Speech Intelligibility Assessment of Dysarthric Speech by using Goodness of Pronunciation with Uncertainty Quantification

Eun Jung Yeo*, Kwanghee Choi*, Sunhee Kim, Minhwa Chung

Interspeech 2023

Contents

Motivation

What is dysarthric speech?

- Dysarthria: A group of motor speech disorders resulting from neuromuscular control disturbances.
- People with dysarthria suffer from degraded speech intelligibility.
- Accurate and reliable speech assessment is essential in the clinical field.

Two approaches of automatic dysarthric speech assessment

- Hand-crafted features provide interpretability with medical implications.
- Neural network-based approaches achieves better performances.

Can we enjoy both interpretability & better performance?

Motivation: Why use Goodness-of-Pronunciation?

What is Goodness-of-Pronunciation (GoP)?

- Degree of similarity between produced and correct pronunciation of phonemes.
- Often used in non-native (L2) speech pronunciation assessment

Advantages of GoP

- Interpretability by showing which phonemes are mispronounced and to what extent each phoneme is atypical.
- Do not need parallel datasets for training and test.
  → Will be covered in the following slides!

Motivation: Why use Goodness-of-Pronunciation?

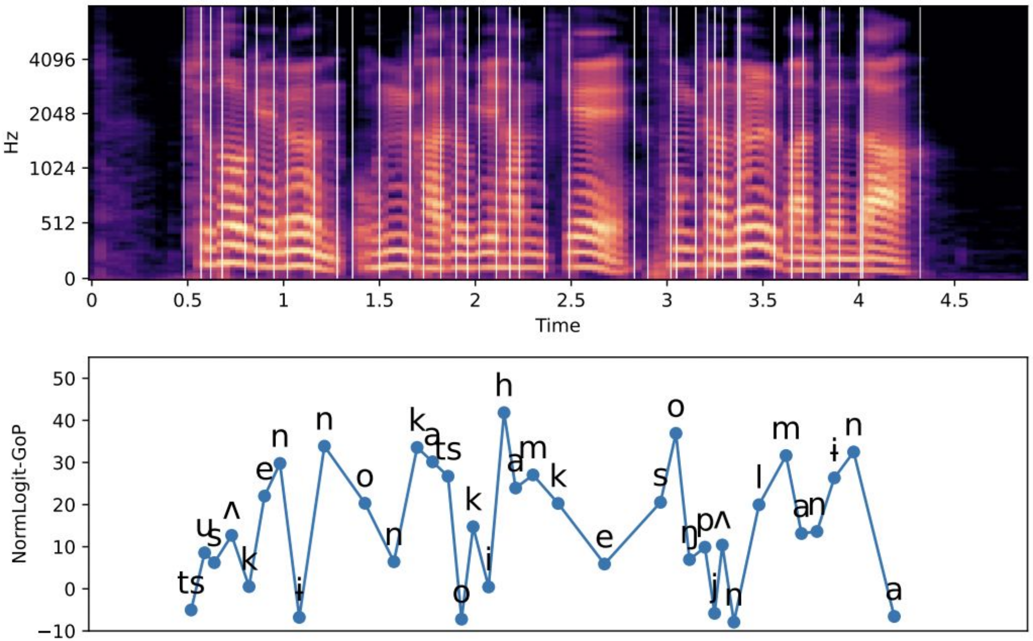
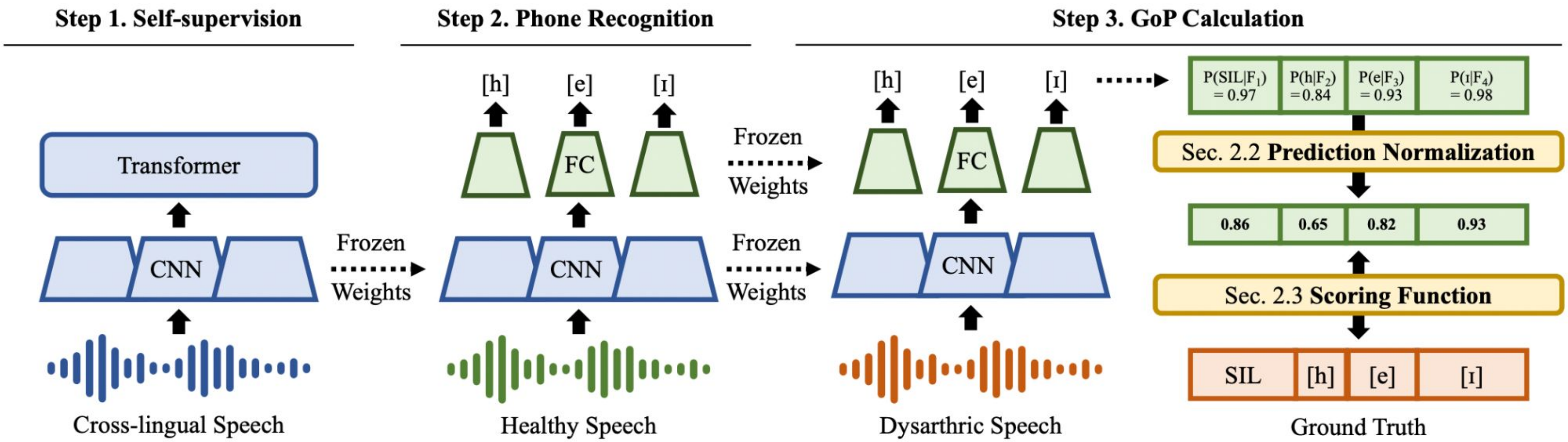


