Synthetic Speech References for Automatic Pathological Speech Intelligibility Assessment

Parvaneh Janbakhshi, Ina Kodrasi, and Hervé Bourlard

Idiap Research Institute
Speech and Audio Processing Group

April 2020

Outline

1. Pathological Speech Intelligibility Assessment
2. State-of-the-art Automatic Intelligibility Measures
3. P-ESTOI with Synthetic Speech References
4. Experimental Results
5. Summary

Pathological speech intelligibility assessment

- Disrupted speech production mechanism due to speech disorders, e.g., Cerebral Palsy (CP)
  - Reduced intelligibility and communicative ability

Pathological speech intelligibility assessment

- Disrupted speech production mechanism due to speech disorders, e.g., Cerebral Palsy (CP)
  - Reduced intelligibility and communicative ability
- Speech intelligibility (degree of understandability of speech)
  - A clinical auditory-perceptual evaluation of pathological speech

Pathological speech intelligibility assessment

- Disrupted speech production mechanism due to speech disorders, e.g., Cerebral Palsy (CP)
  - Reduced intelligibility and communicative ability
- Speech intelligibility (degree of understandability of speech)
  - A clinical auditory-perceptual evaluation of pathological speech

- Subjective listening tests based on judgement of human listeners
  - Labor-intensive
  - Subject to the listener bias and to contextual and linguistic cues

- Automatic intelligibility measures
  - Frequent, economical, and objective assessment
  - Applicable in remote speech therapies

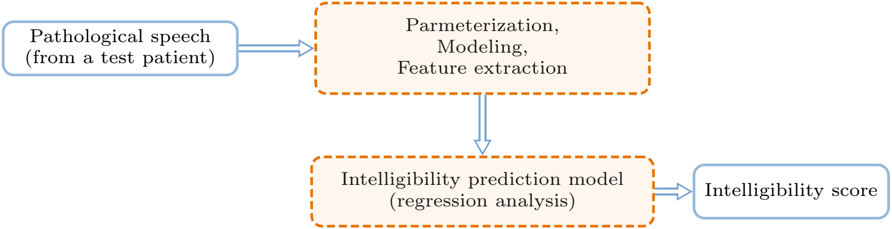
Outline

1. Pathological Speech Intelligibility Assessment
2. State-of-the-art Automatic Intelligibility Measures
3. P-ESTOI with Synthetic Speech References
4. Experimental Results
5. Summary

Automatic pathological intelligibility measures (State-of-the-art)

» Reference-free approaches

» Reference-based approaches



Pathological speech
(from a test patient)

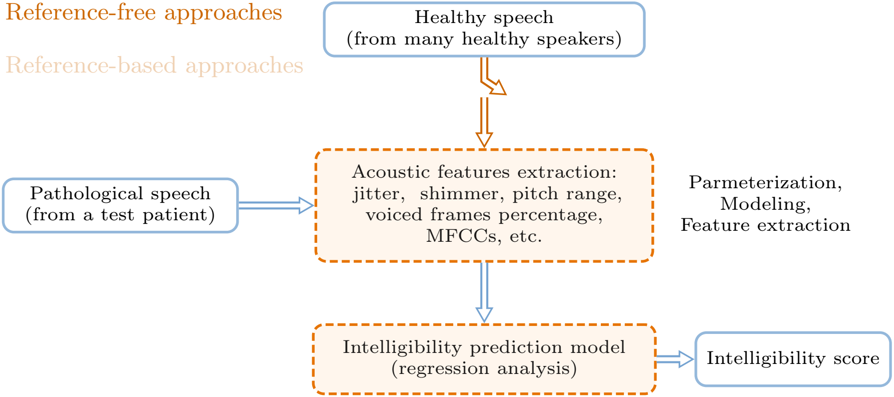
Parmeterization,
Modeling,
Feature extraction

Intelligibility prediction model
(regression analysis)

Intelligibility score

Automatic pathological intelligibility measures (State-of-the-art)

