGLE-SYNTHETIC-CONTROL** STUDY. p-VALUES ARE PRESENTED IN ITALICS FONT.

Posterior feature space	Thr_{cen}		Thr_{inter}	
	r	ρ	r	ρ
Phone	.924	.942	.931	**.961**
	1.44e-7	*8.08e-7*	*1.14e-8*	*4.46e-7*
$\text{AF}_{\text{binary}}$	0.827	.893	.822	.885
	1.40e-4	*7.23e-6*	*1.68e-4*	*1.13e-5*
$\text{AF}_{\text{multi-manner}}$	**.937**	.906	.930	.912
	2.40e-7	*3.09e-6*	*4.78e-7*	*2.13e-6*

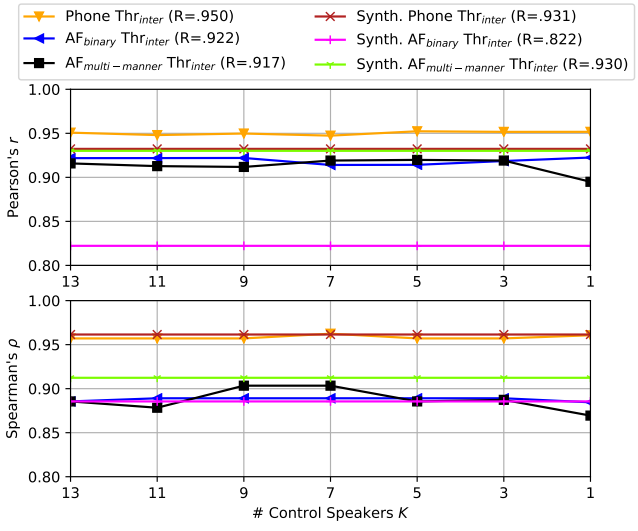

Fig. 2. Pearson's correlation and Spearman's correlation when the number of control speakers K is varied from 13 to 1. Synth. refers to the case of single-synthetic-control.

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 $IntScore$ (presented in the legends); each marker represents one speaker. It can be observed that phone space and $\text{AF}_{multi-manner}$ space are predicting well high intelligibility regions, while AF_{binary} is predicting comparatively well the low intelligibility regions. As a consequence, although AF_{binary} is not the best in terms of r and ρ, it yields the best RMSE of 16.9%. We observe this trend even in the case of synthetic control speech, denoted as Synth AF_{binary}. This is promising as we have not used any dysarthric speech data to build any part of the assessment system. In the previous studies, on the same data set, RMSE ranging from 12% to 18.6% have been reported with the use of dysarthric speech data to build the intelligibility prediction models [5], [10]. Overall, the analysis indicates that $IntScore$ estimation needs to be further improved for low intelligibility regions to take advantage of its interpretability. This is a part of our on-going work.

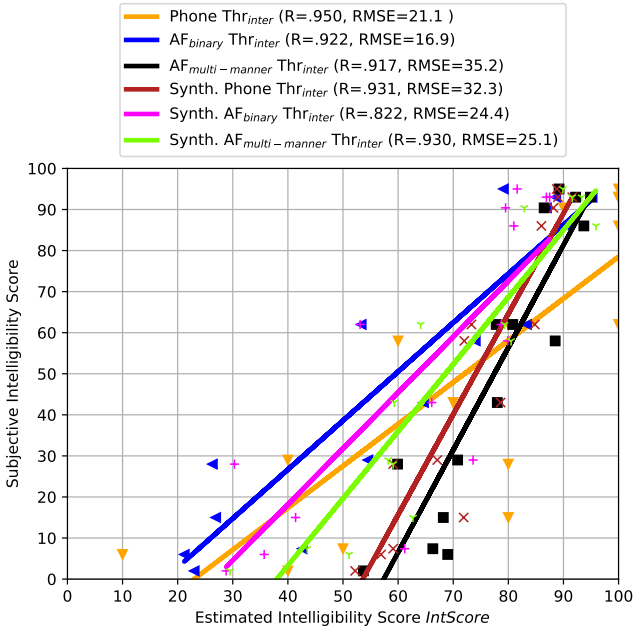

Fig. 3. Pearson's correlation plot obtained from proposed intelligibility assessment systems. Synth refers to the case of single-synthetic-control.

