A comparative study of noise robustness of goodness of pronunciation (GoP) measures and its modifications based on teacher's utterance

Sweekar Sudhakara[1], *Manoj Kumar Ramanathi*[1], *Chiranjeevi Yarra*[1], *Anurag Das*[2], *Prasanta Kumar Ghosh*[1]

[1] Department of Electrical Engineering, Indian Institute of Science (IISc), Bangalore

[2] Department of Computer Science and Engineering, Texas A&M University, College Station

INTRODUCTION

- Goodness of pronunciation (GoP) is effective in evaluating L2 pronunciations in computer-aided pronunciation training (CAPT)
- In real life scenarios, CAPT systems need to deal with noisy conditions
- We propose modifications to the typical lexicon based GoP

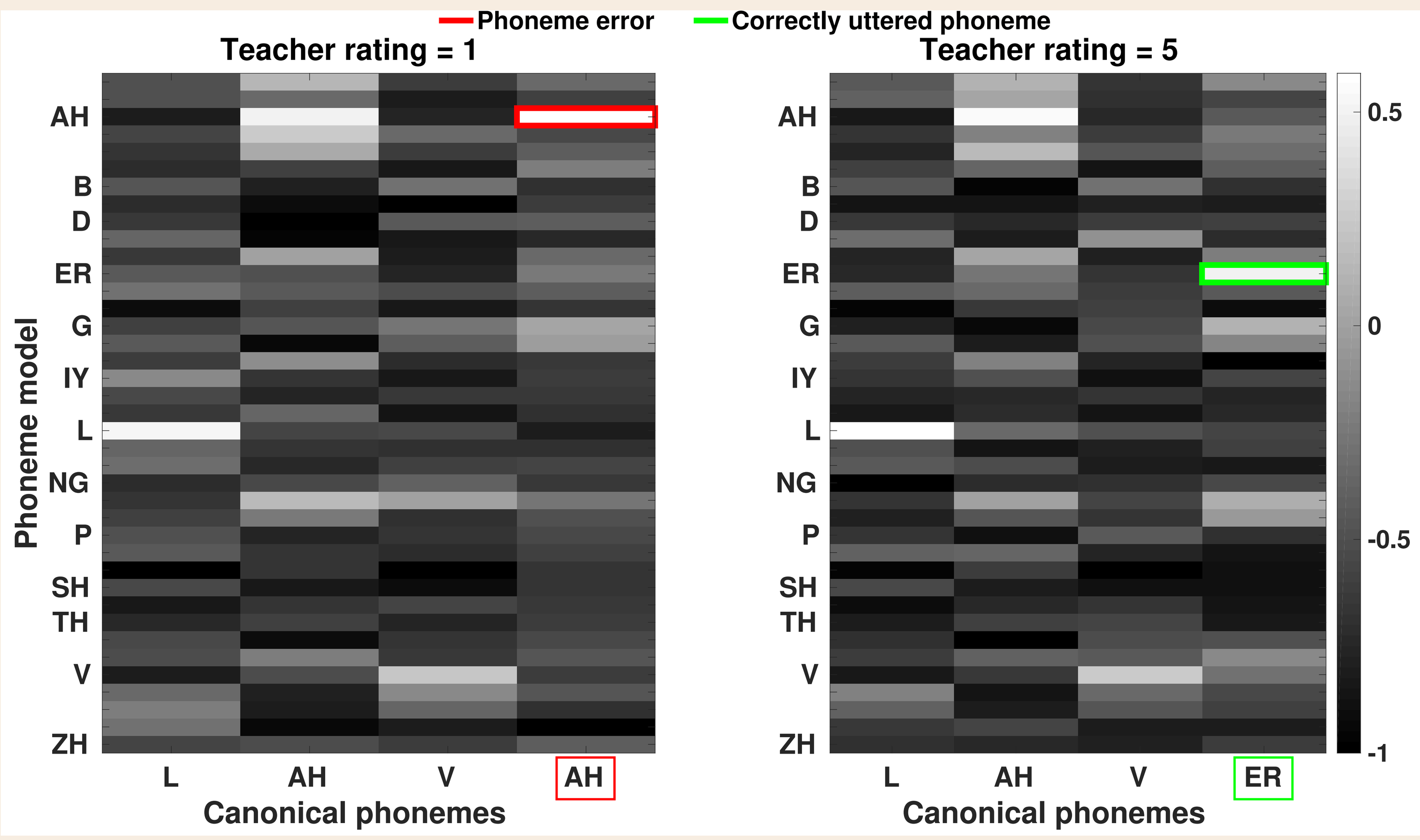
Lexicon based GoP (LGoP):