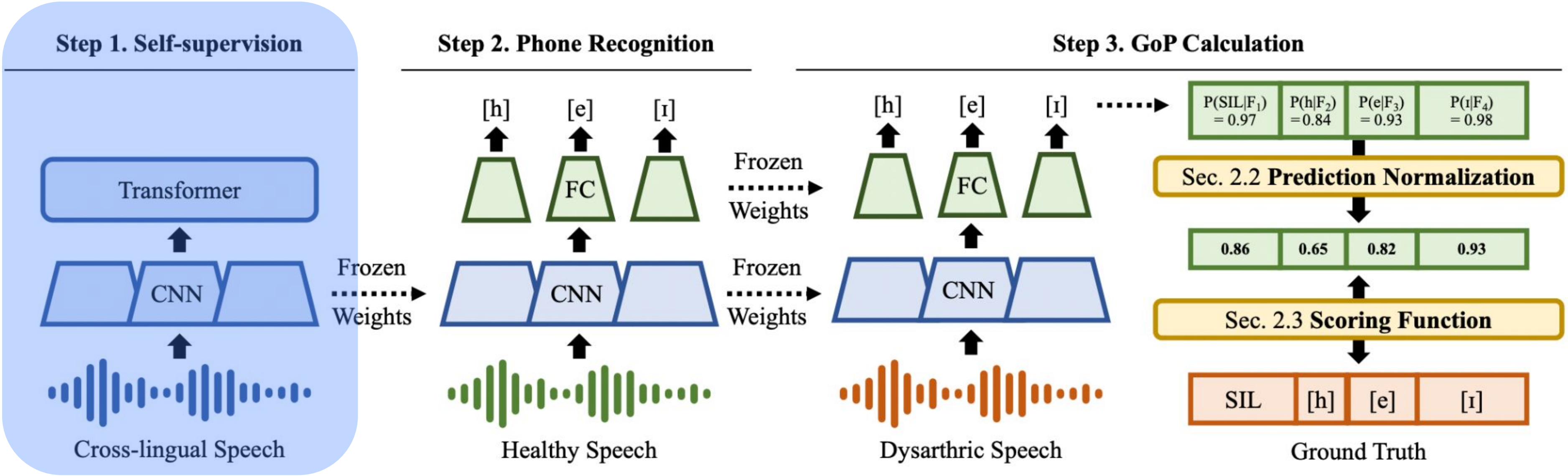
Figure 1: *Example of GoP scores for each phone within an utterance. Higher values indicate greater deviation from the correct pronunciation. GoP scores allow easy identification of mispronounced phonemes.*