- » Reference-free approaches
- » Reference-based approaches

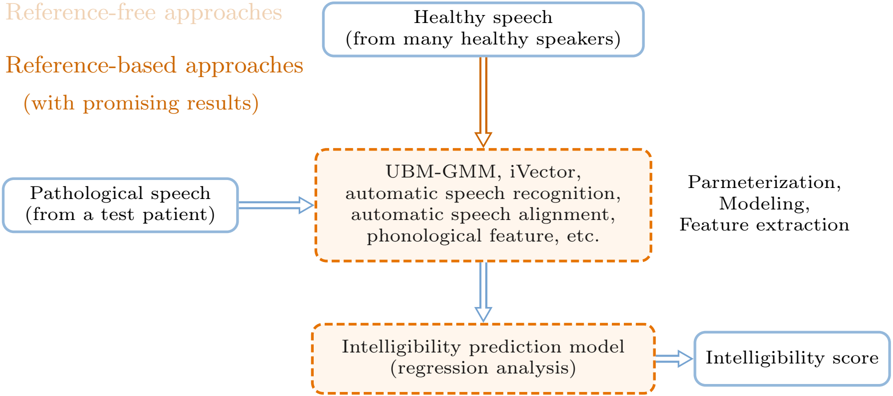


Automatic pathological intelligibility measures (State-of-the-art)

» Reference-free approaches

» Reference-based approaches

(with promising results)



Healthy speech
(from many healthy speakers)

Pathological speech
(from a test patient)

UBM-GMM, iVector,
automatic speech recognition,
automatic speech alignment,
phonological feature, etc.

Parmeterization,
Modeling,
Feature extraction

Intelligibility prediction model
(regression analysis)

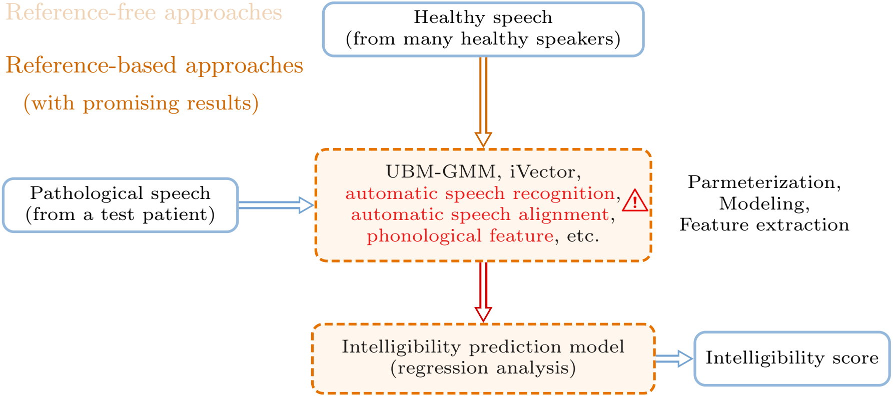
Intelligibility score

Automatic pathological intelligibility measures (State-of-the-art)

» Reference-free approaches

» Reference-based approaches

(with promising results)



Healthy speech
(from many healthy speakers)

Pathological speech
(from a test patient)

UBM-GMM, iVector,
automatic speech recognition,
automatic speech alignment,
phonological feature, etc. ⚠

Parmeterization,
Modeling,
Feature extraction

Intelligibility prediction model
(regression analysis)

