Pronunciation scoring for dysarthric speakers with DNN-HMM based goodness of pronunciation (GoP) measure

Shruti Jeyaraman, Ananthakrishnan K., Vijayalakshmi P.[*], Nagarajan T.[†]
[*]Department of Electronics and Communication Engineering, Sri Sivasubramaniya Nadar College of Engineering
[†]Department of Computer Science and Engineering, Shiv Nadar University
shruti2010202@ssn.edu.in, ananthakrishnan2010011@ssn.edu.in,
vijayalakshmip@ssn.edu.in, nagarajant@snuchennai.edu.in

Abstract

Dysarthria is a neurological motor disorder caused by cranial damage that interferes with the muscles involved in the correct pronunciation of sounds and intelligible speech. Computer Aided Pronunciation Training (CAPT) systems traditionally used for the pronunciation assessment of L2 language learners can offer a method to detect and score mispronounced sounds in dysarthric speakers as a way of evaluation without human intervention. In this work, a phonetic level DNN-HMM based Goodness of Pronunciation (GoP) for pronunciation scoring, on native Tamil Dysarthric speakers corpus is presented. The scores are calculated using the posteriors of the subphonemic elements called senones with a focus on their prevalence across phones and their transitions across HMM states. The phonetic-level scores obtained for speakers of different levels of severity help establish speaker-specific trends in pronunciation through an objective log-likelihood metric, in contrast to subjective evaluations by Speech Language Therapists (SLTs).

1 Introduction

Damage caused to the brain upon injury or through a developmental disorder results in dysarthria, a neuromotor speech disorder leading to the discoordination of the speech production muscles which results in dysarthric speech sounding nasal, breathy, jerky, raspy, harsh and incorrect (Palmer and Enderby, 2007), and suffering reduced intelligibility. Freed (2018) and Clark and Solomon (2012) note that current methods of assessment of dysarthria are performed by speech language therapists (SLTs) through case histories, non-speech examinations of speed, strength and steadiness of speech musculature, and examinations of the phonation, articulation and prosody. Treatment comprises general exercises to strengthen respiration, phonation and articulation to improve the range and strength of muscles involved in speech production (Enderby, 2013) administered by SLTs. Previous work (Mariya Celin et al., 2019) has discussed the development of a speech communicative aid to differentiate pronunciation and system errors, and subsequently correct dysarthric speech electronically. However, this work focuses on analysing speaker-specific weaknesses in their speech production system using place of articulation analysis which can guide SLTs to tailor customized treatment for patients and track their progress over time.

Computer Aided Pronunciation Training (CAPT) systems have been long used for phone-level pronunciation assessment among L2 language learners. Such systems use Automatic Speech Recognizers (ASRs) to decode the phonemes uttered, followed by an evaluation or feedback system to score the learner's pronunciation of the words (Wang et al., 2019). Consistent research has depicted the superiority of deep neural networks-based ASRs (DNN) over Gaussian mixture model-ASRs, with improved word error rates (WERs) (Hu et al., 2013).

Witt and Young (2000) proposed the Goodness of Pronunciation (GoP) metric based on a log- likelihood ratio to evaluate segment-level pronunciation. Following various reformulations and revisions especially suited to DNN-based approach, the evaluation of the GoP using the sub-phonemic posterior probabilities along with state transition probabilities (STPs) as outlined by Sudhakara et al. (2019) has been used in the mispronunciation detection and scoring task of the SSNCE Tamil Dysarthric Speech Corpus (Celin et al., 2020) with a modification to account for the empirical senone distribution as actually appearing in the corpus.

Section 2 of this paper outlines the design of the mispronunciation evaluation system, explaining the development of the ASR, the Goodness of Pronunciation formulation and the data used in the work. Section 3 describes the experimental setup and implementation aspects. The results and subsequent analyses are presented in Section 4.

2 Methods

The mispronunciation evaluation system consists of two main parts, the Automatic Speech Recognition (ASR) phase and the GoP evaluation phase.