Our method



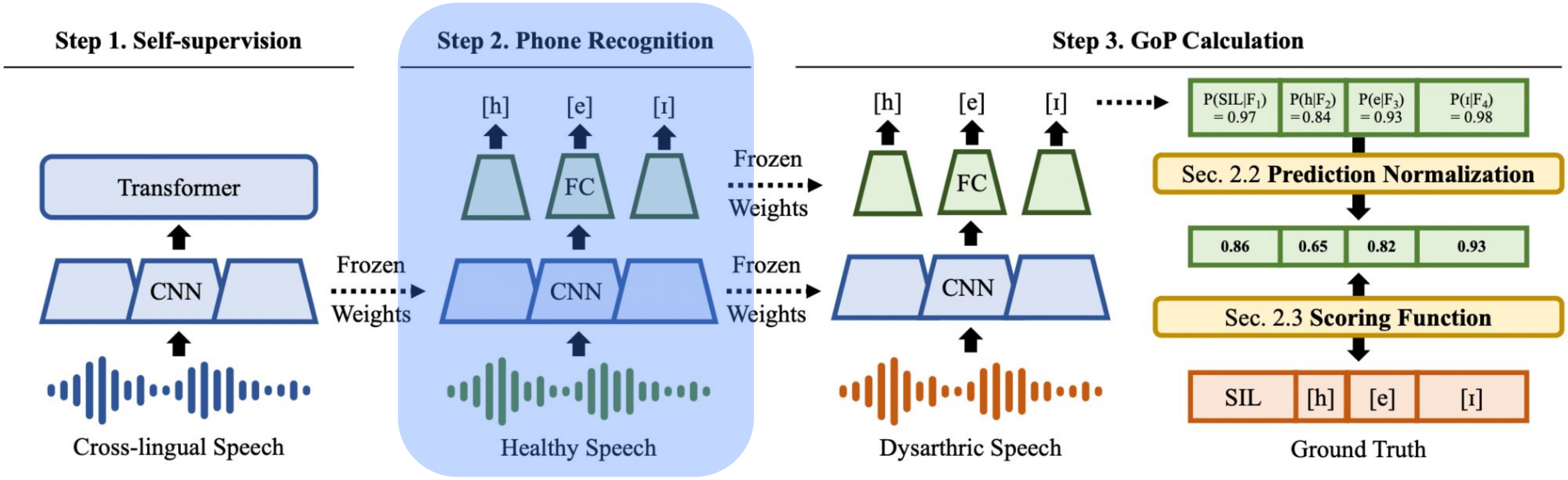
Step 1. Self-supervision
Step 2. Phone Recognition
Step 3. GoP Calculation
Transformer
CNN
Frozen
Weights
Cross-lingual Speech
[h]
[e]
[ɪ]
FC
CNN
Frozen
Weights
Frozen
Weights
Healthy Speech
[h]
[e]
[ɪ]
FC
CNN
Dysarthric Speech
$P(SIL|F_1) = 0.97$
$P(h|F_2) = 0.84$
$P(e|F_3) = 0.93$
$P(ɪ|F_4) = 0.98$
Sec. 2.2 **Prediction Normalization**
0.86
0.65
0.82
0.93
Sec. 2.3 **Scoring Function**
SIL
[h]
[e]
[ɪ]
Ground Truth