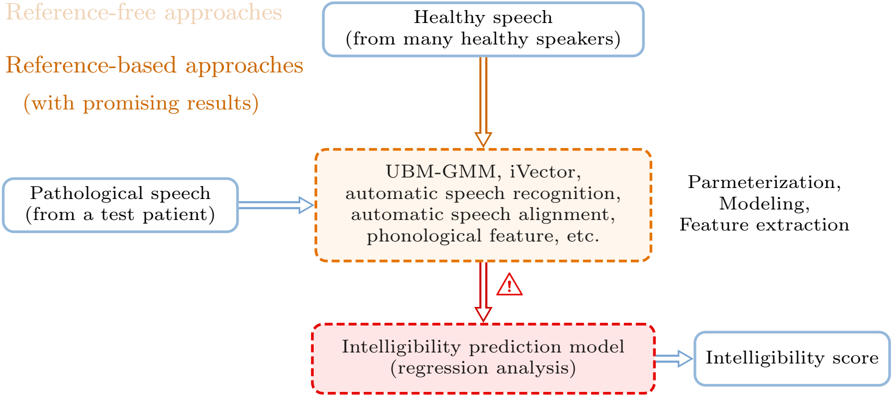
Intelligibility score

⚠ Unpredictability for severe patients

Automatic pathological intelligibility measures (State-of-the-art)

- » Reference-free approaches
- » Reference-based approaches

  (with promising results)



Healthy speech
(from many healthy speakers)

Pathological speech
(from a test patient)

UBM-GMM, iVector,
automatic speech recognition,
automatic speech alignment,
phonological feature, etc.

Parmeterization,
Modeling,
Feature extraction

⚠

Intelligibility prediction model
(regression analysis)

Intelligibility score

⚠ Large number of features
(increasing the risk of over-fitting)

Our previously proposed referenced-based intelligibility measures (State-of-the-art)

- Pathological short-time objective intelligibility (P-ESTOI) measure (Janbakhshi et al., 2019a)
  - Simple structure
  - Based on a single feature without training (no risk of overfitting)
  - Generalizable across languages and neurological diseases

Our previously proposed referenced-based intelligibility measures (State-of-the-art)

- Pathological short-time objective intelligibility (P-ESTOI) measure (Janbakhshi et al., 2019a)
  - Simple structure
  - Based on a single feature without training (no risk of overfitting)
  - Generalizable across languages and neurological diseases
  - Requires healthy recordings matching the phonetic content of the pathological speech to create a reference model (phonetically-balanced scenarios)
    - Not applicable to phonetically-unbalanced scenarios

Our previously proposed referenced-based intelligibility measures (State-of-the-art)

- Pathological short-time objective intelligibility (P-ESTOI) measure (Janbakhshi et al., 2019a)
  - Simple structure
  - Based on a single feature without training (no risk of overfitting)
  - Generalizable across languages and neurological diseases
  - Requires healthy recordings matching the phonetic content of the pathological speech to create a reference model (phonetically-balanced scenarios)
    - Not applicable to phonetically-unbalanced scenarios
- Subspace-based intelligibility measure (SIM) (Janbakhshi et al., 2019b)
  - Applicable to phonetically-unbalanced scenarios
  - Ignores temporal cues for intelligibility assessment ⇒ lower performance than P-ESTOI

Outline

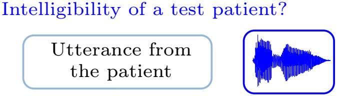
1. Pathological Speech Intelligibility Assessment
2. State-of-the-art Automatic Intelligibility Measures
3. P-ESTOI with Synthetic Speech References
4. Experimental Results
5. Summary

P-ESTOI with synthetic speech references

- Goal ⇒ Flexible version of P-ESTOI (applicable in phonetically-unbalanced scenarios)

P-ESTOI with synthetic speech references

- P-ESTOI$_H$ with healthy speech references (applicable in phonetically-balanced scenarios)

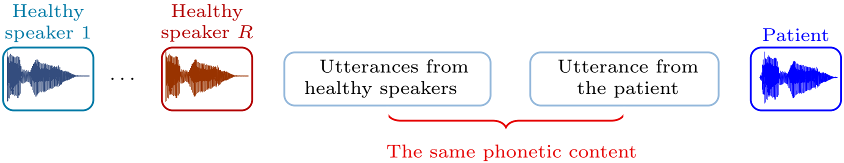


Intelligibility of a test patient?