2.1 Automatic Speech Recognition (ASR)

A phone-level ASR is a statistical recognition system that decodes the sequence of phonemes using phonetic acoustic model and language grammar model.

For the Deep Neural Network (DNN) serving as the acoustic model, the input comes from acoustic features extracted from the speech signals, namely Mel-frequency cepstral coefficients (MFCCs). The language model, modelled using n-grams, handles the sequence of phonemes, while the HMM models the transitions between sub-phonemic states for each phoneme. The combined acoustic and language model is then used to search for the phonetic sequence with the maximum likelihood of occurrence to decode the utterance down to its phonetic makeup.

The outputs of the DNN are the posterior probabilities of the senone (sub-phonemic) state to which the input belongs. Along with the transition probabilities of the triphone HMM states, the output of the DNN-HMM system is the posterior probability of the phones in each utterance. The speech decoding process uses the Viterbi algorithm.

2.2 Goodness of Pronunciation (GoP)

Post decoding, the speech is scored using the Goodness of Pronunciation (GoP) metric that evaluates the ‘correctness’ of pronunciation of the phones in a speech input. It has evolved from the use of the straightforward logarithmic posterior probability of the target phone using GMM-HMM systems, to more sophisticated formulations accounting for the senonic contributions to phonemic posterior probability through DNN-HMMs as well. The utterances of speakers in our Tamil dysarthric speech corpus have been evaluated with GoP considering the log posteriors of the senones making up the phones. The GoP formulation implemented for this work is derived from Sudhakara et al. (2019), with a modification. While the original work assumes that all senones occur with equal probability, the current work modifies the formulation to follow the empirical distribution of senones in the corpus. The proposed GoP formulation is

$$GoP(p) = \frac{1}{T}\left[\begin{array}{l}\sum_{t=1}^{T}\log P(s_t|y_t) \\ +\sum_{t=2}^{T}(\log P(s_t|s_{t-1}) \\ -\log P(s_t))\end{array}\right] \quad (1)$$

The formulation thus accounts for the senonic posterior probabilities as well as their transition probabilities across HMM states and the empirical probability of occurrence of the senones.

2.3 Speech Dataset

The data used in this work is the SSNCE Tamil Dysarthric Speech Corpus (SSNCE-TDSC) (Celin et al., 2020) . The corpus contains approximately 8 hours of time-aligned Tamil dysarthric speech data and metadata collected from a total 30 native Tamil speakers with 7 mild, 10 moderate and 3 severely dysarthric speakers, as well as 10 non-dysarthric speakers who form the ‘normal pronunciation baseline’ in our experiments.

Each speaker has spoken 365 Tamil utterances with at least 25 examples occurring for each phoneme to ensure statistical model training. The utterances include a combination of common and uncommon Tamil phrases. The corpus also contains clinical data, an SLT-ascertained speech intelligibility score and a descriptive speech assessment for every speaker provided by an SLT (Celin et al., 2016).

3 Implementation

3.1 Experimental Setup

Equation 1 is considered the baseline GoP formulation. The DNN-HMM acoustic model used in the GoP extraction is trained on the normal speakers of SSN-TDSC. The GoP score obtained per phone, is then averaged across all phones occurring in an utterance, to get the GoP of the utterance. This helps exhibit the trend in pronunciation from one severity category to another. The Kaldi-toolkit is used in training the DNN-HMM system based on Nnet2 recipe. The number of senones at the output layer of the DNN is 1191.

3.2 Implementation Aspects

13 dimensional Mel Frequency Cepstral Coefficients (MFFCs) of dysarthric speech, computed over all frames with a frame size of 25 ms with a

frame shift of 10 ms, are used to train the DNN-HMM system. First, posteriors are generated for all the senones present in the system for a given utterance. Then, for the given utterance, the composite senone sequence, typically known as transition-ids in Kaldi are obtained using forced Viterbi alignment. A lookup table with mappings between transition-ids and the senones is generated, which is useful in the decoding of the senone sequence and its transition probabilities. Then, the empirical distribution of senones is taken to obtain the senonic prior probabilities per their occurrence in the data. The GoP is then computed using the transition-ids and the probability of the selected senone sequence.