V. CONCLUSIONS

We proposed an approach to assess dysarthric speech intelligibility by matching and verifying the utterances of speakers with dysarthria of a set of words against a set of control speakers' utterances of those words in phone or broad phonetic posterior feature spaces. Our investigations on the UA-Speech corpus using posterior feature estimators trained on auxiliary data and language showed that the proposed approach obtains high correlation with subjective intelligibility scores for both phone and broad-phonetic posterior feature spaces. Our investigations also demonstrated that the proposed approach obtains high correlation even when the control speakers' speech is replaced by speech synthesized by a neural TTS system or the number of control speakers is considerably reduced. Our future work will focus on extending the proposed approach in the framework of KL-HMM [32] to better explain the variations in dysarthric speech in phone and AF spaces.

REFERENCES

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

[2] F. Darley, A. Aronson, and J. Brown, *Motor Speech Disorders*. Saunders, 1975.

[3] S. Legendre, J. Liss, and A. Lotto, "Discriminating dysarthria type and predicting intelligibility from amplitude modulation spectra," *The Journal of the Acoustical Society of America*, vol. 125, p. 2530, 05 2009.

[4] T. H. Falk, R. Hummel, and W. Chan, "Quantifying perturbations in temporal dynamics for automated assessment of spastic dysarthric speech intelligibility," in *Proceedings of ICASSP*, 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 Communication*, 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," *Journal of communication 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 *Proceedings of ICASSP*. IEEE, 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 Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[9] J. Jensen and C. H. Taal, "An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[10] D. Martínez, E. Lleida, P. Green, H. Christensen, A. Ortega, and A. Miguel, "Intelligibility assessment and speech recognizer word accuracy rate prediction for dysarthric speakers in a factor analysis subspace," *ACM Transactions on Accessible Computing*, 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 *Proceedings of Interspeech*, 2019.

[12] M. J. Kim, Y. Kim, and H. Kim, "Automatic Intelligibility Assessment of Dysarthric Speech Using Phonologically-Structured Sparse Linear Model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 694–704, 2015.

[13] C. Middag, J.-P. Martens, G. Van Nuffelen, and M. De Bodt, "Automated intelligibility assessment of pathological speech using phonological features," *EURASIP Journal on Advances in Signal Processing*, 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 and Alternative Communication*, 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 *Proceedings of ICASSP*, 2015.

[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 *Proceedings of Interspeech*, 2015.

[17] R. Kent, G. Weismer, J. Kent, and J. Rosenbek, "Toward Phonetic Intelligibility Testing in Dysarthria," *The Journal of speech and hearing disorders*, vol. 54, pp. 482–99, 12 1989.

[18] ASHA, "Dysarthria in adults," *https://www.asha.org/practice-portal/clinical-topics/dysarthria-in-adults/*.

[19] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-26, no. 1, pp. 43–49, February 1978.

[20] R. Ullmann, R. Rasipuram, M. Magimai.-Doss, and H. Bourlard, "Objective Intelligibility Assessment of Text-to-Speech Systems Through Utterance Verification," in *Proceedings of Interspeech*, 2015.

[21] T. Kailath, "The Divergence and Bhattacharyya Distance Measures in Signal Selection," *IEEE Transactions on Communication*, vol. 15, no. 1, pp. 52–60, 1967.

[22] R. E. Blahut, "Hypothesis Testing and Information Theory," *IEEE Trans. on Information Theory*, vol. IT-20, no. 4, pp. 405–417, 1974.

[23] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[24] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot. Int.*, vol. 5, no. 9, pp. 341–345, 2001.

[25] S. Soldo, M. Magimai.-Doss, J. Pinto, and H. Bourlard, "Posterior Features for Template-based ASR," in *Proceedings of ICASSP*, 2011, pp. 4864–4867.

[26] N. Chomsky and M. Halle, *The Sound Patterns in English*. MIT Press, 1968.

[27] P. Ladefoged, *A Course in Phonetics*. Harcourt Brace College Publishers, 1993.

[28] J. Vásquez-Correa, P. Klumpp, J. R. Orozco-Arroyave, and E. Nöth, "Phonet: a tool based on gated recurrent neural networks to extract phonological posteriors from speech," *Proceedings of Interspeech*, pp. 549–553, 2019.