Utterance from the patient

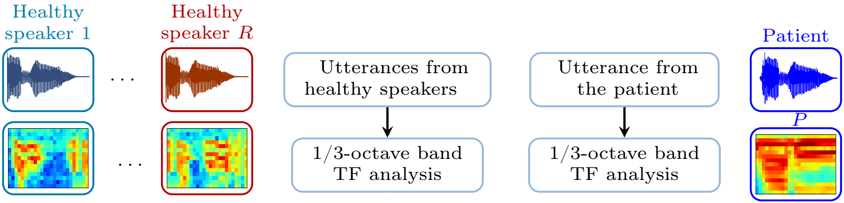
P-ESTOI with synthetic speech references

- P-ESTOI$_H$ with healthy speech references (applicable in phonetically-balanced scenarios)


Healthy speaker 1
. . .
Healthy speaker R
Utterances from healthy speakers
Utterance from the patient
Patient
The same phonetic content

P-ESTOI with synthetic speech references

- P-ESTOI$_H$ with healthy speech references (applicable in phonetically-balanced scenarios)



Healthy speaker 1

. . .

Healthy speaker R

Utterances from healthy speakers

1/3-octave band TF analysis

Utterance from the patient

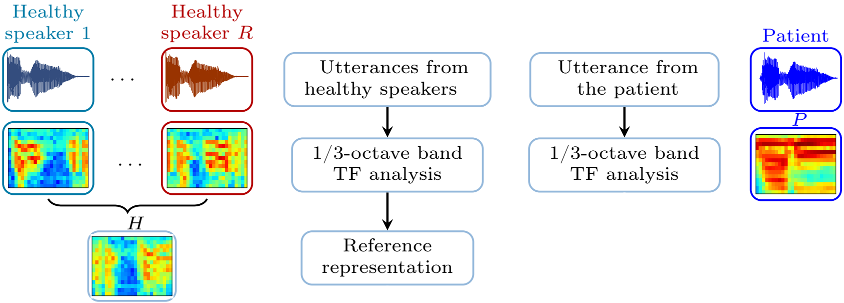
1/3-octave band TF analysis

Patient

P

P-ESTOI with synthetic speech references

- P-ESTOI$_H$ with healthy speech references (applicable in phonetically-balanced scenarios)



Utterances from healthy speakers

1/3-octave band TF analysis

Reference representation

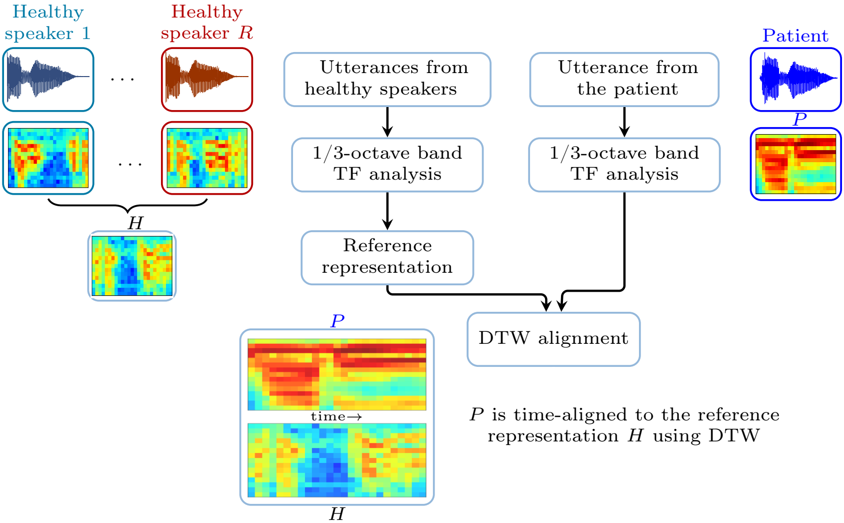
Utterance from the patient

1/3-octave band TF analysis

Creating an utterance-dependent reference H by:
dynamic time warping (DTW) + temporal clustering

P-ESTOI with synthetic speech references

- P-ESTOI$_H$ with healthy speech references (applicable in phonetically-balanced scenarios)



Healthy speaker 1

Healthy speaker R

Patient

. . .