The scores obtained through the proposed implementation are compared against the Kaldi DNN-based Compute GoP implementation which is derived from Hu et al. (2015) as outlined here.

$$GoP(p) = \log \frac{LPP(p)}{\max_{q \in Q} LPP(q)} \qquad (2)$$

LPP is the Log Phone Posterior for a phone which is the logarithmic posterior of the phone averaged over the duration of its frame. This implementation does not account for transition probabilities between the senones.

4 Results

Figure 1 illustrates the spectrogram for an utterance. Presented below are the overall frame-level GoP scores for every phone. The frame-level GoP scores are smoothened to obtain the phone-level GoPs, based on the number of frames allotted to each canonical phone during the forced-alignment phase, which are depicted in the third plot.

A comparison of phone-level GoP scores is shown in the time-normalised plots Figure 2, presenting the variations in scores across contiguous phones of the same utterance between a mild and moderate dysarthric speaker. It can be seen that for the mild speaker (blue), the variations in pronunciation within an utterance are lesser than that of a moderate speaker (green).

Further analysis of speaker-specific phonetic-level pronunciation is performed for all 20 dysarthric speakers for five classes of phones based on their places of articulation. The phone scores based on the proposed approach, for selected speakers in the three dysarthric categories are presented in Table 1. The scores offer insights into speaker-specific differences in the articulation of different phones and inter-speaker pronunciation variations within the same dysarthric category.

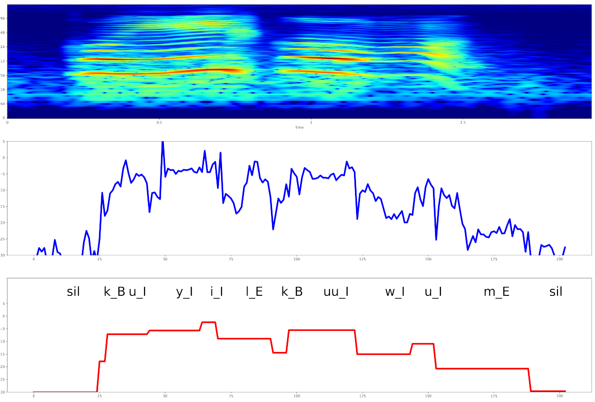

Figure 1: Spectrogram for utterance "*kuyil kuuwum*" spoken by moderate speaker FGA along with per-frame, and per-phone GoP scores

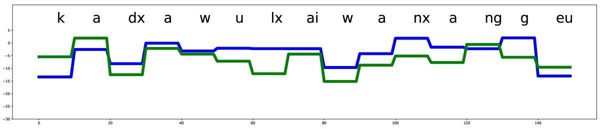

Figure 2: A comparison of a mild (blue) and a moderate (green) speakers' phone-level scores for the utterance "*kadxawulxai wanxanggeu*"

- There is a higher degree of mispronunciation of bilabial sounds for speakers in all categories. This is consistent with the SLP assessment, which noted a predominant drooling and lip closure problem across most speakers in the corpus.

- Horizontally, the scores get progressively lower (more negative) from mild to moderate to severe speakers, indicating the overall higher rate of erroneous pronunciation by severe speakers for a larger subset of phones in the language.

- For mild speakers FSI and MAK, phones p, b, m, t, tx, dx, k are some of the most mispronounced. This indicates that these speakers may possibly suffer from a tongue movement problem and might need specific attention in pronunciation training for these phones.