Our method



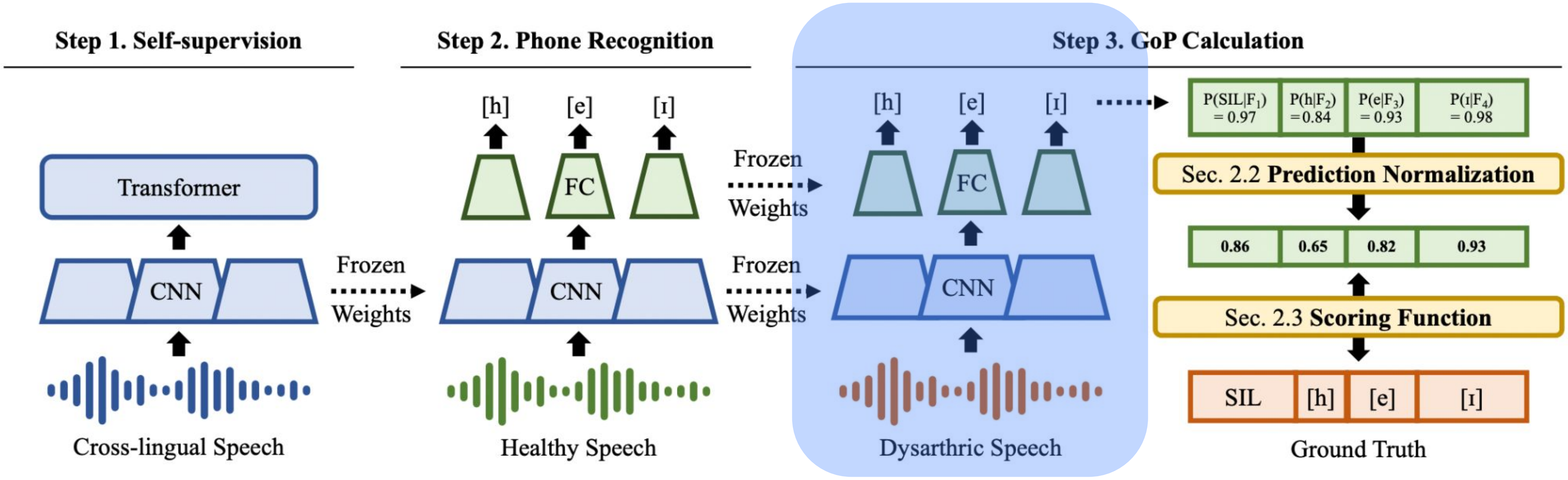
Pre-trained
self-supervised model [1]

Our method



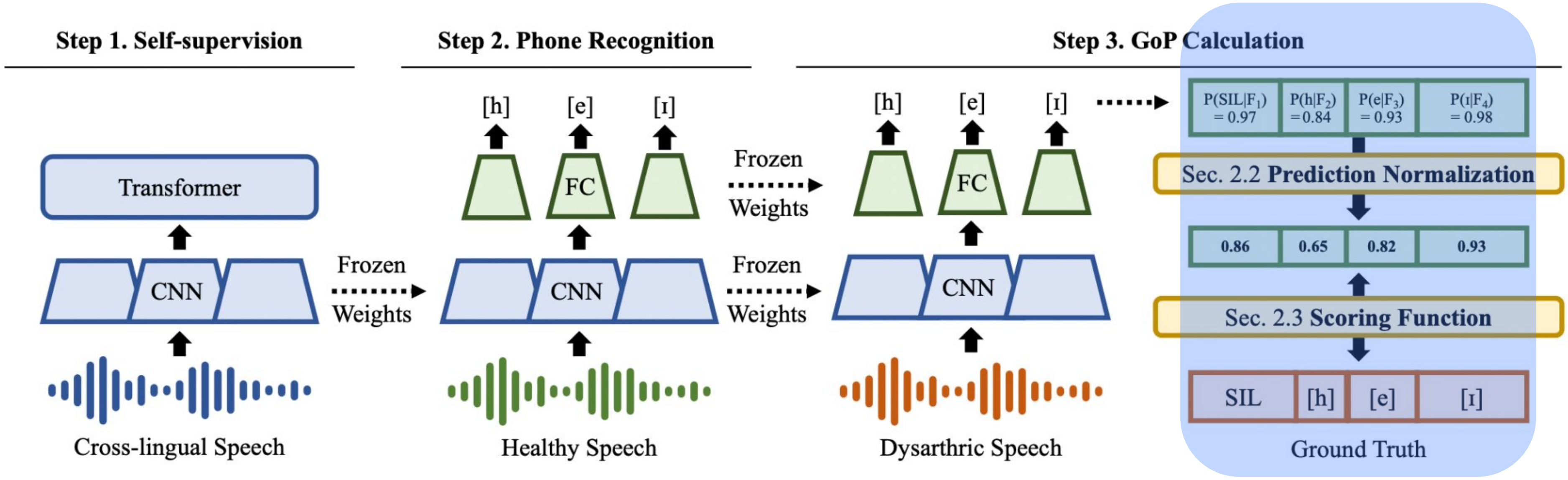
Train a phoneme recognition model using **healthy speech**

Our method



Infer the trained model on
dysarthric speech dataset

Our method



Calculate GoP scores
by comparing
phone predictions with
ground truth phones

Motivation: Why use Uncertainty Quantification?