. . .

H

P

Utterances from healthy speakers

1/3-octave band TF analysis

Reference representation

Utterance from the patient

1/3-octave band TF analysis

DTW alignment

P

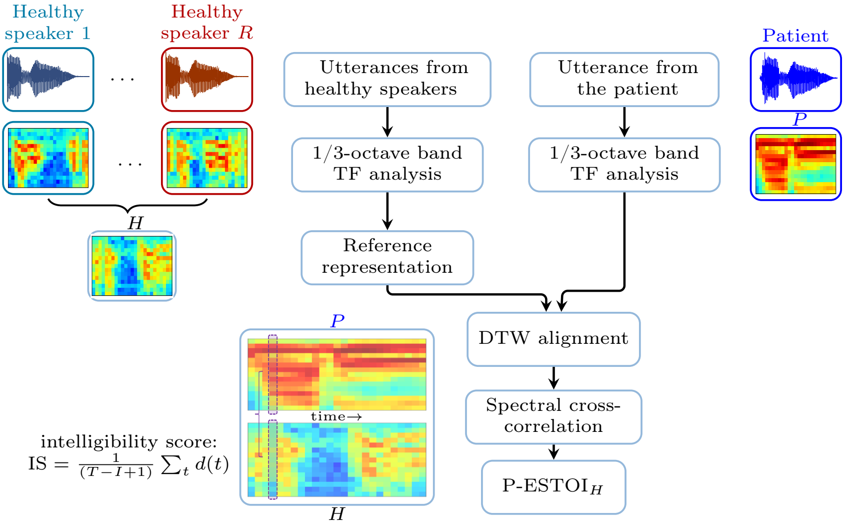
time→

H

P is time-aligned to the reference representation H using DTW

P-ESTOI with synthetic speech references

- P-ESTOI$_H$ with healthy speech references (applicable in phonetically-balanced scenarios)



Healthy speaker 1

. . .

Healthy speaker R

H

Patient

P

Utterances from healthy speakers

1/3-octave band TF analysis

Reference representation

Utterance from the patient

1/3-octave band TF analysis

DTW alignment

Spectral cross-correlation

P-ESTOI$_H$

P

time→

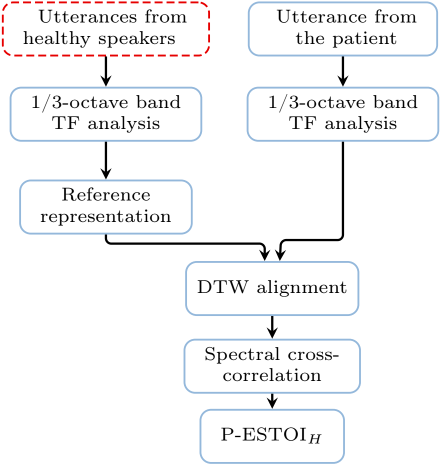
H

intelligibility score:

$$\text{IS} = \frac{1}{(T-I+1)} \sum_t d(t)$$

P-ESTOI with synthetic speech references

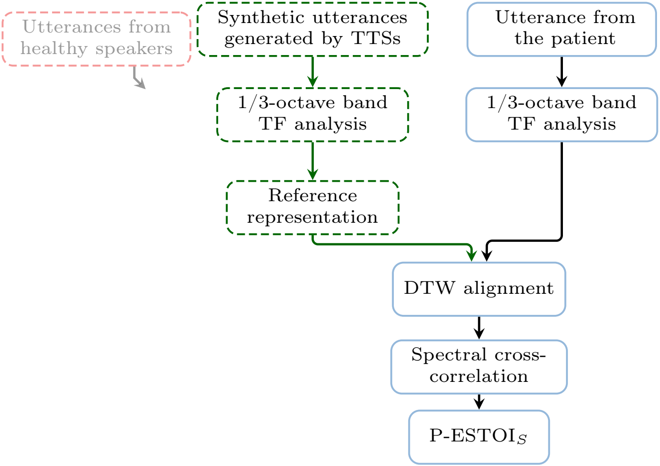
- Goal ⇒ Flexible version of P-ESTOI (applicable in phonetically-unbalanced scenarios)