- The mild speaker MPA's clinical information presented in (Celin et al., 2016) states that

		Mild		**Moderate**		**Severe**	
Place of Articulation	consonant	FSI	MAK	FGA	MKA	MRI	MMA
Bilabial	/p/	-16.760	-17.974	-16.709	-17.301	-18.315	-17.106
	/b/	-11.406	-11.887	-11.218	-9.886	-16.592	-10.068
	/w/	-7.800	-8.033	-10.092	-11.358	-13.113	-11.330
	/m/	-11.704	-11.180	-12.338	-10.434	-13.472	-12.209
Dental	/t/	-12.928	-12.433	-13.525	-12.894	-13.563	-13.624
	/d/	-5.112	-7.095	-11.254	-8.807	-11.314	-7.666
Alveolar	/s/	-5.318	-6.599	-6.9	-5.339	-9.022	-8.33
	/n/	-6.077	-5.657	-9.435	-4.45	-11.018	-7.192
	/l/	-4.168	-4.153	-9.348	-7.765	-9.976	-7.26
	/r/	-5.259	-5.378	-8.766	-6.985	-8.409	-6.713
	/tx/	-14.847	-14.575	-14.875	-12.887	-15.67	-14.88
	/dx/	-9.35	-8.936	-13.381	-7.822	-13.041	-10.859
Palatal	/c/	-5.151	-7.099	-7.096	-3.117	-8.456	-8.201
	/sx/	-5.429	-5.038	-7.513	-5.502	-7.592	-6.209
	/zh/	-1.392	-4.048	-2.236	-1.393	-8.673	-5.093
	/j/	-4.515	-4.84	-7.549	-6.418	-12.467	-5.858
Velar	/k/	-11.369	-11.603	-11.724	-11.779	-11.909	-13.853
	/g/	-6.222	-6.316	-11.393	-7.298	-11.154	-8.35
	/ng/	-0.466	-1.458	-4.911	-0.375	-8.804	-3.131

Table 1: GoP scores obtained for speakers from mild, moderate and severe categories for consonant phones of 5 classes. The more negative the score, the worse the pronunciation.

Utterances	**Kaldi Compute GoP**			**Proposed GoP formulation**		
	MAK **(mild)**	**FBL** **(mod)**	**MRI** **(sev)**	**MAK** **(mild)**	**FBL** **(mod)**	**MRI** **(sev)**
kadxawulxai wanxanggeu	-1.137	-1.138	-1.1364	-5.787	-7.946	-9.092
mayil agawum	-0.809	-0.814	-0.814	-6.184	-9.185	-10.764
manam tirxanddeu peeseu	-1.122	-1.124	-1.125	-7.001	-8.962	-10.436

Table 2: Scores for a few utterances obtained from Kaldi's compute-gop and the proposed approach, categorized by speaker-severity

he exhibits tongue protrusion and stressful speech. Interestingly, his GoP scores for palatal phones are closer to zero, rather than being skewed toward the more negative range. Given his tongue protrusion issue, his scores for both the alveolar and dental classes are the most negative among the mild category of speakers.

- For the moderate speakers MVI and FGA present restrictions in their tongue movements (Celin et al., 2016), alveolar and dental scores are the most negative within their speaker class. Interestingly, high recurrence of more negative scores in FGA could project that she strongly leans towards the 'severe' category of speakers.
- The moderate speakers MGN and FBL are noted to have an absence of lip closure (Celin et al., 2016), with FBL's bilabial scores being the lowest in the bilabial class.
- SLT assessment of severe speaker MRI indicates lip closure issue and severe drooling and the clinical diagnosis for MMA indicates his frequent drooling. MER has microcephaly condition (Celin et al., 2016). Such factors reason why the three severe speakers often have the most negative GoP scores across all classes of phones.

Table 2 presents the sentence level GoP scores from Kaldi's compute GoP formulation and the proposed formulation. The sentence-level scores are obtained by averaging the scores of the phones present in a given utterance. While the scores follow a generally decreasing trend from mild to severe speakers, the proposed formulation better discriminates the pronunciations between speakers across severity categories. Coupled with the place of articulation analysis outlined above, the GoP system illuminates challenges faced by every speaker with phones that appear in different contexts across various sentences. The specific phonetic-level feedback can help devise speaker-specific articulation treatments designed to improve treatments for dysarthric speech.