[29] J. Fritsch, S. P. Dubagunta, and M. Magimai.-Doss, "Estimating the degree of sleepiness by integrating articulatory feature knowledge in raw waveform based CNNs," in *Proceedings of ICASSP*, 2020, pp. 6534–6538.

[30] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," in *Proceedings of ICASSP*, 2018, pp. 4779–4783.

[31] M. S. Paja and T. H. Falk, "Automated dysarthria severity classification for improved objective intelligibility assessment of spastic dysarthric speech," in *Proceedings of Interspeech*, 2012.

[32] G. Aradilla, J. Vepa, and H. Bourlard, "An Acoustic Model Based on Kullback-Leibler Divergence for Posterior Features," in *Proceedings of ICASSP*, 2007, pp. 657–660.

[33] D. Johnson, D. Ellis, C. Oei, C. Wooters, and P. Faerber, "Quicknet," *www1.icsi.berkeley.edu/Speech/qn.html*, 2004.

[34] J. Carletta *et al.*, "The AMI meeting corpus: A pre-announcement," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2005, pp. 28–39.

[35] R. Rasipuram and M. Magimai.-Doss, "Articulatory feature based continuous speech recognition using probabilistic lexical modeling," *Computer Speech & Language*, vol. 36, pp. 233–259, 2016.

VI. Supplementary Material

A. Phonetic Posterior Feature Representations

In this section, we provide further details about the different posterior feature estimators

Phone space consists of 45 dimensional context-independent phoneme posterior probabilities estimated by an off-the-shelf multilayer perceptron (MLP). The MLP takes as input 39-dimensional perceptual linear predictive cepstral features with (frame size is 25 ms, frame shift 10 ms) a nine frame temporal context (i.e. four frames preceding and four frames following). The MLP had a single hidden layer with 5000 units. The output layer consisted of 44 English phonemes (based on UniSyn dictionary) and silence, i.e., $D = 45$. The MLP has been trained on 232 hours of conversational telephone speech with the QuickNet tool [33] by minimizing the frame-level cross entropy. This MLP was originally trained for template-based speech recognition [25], and has been later used in the speech intelligibility prediction studies reported in [15], [20].

$\mathbf{AF}_{binary}$ space consists of 18 binary valued AFs, namely, $\{$`pause`, `consonantal`, `back`, `anterior`, `open`, `close`, `nasal`, `stop`, `continuant`, `lateral`, `flap`, `trill`, `voice`, `strident`, `labial`, `dental`, `velar`, `vocalic`$\}$. In the Phonet toolkit[1] [28], these AFs are modeled by 18 off-the-shelf recurrent neural network (RNN) based binary classifiers, i.e. $D = 18 \times 2$. The RNNs takes as input log-energies of 33-dimensional Mel filterbank energies. The RNN classifiers have been trained on 17 hours of clean FM podcasts Mexican Spanish with a cost function based on cross entropy. For more details, related to the mapping between Spanish phones and the AFs and training of RNNs, the reader is referred to [28].

$\mathbf{AF}_{multi-manner}$ space consists of nine "manner of articulation" category AFs, namely, $\{$`silence`, `aspirated`, `approximant`, `fricative`, `nasal`, `voiced-fricative`, `voiced-stop`, `stop`, `vowel`$\}$. These AFs were modeled by an off-the-shelf convolution neural network (CNN) that takes raw waveform as input and predicts the posterior probabilities of the nine manner of articulation category AFs, i.e. $D = 9$. The CNN has been trained on the 77 hour AMI corpus [34] with a cost function based on cross entropy. The mapping between the English phones and the AFs was based on a previous work on automatic speech recognition [35]. For further details about the architecture and training of the CNN, the reader is referred to [29].

B. Synthetic references from Tacotron2

We used an off-the-shelf neural TTS system Tacotron2 [30] to obtain synthetic references. The synthesizer has been originally trained on the LJSpeech corpus[2], which is an annotated English corpus including 13,100 short audio clips of a single speaker reading passages from 7 non-fiction books. The system has been rated with a mean opinion score of 4.526 ± 0.066 on scale of 1 to 5. During synthesis, each word from the UA-Speech word-list was converted into a phoneme sequence based on CMUDict[3]. For more information about the TTS system, the reader is referred to [30].

[1] https://github.com/jcvasquezc/phonet

[2] https://keithito.com/LJ-Speech-Dataset/

[3] http://www.speech.cs.cmu.edu/cgi-bin/cmudict