Pathological Speech Intelligibility Assessment based on the Short-Time Objective Intelligibility Measure

Parvaneh Janbakhshi[1,2], Ina Kodrasi[1], Hervé Bourlard[1,2]

[1]Idiap Research Institute, Speech and Audio Processing Group, Martigny, Switzerland [2]École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

{parvaneh.janbakhshi, ina.kodrasi, herve.bourlard}@idiap.ch

Motivation

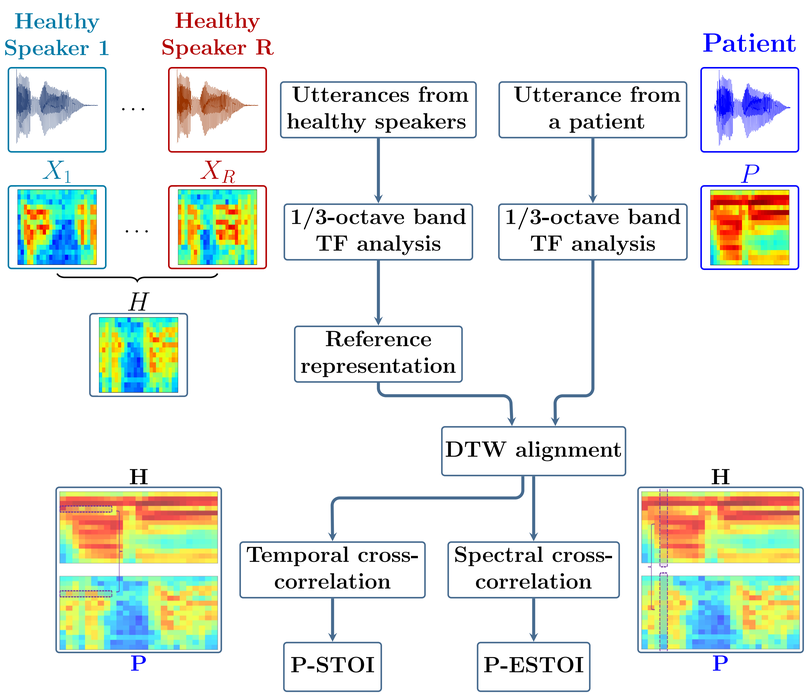
- Objective pathological speech intelligibility assessment can be crucial for the management of speech disorders
- Short-Time Objective Intelligibility (STOI) and Extended STOI (ESTOI) are successful objective intelligibility measures used in speech enhancement

Aim

- Develop reliable measures to automatically assess pathological speech intelligibility based on STOI and ESTOI

Challenge

- Enhancement objective measures → comparing time-aligned noisy and reference (clean) signals
- A time-aligned fully intelligible (reference) version of the patients' speech signal is not available



Proposed method

1. Time alignment using Dynamic Time Warping (DTW)
2. Creating utterance-dependent reference representations from multiple healthy speakers
3. Intelligibility assessment
   - **P-STOI** and **P-ESTOI** computed based on the **temporal** or **spectral cross-correlation** of the aligned pathological and reference signals

P-STOI and P-ESTOI

- $H_j(t)$ and $P_j(t)$: time-frequecny units of the aligned healthy reference and pathological test representations
- j: octave band index, t: frame index
- $\overline{H_j(i)} = \frac{1}{I}\sum_{i=t}^{t+I-1} H_j(i)$ and $\overline{P_j(i)}$ is similarly defined

$$\textbf{P-STOI} = \frac{1}{(T-I+1)J}\sum_{j,t} d_j^S(t),$$

$$d_j^S(t) = \frac{\sum_{i=t}^{t+I-1}\left(H_j(i)-\overline{H_j(i)}\right)\left(P_j(i)-\overline{P_j(i)}\right)}{\sqrt{\sum_{i=t}^{t+I-1}\left(H_j(i)-\overline{H_j(i)}\right)^2\sum_{i=t}^{t+I-1}\left(P_j(i)-\overline{P_j(i)}\right)^2}}$$

- $\overline{H_j(i)} = \frac{1}{J}\sum_{j=1}^{J} H_j(i)$ and $\overline{P_j(i)}$ is similarly defined

$$\textbf{P-ESTOI} = \frac{1}{(T-I+1)}\sum_{t} d^E(t),$$

$$d^E(t) = \frac{1}{I}\sum_{i=t}^{t+I-1}\frac{\sum_{j=1}^{J}\left(H_j(i)-\overline{H_j(i)}\right)\left(P_j(i)-\overline{P_j(i)}\right)}{\sqrt{\sum_{j=1}^{J}\left(H_j(i)-\overline{H_j(i)}\right)^2\sum_{j=1}^{J}\left(P_j(i)-\overline{P_j(i)}\right)^2}}$$

Evaluation

Databases

- 10 English-speaking Cerebral Palsy (CP) patients and 13 healthy speakers
- 10 French-speaking (Amyotrophic Lateral Sclerosis) ALS patients and 41 healthy speakers

Criteria

- Pearson (R) and Spearman rank (R_s) correlation coefficients between estimated scores and the subjective intelligibility scores (along with p-values)

Comparison

- State-of-the-art feature-based measures such as Low-to-High Modulation energy Ratio (LHMR) and standard deviation of the zeroth order delta coefficient σ_Δ

Results

Measures	R	p	R_S	p
English CP database				
P-STOI	**0.90**	5e−4	**0.82**	7e−3
P-ESTOI	**0.95**	4.3e−5	**0.91**	2e−4
σ_Δ	0.45	0.20	0.51	0.13
$LHMR$	-0.55	0.09	-0.54	0.10
French ALS database				
P-STOI	**0.87**	2e−3	0.37	0.33
P-ESTOI	**0.95**	5.6e−5	0.43	0.32
σ_Δ	**0.76**	0.01	0.48	0.16
$LHMR$	**-0.69**	0.03	-0.46	0.18

- P-STOI and P-ESTOI achieve a high and significant Pearson correlation
- P-ESTOI yields the best performance on both databases (capturing impact of spectral distortions is more important than temporal distortions)

Conclusion

- P-STOI and P-ESTOI can be used as reliable objective intelligibility measures for pathological speech, independently of the language or of the disease