- GoP of phoneme p over the segment containing acoustic observation $\mathbf{O} = \{O_t, \forall\, 1 \le t \le T\}$ is defined as $GoP(p) = \frac{1}{T}\left|\log \mathcal{P}(p|\mathbf{O})\right|$ where T is the total number of frames in the phoneme segment[1].
- Phoneme boundaries are obtained by forced-alignment with native lexicon.

PROPOSED STUDY

Teacher's utterance based GoP (TGoP):

- Phoneme transcriptions from forced-alignment might have phoneme errors



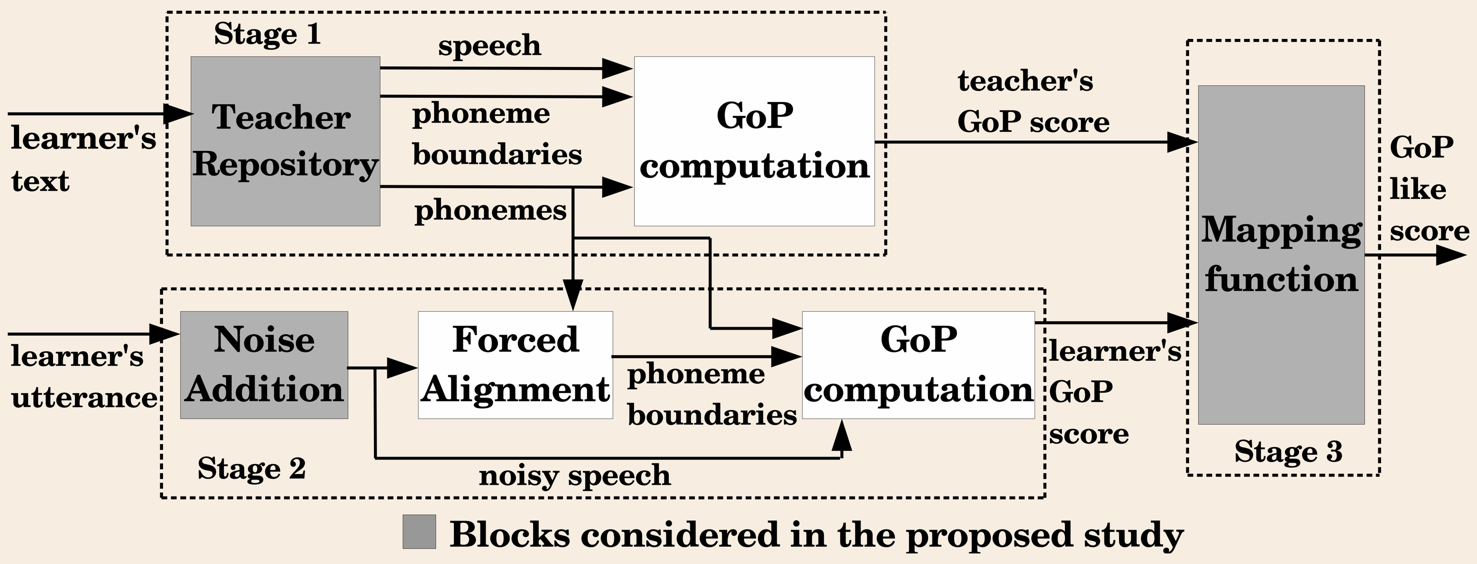
- GoP scores are closer but teacher ratings are far apart
- Propose to do forced-alignment of learner's utterance using phonemes in the teacher's utterance and then compute GoP