What is Uncertainty Quantification (UQ)?

- Modern NNs are reported to be overconfident; often generate probabilities close to 1.0 even when their predictions are incorrect [2].
- NN suffers especially when the model encounters out-of-distribution (OOD) inputs; data that differ significantly from the training data.

→ UQ techniques directly combat the above problem.

We use UQ on phoneme probabilities to improve the GoP scores!

- Dysarthric speech is OOD for the phoneme predictor, as it is trained with healthy speech only.

Our method: Apply conventional UQ approaches

Baselines

- **[GMM-GoP]** [3] directly uses the phoneme predictions, i.e. $P(p|x)$.
- **[NN-GoP]** [4] and **[DNN-GoP]** [5] leveraged the development of NNs + UQ

UQ #1. Normalizing the predictions

- **[Scale]** reduces the peakiness by temperature scaling.
- **[Prior]** removes the influence of prior phoneme distribution.

UQ #2. Modifying the scoring function

- **[Entropy]** or **[Margin]** also leverages the prediction of other phonemes.
- **[MaxLogit]** and **[LogitMargin]** utilizes the pre-softmax logits.

Datasets

Train the phone predictor using:

- **[Common Phone]** is a gender-balanced, multilingual corpus.
- **[L2-ARCTIC]** contains non-native English speakers for L2 research.

Evaluate our method using:

- **[UASpeech]** is an English dataset with 15 patients & 13 healthy speakers.
- **[QoLT]** is a Korean dataset with 70 patients & 10 healthy speakers.
- **[SSNCE]** is a Tamil dataset with 20 patients & 10 healthy speakers.

All the datasets are publicly available / available upon request!

Results: Comparing UQ methods

- We measure correlations between GoP scores and intelligibility scores.
- Best score: MaxLogit

Table 1: *Kendall's rank coefficient between GoP & intelligibility severity levels. A higher absolute value indicates a stronger correlation between the two variables.*

Method	Norm.	Scoring Func.	English	Korean	Tamil
Baseline	None	GMM [18, 39]	-0.2049	-0.5237	-0.3571
	None	NN [21]	-0.1536	-0.4687	-0.4003
	Prior	DNN-GoP [21]	-0.1836	-0.4237	-0.4681
Proposed	None	Entropy	-0.1831	-0.2643	-0.3251
		Margin	-0.1628	-0.4434	-0.4445
		MaxLogit	-0.2164	-0.5440	-0.5786
		LogitMargin	-0.1732	-0.4753	-0.5158
	Scale	Entropy	-0.1755	-0.1974	-0.2263
		Margin	-0.1260	-0.4444	-0.4210
		MaxLogit	-0.2164	-0.5440	-0.5786
		LogitMargin	-0.1732	-0.4753	-0.5158
	Prior	Entropy	-0.1833	-0.2645	-0.3254
		Margin	-0.1630	-0.4432	-0.4447
		MaxLogit	**-0.2165**	**-0.5442**	**-0.5788**
		LogitMargin	-0.1733	-0.4753	-0.5160

Results: Comparing UQ methods

- Experiments without normalization **[None]** generally showed lower performance than the baseline, except for MaxLogit-based GoP.

Table 1: *Kendall's rank coefficient between GoP & intelligibility severity levels. A higher absolute value indicates a stronger correlation between the two variables.*