Utterances from healthy speakers
Utterance from the patient
1/3-octave band TF analysis
1/3-octave band TF analysis
Reference representation
DTW alignment
Spectral cross-correlation
P-ESTOI$_H$

P-ESTOI with synthetic speech references

- Proposing to use synthetic speech generated by high-quality text-to-speech (TTS) systems to create intelligible references models

Using TTS systems as "average" intelligible speakers


Utterances from healthy speakers
Synthetic utterances generated by TTSs
Utterance from the patient
1/3-octave band TF analysis
1/3-octave band TF analysis
Reference representation
DTW alignment
Spectral cross-correlation
P-ESTOI$_S$

P-ESTOI with synthetic speech references

- Synthetic speech references generated with state-of-the-art TTS system
- Speaker-dependent TTS systems trained on multiple healthy speakers
  - Deep Neural Network (DNN)-based TTS system inspired by the state-of-the-art Merlin TTS system (Wu et al., 2016)
  - Festival front-end, two Bidirectional long short-term memory networks as duration and acoustic models, and the WORLD vocoder (Schnell and Garner, 2018)

Outline

1. Pathological Speech Intelligibility Assessment
2. State-of-the-art Automatic Intelligibility Measures
3. P-ESTOI with Synthetic Speech References
4. Experimental Results
5. Summary

Experimental results

- Databases
  - English Universal Access database (Kim et al., 2008)
    - Recordings of 764 word utterances from 15 CP patients and 4 healthy speakers
  - English CMU ARCTIC database (Kominek and Black, 2004)
    - Recordings of 1132 utterances from 4 healthy speakers ⇒ 4 TTS systems

Experimental results

- Databases
  - English Universal Access database (Kim et al., 2008)
    - Recordings of 764 word utterances from 15 CP patients and 4 healthy speakers
  - English CMU ARCTIC database (Kominek and Black, 2004)
    - Recordings of 1132 utterances from 4 healthy speakers ⇒ 4 TTS systems
- State-of-the-art intelligibility measures
  - P-ESTOI$_H$ with natural speech references (Janbakhshi et al., 2019a)
  - SIM (Janbakhshi et al., 2019b)
  - iVector and ASR-based approaches (Martínez et al., 2015)

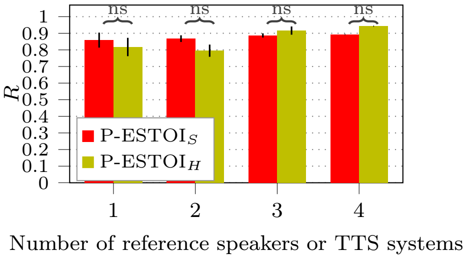
Experimental results

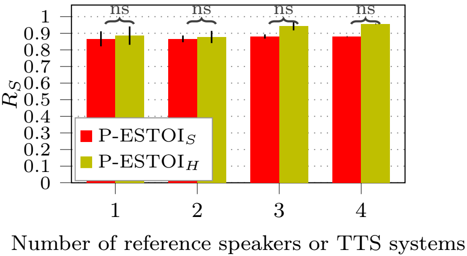
- Databases
  - English Universal Access database (Kim et al., 2008)
    - Recordings of 764 word utterances from 15 CP patients and 4 healthy speakers
  - English CMU ARCTIC database (Kominek and Black, 2004)
    - Recordings of 1132 utterances from 4 healthy speakers ⇒ 4 TTS systems
- State-of-the-art intelligibility measures
  - P-ESTOI$_H$ with natural speech references (Janbakhshi et al., 2019a)
  - SIM (Janbakhshi et al., 2019b)
  - iVector and ASR-based approaches (Martínez et al., 2015)
- Evaluation
  - Pearson correlation coefficient (R)
  - Spearman rank correlation coefficient (R_S)