GoP like (GL) score:

- GoP is computed using native acoustic models. Acoustic differences might lead to poor performance
- Propose to compute score based on relative difference between GoP score of learner's utterance $GoP_l(p)$ and that of teacher's utterance $GoP_t(p)$

$$GL(p) = 1 - \tanh\left(k \times \left|\left(GoP_t(p) - GoP_l(p)\right)\Big/GoP_t(p)\right|\right)$$

- k is an empirically chosen parameter to control strictness of scoring
- $GL(p)$ is close to 1 when $GoP_t(p) \approx GoP_l(p)$



EXPERIMENTAL SETUP

- **GoP formulations:** Q is phoneme set, s is sub-phoneme (senone) and n is the number of senones

$$\textbf{E1: } \frac{1}{T}\left|\log \frac{\mathcal{P}(\mathbf{O}|p)\mathcal{P}(p)}{\sum_{q\in Q}\mathcal{P}(\mathbf{O}|q)\mathcal{P}(q)}\right|, \quad \textbf{E2: } \frac{1}{T}\left|\log \frac{\mathcal{P}(\mathbf{O}|p)}{\max_{q\in Q}\mathcal{P}(\mathbf{O}|q)}\right|, \quad \textbf{E3: } \frac{\mathcal{P}(\mathbf{O}|p)\mathcal{P}(p)}{\sum_{q\in Q}\mathcal{P}(\mathbf{O}|q)\mathcal{P}(q)},$$

$$\textbf{E4: } \frac{1}{T}\left[\sum_{t=1}^{T}\log \mathcal{P}(O_t|p) - \max_{\{q\in Q, q\neq p\}}\sum_{t=1}^{T}\log \mathcal{P}(O_t|q)\right], \quad \textbf{E5: } \frac{1}{T}\sum_{t=1}^{T}\log\frac{\mathcal{P}(s_t|O_t^{(p)})}{\mathcal{P}(s_t)},$$

$$\textbf{E6}^{2}\textbf{: } \frac{1}{T}\left[\sum_{t=1}^{T}\log \mathcal{P}(s_t|O_t^{(p)}) + \sum_{t=2}^{T}\log \mathcal{P}(s_t|s_{t-1}) + (T-1)\log n\right]$$

- **Additive noises:** babble, white Gaussian, f-16 at 0 dB, 10 dB and 20 dB
- **Evaluation metric:** Pearson correlation coefficient between utterance level GoP scores and the expert ratings
- **DNN-HMM based acoustic model**: trained on LibriSpeech corpus

DATABASE

- Read English corpus collected from 16 Indian English learners (L)
- Each learner reads 415 single words and 385 multiple words stimuli
- Learners belong to 6 different native languages - Malayalam (4L), Kannada (5L), Telugu (3L), Tamil (2L), Hindi (1L) and Gujarati (1L)
- A spoken English expert manually rated each utterance on a scale of 5 to 1 based on native language influence
- Recordings of noises from NOISEX-92 database were used