Limitations

In a DNN-HMM system, a few senones could have never occurred or occurred, albeit rarely, during the training phase. This affects the third term in the Goodness of Pronunciation formulation i.e log P(St), which leads to a small positive GoP score for some phones. However, generally the GoP score should always be less than or equal to zero. Hence, DNN-HMMs will have to be built after removing the rare senones or further tying them with closest senone(s) which are not as rare. Another limitation was evident during forced alignment, when compared with the spectrograms of the utterances. Often, silence is under-estimated during forced alignment. This leads the model to compute GoP score of a phone starting from its preceding silence phase. Hence, forced alignment algorithms may have to be revised taking into account sharp changes in spectrogram boundaries. Finally, it is to be noted that the reliability of the GoP evaluation system is dependent on the accuracy of the ASR in correctly decoding phones. Thus, improving the ASR to achieve lower Word Error Rates (WERs) would lead to more credible GoP scores.

Acknowledgement

The authors would like to thank the Ministry of Electronics and Information Technology (MeitY), Government of India, for funding the project on ’Assistive Speech technologies’ under the Project titled ’National Language Translation Mission (NLTM): BHASHINI’, Ref. No. 11(1)/2022-HCC(TDIL).

References

T.A Mariya Celin, T Nagarajan, and P Vijayalakshmi. 2016. Dysarthric speech corpus in tamil for rehabilitation research. In *2016 IEEE Region 10 Conference (TENCON)*, pages 2610–2613.

T.A. Mariya Celin, T. Nagarajan, and P. Vijayalakshmi. 2020. Data augmentation using virtual microphone array synthesis and multi-resolution feature extraction for isolated word dysarthric speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 14(2):346–354.

Heather M. Clark and Nancy Pearl Solomon. 2012. Muscle tone and the speech-language pathologist: Definitions, neurophysiology, assessment, and interventions. *Perspectives on Swallowing and Swallowing Disorders (Dysphagia)*, 21(1):9–14.

Pamela Enderby. 2013. Chapter 22 - disorders of communication: dysarthria. In *Neurological Rehabilitation*, volume 110, pages 273–278. Elsevier.

Donald B. Freed. 2018. *Motor Speech Disorders: Diagnosis and Treatment*, volume 3. Plural Publishing Inc.

Wenping Hu, Yao Qian, and Frank K. Soong. 2013. A new dnn-based high quality pronunciation evaluation for computer-aided language learning (call). In *Interspeech 2013*, pages 1886–1890.

Wenping Hu, Yao Qian, Frank K. Soong, and Yong Wang. 2015. Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers. *Speech Communication*, 67:154–166.

T. A. Mariya Celin, G. Anushiya Rachel, T. Nagarajan, and P. Vijayalakshmi. 2019. A weighted speaker-specific confusion transducer-based augmentative and alternative speech communication aid for dysarthric speakers. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(2):187–197.

Rebecca Palmer and Pamela Enderby. 2007. Methods of speech therapy treatment for stable dysarthria: A review. *Advances in Speech Language Pathology*, 9(2):140–153.

Sweekar Sudhakara, Manoj Kumar Ramanathi, Chiranjeevi Yarra, and Prasanta Kumar Ghosh. 2019. An improved goodness of pronunciation (gop) measure for pronunciation evaluation with dnn-hmm system considering hmm transition probabilities. In *Interspeech 2019*, pages 954–958.

Hongyang Wang, Jie Xu, Hai Ge, and Yufeng Wang. 2019. Design and implementation of an english pronunciation scoring system for pupils based on dnn-hmm. In *2019 10th International Conference on Information Technology in Medicine and Education (ITME)*, pages 348–352, Los Alamitos, CA, USA. IEEE Computer Society.

S.M Witt and Steve J. Young. 2000. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, 30(2):95–108.