Kannada Kali: A Smartphone Application for Evaluating Spoken Kannada Words and Detecting Mispronunciations Using Self Organizing Maps

Conference Paper · December 2018

DOI: 10.1109/T4E.2018.00009

10 authors, including:

Savitha Murthy
PES Institute of Technology
5 PUBLICATIONS **14** CITATIONS
SEE PROFILE

Avinash Kumar
Zscaler
21 PUBLICATIONS **15** CITATIONS
SEE PROFILE

Viraj Kumar
Dayananda Sagar University
51 PUBLICATIONS **184** CITATIONS
SEE PROFILE

Dinkar Sitaram
PES Institute of Technology
101 PUBLICATIONS **2,553** CITATIONS
SEE PROFILE

Kannada Kali: A Smartphone Application for Evaluating Spoken Kannada Words and Detecting Mispronunciations using Self Organizing Maps

Savitha Murthy*, Ankit Anand, Avinash Kumar
Ajay Cholin, Ankita Shetty, Aditya Bhat
Akshay Venkatesh, Lingaraj Kothiwale
Dinkar Sitaram†
PES University, Bangalore, India
*savithamurthy@pes.edu, †dinkars@pes.edu

Viraj Kumar
Indian Institute of Science, Bangalore, India
viraj.kumar.cs@gmail.com

***Abstract*—Computer Aided Pronunciation Training (CAPT) systems can assist people learning to speak new languages by detecting and correcting mispronunciations. "Kannada Kali" is a prototype Android application that leverages learners' increasing access to smartphones to evaluate the pronunciation of Kannada words and provide feedback using a cloud-based framework. A CAPT system typically uses an Automatic Speech Recognition (ASR) sub-system. For sufficient accuracy, ASR systems need to be trained using speech data from both native (L1) and non-native (L2) speakers. Since the latter type of data is particularly difficult to gather, we follow recent research efforts that seek to minimize the dependency on large speech corpora. We recorded 21 Kannada words (two to five syllables long) pronounced correctly by a Kannada teacher as templates, and 1169 samples of these words spoken by 19 native and non-native Kannada speakers aged 18 to 25 years. These samples were manually rated on a 5-point Likert scale by the Kannada teacher and used to train a neural network classifier for our application. "Kannada Kali" provides learners feedback that matches the teacher ratings with an accuracy of 86% on binary classification and 68% on multi-class classification. We also propose a novel approach for detecting mispronunciations using Self Organizing Maps (SOM) and report promising initial results.**

***Index Terms*—pronunciation evaluation, CAPT, Kannada, mispronunciation detection, SOM**

I. Introduction

A growing number of people are learning to speak new languages, either as part of their formal education or because they have migrated to a region where a different language is predominant. Thanks to advances in machine learning and increasing access to smartphones, non-native (L2) learners have access to sophisticated Computer Aided Language Learning (CALL) systems and Computer Aided Pronunciation Training (CAPT) systems that allow them to learn new languages at their own pace. As their names suggest, a CALL system acts as a guide for learning languages (including linguistic features), whereas a CAPT system focuses on pronunciation. We will focus on the latter system, which assigns a score to the learner's pronunciation. To aid learning, a CAPT system may also provide feedback indicating where the learner's pronunciation deviates from the canonical pronunciation. In this paper, we describe '*Kannada Kali*', an Android-based CAPT for learning Kannada pronunciations. Our prototype communicates with a cloud-based framework for pronunciation evaluation. We also propose a novel technique for mispronunciation detection using Self Organizing Maps (SOM).

II. Related Work

CAPT systems traditionally make use of an Automatic Speech Recognition (ASR) system that is trained on speech data from native (L1) and non-native (L2) speakers. ASR systems require resources such as lexicons (pronunciation dictionaries), transcripts, language model (language grammar) and adequate speech data (from both L1 and L2 speakers) for training. They detect mispronunciations based on the posterior probabilities or Goodness of Pronunciation (GOP) scores at word, sub-word or phone levels. The following section gives an overview of the related work in pronunciation evaluation and mispronunciation detection based on the model used for evaluation, the size of data required and comparison techniques without the use of a corpus.

