An improved goodness of pronunciation (GoP) measure for pronunciation evaluation with DNN-HMM system considering HMM transition probabilities

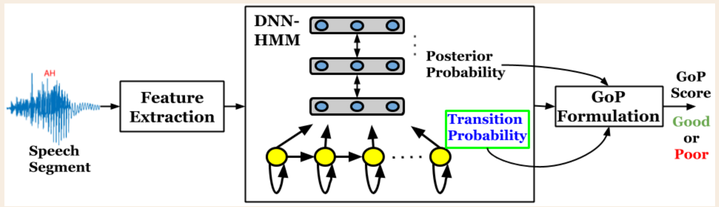
Sweekar Sudhakara, Manoj Kumar Ramanathi, Chiranjeevi Yarra and Prasanta Kumar Ghosh

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

Introduction

- Computer-aided pronunciation training (CAPT) helps non-native English learners in learning English.
- A technique known as Goodness of pronunciation (GoP) is shown to be effective in measuring pronunciation quality in CAPT.
- It is computed using Deep neural network-hidden Markov model (DNN-HMM) based acoustic model.

Proposed GoP formulation:



- We derive a formulation for GoP without any assumptions on sub-phonemic (senone) posterior probabilities and state transition probabilities (STPs).
- Existing works have neglected STPs and not explored their impact.

Database

- Read English corpus collected from 8 male (M) and 8 female (F) Indian English learners.
- Each learner reads 415 single words and 385 multiple words stimuli.
- Learners belong to 6 different native languages - Malayalam (3M+1F), Kannada (1M+4F), Telugu (2M+1F), Tamil (2M+0F), Hindi (0M+1F) and Gujarati (0M+1F).
- A spoken English expert manually rated each utterance on a scale of 5 (excellent) to 1 (poor) based on native language influence.

References

1. W. Hu, Y. Qian, and F. K. Soong, "A New DNN-based High Quality Pronunciation Evaluation for Computer-Aided Language Learning (CALL)" in INTERSPEECH, 2013, pp. 1886—1890.
2. W. Hu, Y. Qian, and F. K. Soong, "An Improved DNN-based Approach to Mispronunciation Detection and Diagnosis of L2 Learners' Speech." in SLaTE, 2015, pp. 71—76.
3. H. Huang, H. Xu, Y. Hu, and G. Zhou, "A transfer learning approach to goodness of pronunciation based automatic mispronunciation detection," The Journal of the Acoustical Society of America, vol. 142, no. 5, pp. 3165—3177, 2017.

GoP definition and its formulation

- GoP of phoneme p is defined as $GoP(p) = \frac{1}{T}\left|\log \mathcal{P}(p|\mathbf{O})\right|$. $\mathbf{O}$ is the acoustic observation and T is the total number of frames in the phoneme segment.
- Let the senone sequence, $\mathbf{s} = \{s_t, \forall\ 1 \le t \le T\}$ in a phoneme segment p and is assumed to be known. Thus, $\mathcal{P}(p|\mathbf{O}) = \mathcal{P}(\mathbf{s}|\mathbf{O}) = \mathcal{P}(s_1, s_2, ..., s_T|O_1, O_2, ..., O_T)$.
- In the left-to-right HMM, current state only depends on previous state and current observation is associated only with the current state.

Proposed GoP (PGoP):

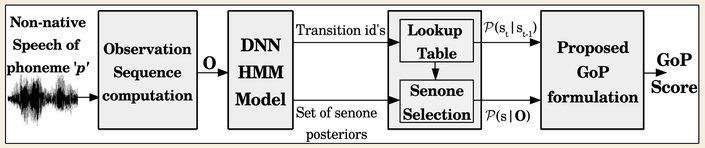
- $\mathcal{P}(\mathbf{s}|\mathbf{O})$ can be expressed in terms of **senone posteriors** $\mathcal{P}(\mathbf{s_t}|\mathbf{O_t})$, **state transition probabilities** $\mathcal{P}(\mathbf{s_t}|\mathbf{s_{t-1}})$ and **senone priors** $\mathcal{P}(\mathbf{s_t})$ as:

$$\mathcal{P}(p|\mathbf{O}) = \frac{\prod_{t=1}^{T}\mathcal{P}(s_t|O_t)\prod_{t=2}^{T}\mathcal{P}(s_t|s_{t-1})}{\prod_{t=2}^{T}\mathcal{P}(s_t)} \quad (1)$$

- Applying $\log$ on Equation 1 and assuming that all senones are equally likely:

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

where n is the total number of senones.



Experimental setup

- **Baseline GoP formulations:**

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

$$\textbf{(BL-2)}^{[2]},\ GoP(p) = \frac{1}{T}\sum_{t=1}^{T}\log\frac{\mathcal{P}(s_t|O_t^{(p)})}{\mathcal{P}(s_t)} \quad \&\quad \textbf{(BL-3)}^{[3]},\ GoP(p) = \frac{1}{T}\sum_{t=1}^{T}\log\mathcal{P}(s_t|O_t^{(p)})$$

- **Utterance level score:**
  - **Single word level score:** Average of GoP scores across all phonemes in the word.
  - **Multiple word level score:** Average of GoP scores across all words in the utterance.
- **Evaluation metric:** Pearson correlation co-efficient between utterance level GoP scores and the expert ratings.
- **DNN-HMM based acoustic models**: LibriSpeech (LS) and Fisher-English (FE) acoustic models trained with LS and FE data respectively.

Results & Discussion

Comparison of GoP formulations:

	BL-1		BL-2		BL-3		PGoP	
	LS	FE	LS	FE	LS	FE	LS	FE
Male Speakers	0.468	0.305	0.623	0.358	0.637	0.401	**0.653**	**0.452**
Female Speakers	0.434	0.266	0.593	0.306	0.605	0.343	**0.624**	**0.396**
All Speakers	0.453	0.273	0.606	0.316	0.619	0.356	**0.637**	**0.409**

- PGoP performs better than BL-1, BL-2 & BL-3.

Word specific comparison:

	BL-1		BL-2		BL-3		PGoP	
	LS	FE	LS	FE	LS	FE	LS	FE
Single Word	0.5229	0.4314	0.6111	0.4914	0.6263	0.5015	**0.6272**	**0.5072**
Multiple Words	0.4687	0.3603	**0.5286**	0.3913	0.5283	0.4002	0.5210	**0.4099**

- PGoP performs better in single word case than in multiple words case.

Language specific comparison:

	BL-1		BL-2		BL-3		PGoP	
	LS	FE	LS	FE	LS	FE	LS	FE
Malayalam	0.425	0.221	0.585	0.289	0.606	0.334	**0.631**	**0.394**
Kannada	0.421	0.241	0.592	0.271	0.605	0.317	**0.621**	**0.368**
Tamil	0.442	0.230	0.603	0.281	0.619	0.340	**0.650**	**0.418**
Telugu	0.515	0.344	0.663	0.409	0.671	0.436	**0.679**	**0.475**
Hindi	0.439	0.312	0.554	0.329	0.563	0.359	**0.584**	**0.412**
Gujarati	0.398	0.275	0.551	0.241	0.550	0.271	**0.561**	**0.316**

- PGoP is not influenced by learner's native language.

Conclusion

- Proposed GoP formulation is a function of senone posteriors and state transition probabilities.
- It correlates better with expert ratings compared with the three baselines.
- **Future work:** To analyze the trade-offs between the improvements and computational efforts involved in the GoP formulations using acoustic models trained on different corpora.

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

{sweekars, manojkumar, chiranjeeviy, prasantg}@iisc.ac.in