Method	Norm.	Scoring Func.	English	Korean	Tamil
Baseline	None	GMM [18, 39]	-0.2049	-0.5237	-0.3571
	None	NN [21]	-0.1536	-0.4687	-0.4003
	Prior	DNN-GoP [21]	-0.1836	-0.4237	-0.4681
Proposed	None	Entropy	-0.1831	-0.2643	-0.3251
		Margin	-0.1628	-0.4434	-0.4445
		MaxLogit	-0.2164	-0.5440	-0.5786
		LogitMargin	-0.1732	-0.4753	-0.5158
	Scale	Entropy	-0.1755	-0.1974	-0.2263
		Margin	-0.1260	-0.4444	-0.4210
		MaxLogit	-0.2164	-0.5440	-0.5786
		LogitMargin	-0.1732	-0.4753	-0.5158
	Prior	Entropy	-0.1833	-0.2645	-0.3254
		Margin	-0.1630	-0.4432	-0.4447
		MaxLogit	**-0.2165**	**-0.5442**	**-0.5788**
		LogitMargin	-0.1733	-0.4753	-0.5160

Results: Comparing UQ methods

- Experiments with scaling normalization **[scale]**, has minimal impact on improvements.

Table 1: *Kendall's rank coefficient between GoP & intelligibility severity levels. A higher absolute value indicates a stronger correlation between the two variables.*

Method	Norm.	Scoring Func.	English	Korean	Tamil
Baseline	None	GMM [18, 39]	-0.2049	-0.5237	-0.3571
	None	NN [21]	-0.1536	-0.4687	-0.4003
	Prior	DNN-GoP [21]	-0.1836	-0.4237	-0.4681
Proposed	None	Entropy	-0.1831	-0.2643	-0.3251
		Margin	-0.1628	-0.4434	-0.4445
		MaxLogit	-0.2164	-0.5440	-0.5786
		LogitMargin	-0.1732	-0.4753	-0.5158
	Scale	Entropy	-0.1755	-0.1974	-0.2263
		Margin	-0.1260	-0.4444	-0.4210
		MaxLogit	-0.2164	-0.5440	-0.5786
		LogitMargin	-0.1732	-0.4753	-0.5158
	Prior	Entropy	-0.1833	-0.2645	-0.3254
		Margin	-0.1630	-0.4432	-0.4447
		MaxLogit	**-0.2165**	**-0.5442**	**-0.5788**
		LogitMargin	-0.1733	-0.4753	-0.5160

Results: Comparing UQ methods

- Experiments with prior normalization **[prior]**, has minimal impact on improvements.

Table 1: *Kendall's rank coefficient between GoP & intelligibility severity levels. A higher absolute value indicates a stronger correlation between the two variables.*

Method	Norm.	Scoring Func.	English	Korean	Tamil
Baseline	None	GMM [18, 39]	-0.2049	-0.5237	-0.3571
	None	NN [21]	-0.1536	-0.4687	-0.4003
	Prior	DNN-GoP [21]	-0.1836	-0.4237	-0.4681
Proposed	None	Entropy	-0.1831	-0.2643	-0.3251
		Margin	-0.1628	-0.4434	-0.4445
		MaxLogit	-0.2164	-0.5440	-0.5786
		LogitMargin	-0.1732	-0.4753	-0.5158
	Scale	Entropy	-0.1755	-0.1974	-0.2263
		Margin	-0.1260	-0.4444	-0.4210
		MaxLogit	-0.2164	-0.5440	-0.5786
		LogitMargin	-0.1732	-0.4753	-0.5158
	Prior	Entropy	-0.1833	-0.2645	-0.3254
		Margin	-0.1630	-0.4432	-0.4447
		MaxLogit	**-0.2165**	**-0.5442**	**-0.5788**
		LogitMargin	-0.1733	-0.4753	-0.5160

Results: Comparing UQ methods

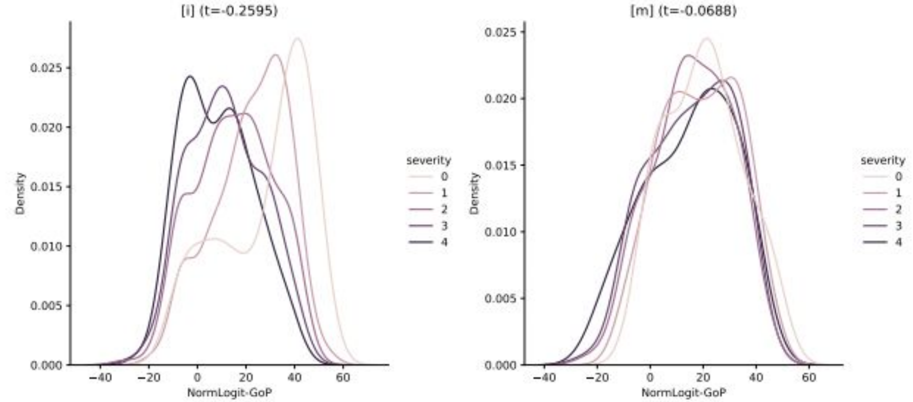
- **[Entropy]** generally showed the lowest performance.
- **[MaxLogit]** and **[LogitMargin]** showed the highest correlation to the intelligibility scores.