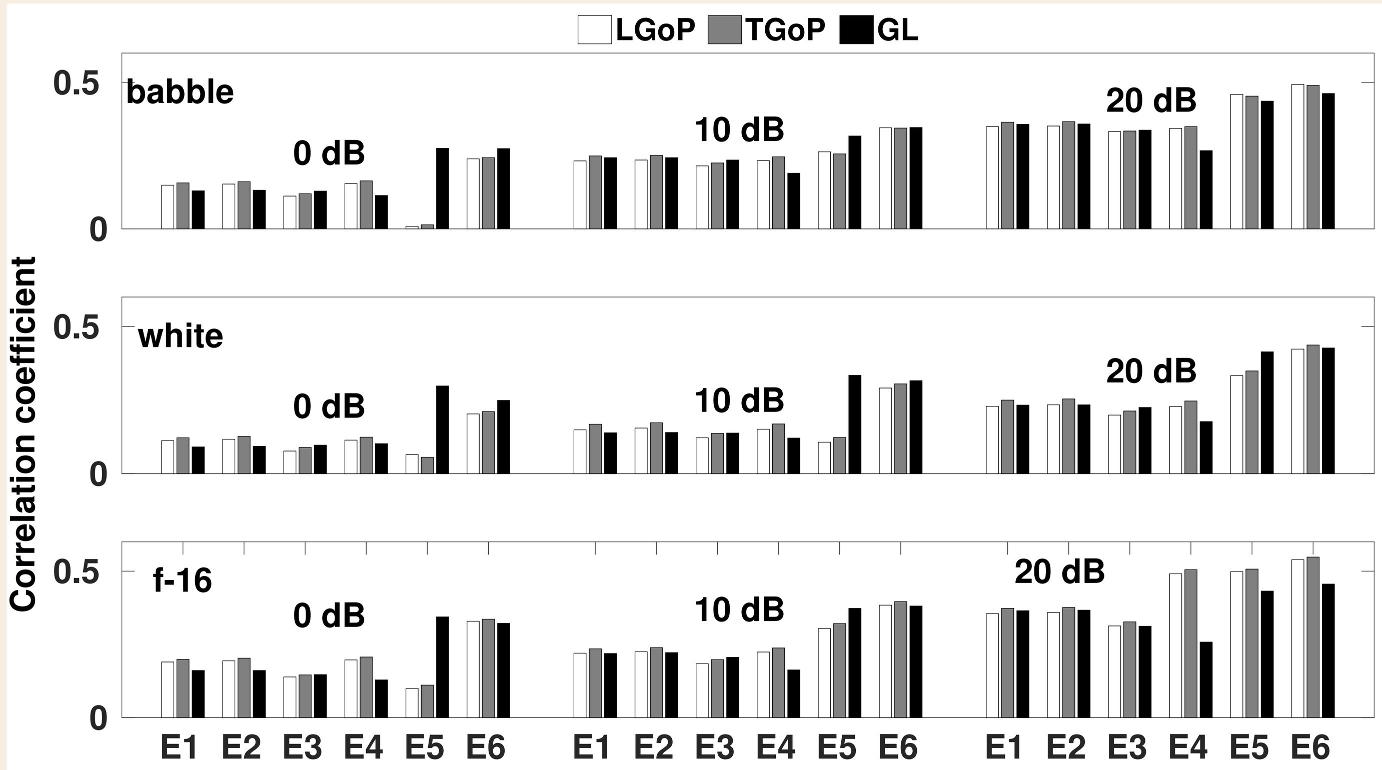
RESULTS & DISCUSSION

Comparison across GoPs with clean speech:

	E1	E2	E3	E4	E5	E6
LGoP	0.4423	0.4450	0.4223	0.4504	0.5658	0.6245
TGoP	**0.4702**	**0.4726**	**0.4488**	**0.4806**	**0.5808**	**0.6399**
GL	0.4587	0.4582	0.4106	0.3201	0.5234	0.5681