Experimental results (phonetically-balanced scenarios)

» Assuming all speakers (healthy and pathological) utter the same utterances

- Considering 764 utterances for each speaker



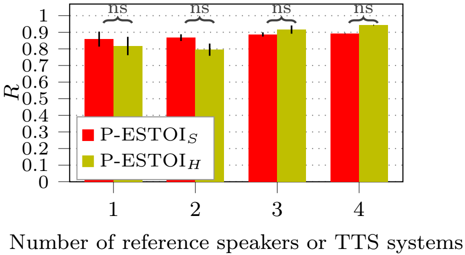


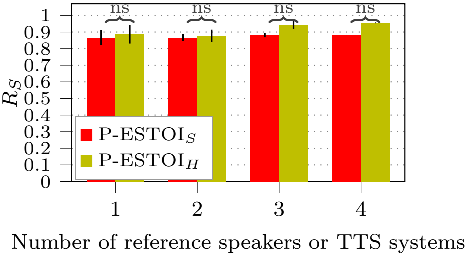
- P-ESTOI$_S$: P-ESTOI with *synthetic speech* references
- P-ESTOI$_H$: P-ESTOI with natural *healthy speech* references

Experimental results (phonetically-balanced scenarios)

» Assuming all speakers (healthy and pathological) utter the same utterances

- Considering 764 utterances for each speaker



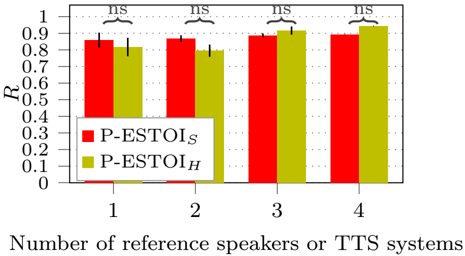


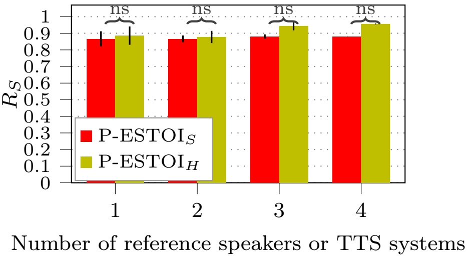
Measures	R	p	R_S	p
P-ESTOI$_S$	0.89	1e−5	0.88	6e−5
SIM	0.77	9e−4	0.84	7e−5
iVector	0.74	–	–	–
ASR	0.55	–	–	–

Experimental results (phonetically-balanced scenarios)

» Assuming all speakers (healthy and pathological) utter the same utterances

- Considering 764 utterances for each speaker



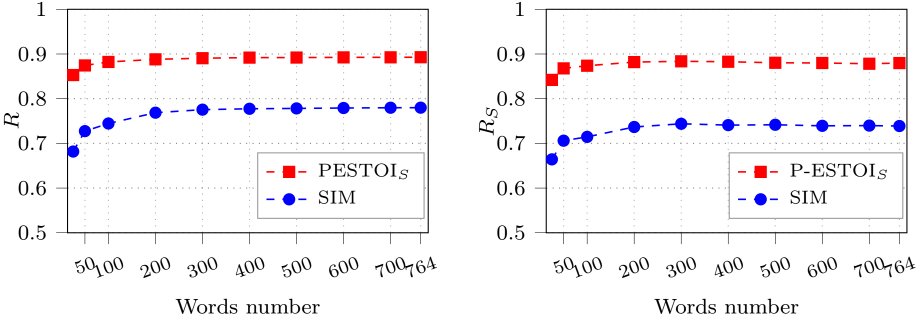


Measures	R	p	R_S	p
P-ESTOI$_S$	0.89	1e−5	0.88	6e−5
SIM	0.77	9e−4	0.84	7e−5
iVector	0.74	–	–	–
ASR	0.55	–	–	–