Table 1: *Kendall's rank coefficient between GoP & intelligibility severity levels. A higher absolute value indicates a stronger correlation between the two variables.*

Method	Norm.	Scoring Func.	English	Korean	Tamil
Baseline	None	GMM [18, 39]	-0.2049	-0.5237	-0.3571
	None	NN [21]	-0.1536	-0.4687	-0.4003
	Prior	DNN-GoP [21]	-0.1836	-0.4237	-0.4681
Proposed	None	Entropy	-0.1831	-0.2643	-0.3251
		Margin	-0.1628	-0.4434	-0.4445
		MaxLogit	-0.2164	-0.5440	-0.5786
		LogitMargin	-0.1732	-0.4753	-0.5158
	Scale	Entropy	-0.1755	-0.1974	-0.2263
		Margin	-0.1260	-0.4444	-0.4210
		MaxLogit	-0.2164	-0.5440	-0.5786
		LogitMargin	-0.1732	-0.4753	-0.5158
	Prior	Entropy	-0.1833	-0.2645	-0.3254
		Margin	-0.1630	-0.4432	-0.4447
		MaxLogit	**-0.2165**	**-0.5442**	**-0.5788**
		LogitMargin	-0.1733	-0.4753	-0.5160

Analyses on phonemes

- **Certain phones have more impact** on intelligibility scores [5].
- While the distribution of /i/ differs significantly, the distribution of /m/ is similar across all severity.



(a) *Korean phoneme /i/.* (b) *Korean phoneme /m/.*

Figure 3: *Kendall's τ distributions for two phonemes by severity. 0:healthy, 1:mild, 2:mild-to-mod., 3:mod.-to-sev., 4:severe.*

Analyses on phonemes

- **Fricatives, Affricates and diphthongs for English and Tamil:** possibly due to their complexity leading to difficulties.
- **Fricative, Nasal and monophthongs for Korean:** possibly related to the movement of the articulators.

	Top-5 Phonemes
English	/aɪ/, /ʃ/, /aʊ/, /z/, /ʤ/
Korean	/i/, /s/, /n/, /a/, /ʌ/
Tamil	/ʂ/, /h/, /ʧ/, /z/, /aɪ/

Takeaways

1. **Improved GoP for dysarthric speech intelligibility assessment**
   → Dysarthric speech is very different from healthy speech!
   → Uncertainty Quantification (UQ) techniques

2. **UQ Techniques**
   → Normalization of phoneme prediction
   → Modification of the scoring function

3. **Usefulness of GoP**
   → *Quantify* how distinct each phoneme is from healthy phonemes.
   → *Quantify* the impact of each phoneme on speech intelligibility.

Selected References

[1] Xu, X., Kang, Y., Cao, S., Lin, B., & Ma, L. (2021, August). Explore wav2vec 2.0 for Mispronunciation Detection. *In Interspeech* (pp. 4428-4432).

[2] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017, July). On calibration of modern neural networks. *In International conference on machine learning* (pp. 1321-1330).

[3] Witt, S. M., & Young, S. J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech communication*, 30(2-3), 95-108.

[4] Hu, W., Qian, Y., Soong, F. K., & Wang, Y. (2015). Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers. *Speech Communication*, 67, 154-166.

[5] Quintas, S., Mauclair, J., Woisard, V., & Pinquier, J. (2022, September). Automatic assessment of speech intelligibility using consonant similarity for head and neck cancer. *In Interspeech*.

Thank you for your attention!

Q & A

- Source code: https://github.com/juice500ml/dysarthria-gop