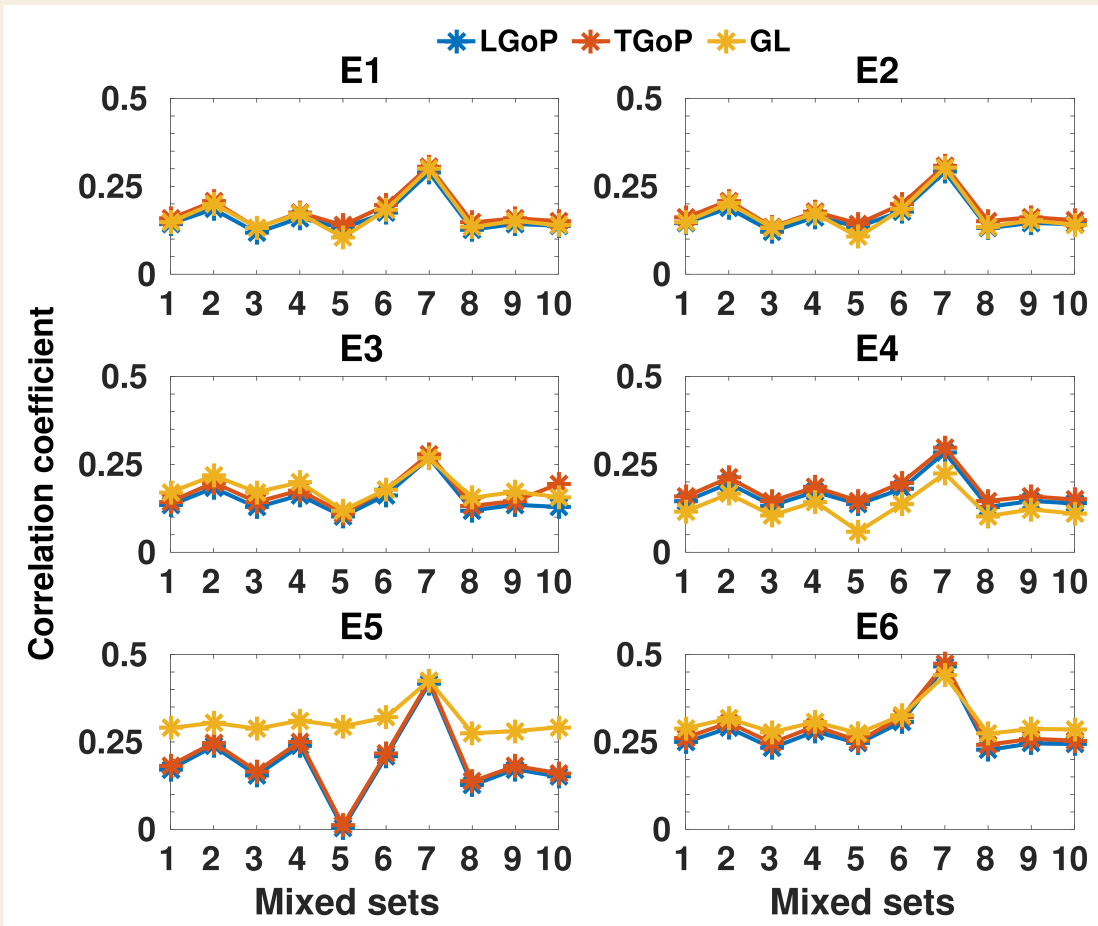
- Correlation coefficient obtained with TGoP is higher than that with LGoP for all the six GoP formulations

Comparison across GoPs with noisy speech:



- Correlation coefficient increases with increasing SNR
- Correlation coefficient obtained with TGoP and GL are higher than that with LGoP for E3, E5 and E6

Comparison across GoPs with mixed speech:



- Set 1: equal amount of recordings from clean speech data and noisy speech data under all three noises at all three SNRs
- Set 2, 3 & 4: babble, white and f-16 under all three SNRs
- Set 5, 6 & 7: 0 dB, 10 dB and 20 dB SNRs under all three noises
- Set 8, 9 & 10: babble & white, white & f-16 and babble & f-16

- Correlation coefficient obtained with TGoP is higher than that with LGoP in all sets and all GoP formulations

CONCLUSION

- Studied the variations in performance of GoP under noisy speech conditions
- Proposed TGoP and GL score as modifications to GoP for noise robustness

ACKNOWLEDGEMENT: Authors thank the **Department of Science & Technology, Government of India** and the **Pratiksha Trust** for their support

REFERENCES

1. S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning", Speech communication, vol. 30, no. 2-3, pp. 95—108, 2000
2. S. Sudhakara, M. K. Ramanathi, C. Yarra, P. K. Ghosh, "An improved goodness of pronunciation (GoP) measure for pronunciation evaluation with DNN-HMM system considering HMM transition probabilities", accepted in INTERSPEECH, 2019

{sweekars[1], manojkumar[1], chiranjeeviy[1], prasantg[1]}@iisc.ac.in, das.anurag2012@gmail.com[2]