- No significant difference between P-ESTOI$_S$ and P-ESTOI$_H$
- P-ESTOI$_S$ yields high and significant correlations outperforming other measures

Experimental results (phonetically-unbalanced scenarios)

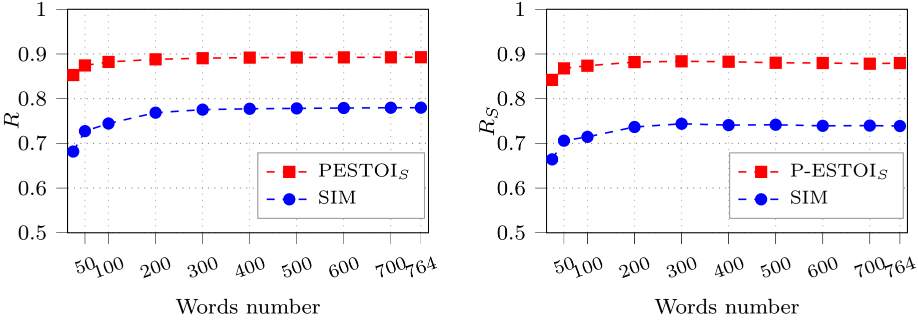
- Assuming all speakers (healthy and pathological) utter different set of utterances
  - Random selection of sets of utterances from the 764 available utterances for each speaker



Experimental results (phonetically-unbalanced scenarios)

» Assuming all speakers (healthy and pathological) utter different set of utterances

- Random selection of sets of utterances from the 764 available utterances for each speaker



» P-ESTOI$_S$ outperforms SIM in phonetically-unbalanced scenarios

Outline

1. Pathological Speech Intelligibility Assessment

2. State-of-the-art Automatic Intelligibility Measures

3. P-ESTOI with Synthetic Speech References

4. Experimental Results

5. Summary

Summary

- Creating the reference representations required in P-ESTOI using synthetic utterances generated by state-of-the-art TTS systems
    - Making P-ESTOI a flexible measure to be also used in phonetically-unbalanced scenarios
- Based on experimental results on CP patients, the performance of P-ESTOI using synthetic speech references is comparable to using natural speech references
- P-ESTOI using synthetic speech references outperforms state-of-the-art automatic intelligibility measures

Thank You

Reference

Janbakhshi, P., Kodrasi, I., and Bourlard, H. (2019a). Pathological speech intelligibility assessment based on the short-time objective intelligibility measure. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Brighton , UK.

Janbakhshi, P., Kodrasi, I., and Bourlard, H. (2019b). Spectral subspace analysis for automatic assessment of pathological speech intelligibility. In *Proc. 20th Annual Conference of the International Speech Communication Association*, Graz, Austria.

Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J., Huang, T., Watkin, K., and Frame, S. (2008). Dysarthric speech database for universal access research. In *Proc. 9th Annual Conference of the International Speech Communication Association*, pages 1741–1744, Brisbane, Australia.

Kominek, J. and Black, A. (2004). The CMU Arctic speech databases. In *Proc. 5th International Speech Communication Association Speech Synthesis Workshop*, pages 223–224, Pittsburgh, USA.

Martínez, D., Lleida, E., Green, P., Christensen, H., Ortega, A., and Miguel, A. (2015). Intelligibility assessment and speech recognizer word accuracy rate prediction for dysarthric speakers in a factor analysis subspace. *ACM Transactions on Accessible Computing*, 6(3):10:1–10:21.

Schnell, B. and Garner, P. N. (2018). A neural model to predict parameters for a generalized command response model of intonation. In *Proc. 19th Annual Conference of the International Speech Communication Association*, pages 3147–3151, Hyderabad, India.

Wu, Z., Watts, O., and King, S. (2016). Merlin: An open source neural network speech synthesis system. In *Proc. 9th International Speech Communication Association Speech Synthesis Workshop*, pages 202–207, Sunnyvale, USA.