As noted earlier, pronunciation evaluation is only concerned with producing an assessment score for the learner's speech without necessarily identifying places of mispronunciation. Initial work on pronunciation evaluation made use of GMM-HMM based acoustic models – Gaussian Mixture Models (GMM) were used to determine the state probabilities of the Hidden Markov Model (HMM) that model phonemes (the basic sounds of speech). The confidence scores of HMM prediction is used to assess the pronunciation quality [1], [2], [3]. With advances in deep learning, Deep Neural Network (DNN) based pronunciation evaluation systems have proved more efficient and extensible than GMM-HMM models [4], [5], [6], [7].

The intent of mispronunciation detection is to inform the learner exactly where mispronunciations have occurred. This requires error detection at phone level and is normally done using forced alignment by a speech recognition system to determine phonetic boundaries. These boundaries are needed to indicate which sound (with the label) was wrongly pronounced. Here again HMM [8], [9] and DNN models [10], [5], [11], [12], [13], [14], [15], [16], [17] have been employed. (See [18] for an overview of various approaches to mispronunciation detection.)

Pronunciation evaluation and mispronunciation detection systems normally involve training a speech recognition system with both native (L1) and non-native (L2) speech data [10], [8], [11], [12], [13], [6], [17], [7]. Non-native speech data is difficult to gather, and such data is particularly scarce for many Indian languages including Kannada. Thus, there has been considerable research on minimizing the dependence on large speech corpora for pronunciation training [1], [19], [20], [14], [9]. These approaches do not require L2 data, but they require sufficient L1 data for training.

III. Our Approach

In this paper, we follow the approach taken by Lee and Glass [21], [10], which uses a comparison based pronunciation evaluation that eliminates the need to train a traditional ASR system (and hence eases the requirement of extensive training data). Instead, this approach compares the utterances of a teacher against the utterances of students using Dynamic Time Warping (DTW), and then uses a suitably trained Support Vector Machine (SVM) classifier on phoneme-based and word-based features to evaluate students' speech. Their experiments also involve the TIMIT corpus as canonical pronunciations for 630 speakers and the CU-CHLOE corpus of 100 speakers for evaluations. We perform our experiments with samples of canonical pronunciations from just one speaker, and manually rated samples of 19 additional speakers.

A cross-platform mobile application for pronunciation training has been developed [22] where the back-end is hosted on a cloud. This approach makes use of a traditional ASR system to detect the place of mispronunciations and displays these on the mobile application. We implement a similar cloud based prototype, but we do not employ an ASR system for mispronunciation detection.

We evaluate pronunciations at the word level using a DNN classifier trained against human ratings. Since DNN requires a lot of data and our dataset consists of only 1169 non-native words, following [21] we segment these words and obtain comparison scores for each of these segments. This finer granularity increases the amount of data available to train the DNN classifier. Even so, since an appropriate Kannada speech corpus is not readily available, we do not have sufficient data for forced alignment. Hence, we propose a novel approach using Self Organizing Maps (SOM) [23], an effective unsupervised clustering algorithm to reduce high-dimensional non-linear data into two-dimensional data (a similarity graph) while also preserving topological relationships. Since SOM is efficient in learning continuous data, we experiment by training different SOMs on the canonical pronunciation for each segment and determining the deviations of the non-native speech segments.

TABLE I: Words according to their categories

Sl. No.	Category	**Words**	
		In English	Kannada Pronunciation
1.	Fruits	apple pineapple orange	/se:/ /bu/ /a/ /na:/ /nas/ /ki/ /tta/ /ḷe/
2.	Colors	white purple yellow	/bi/ /ḷi/ /ne:/ /ra/ /ḷe/ /ha/ /ḷa/ /di/
3.	Animals	cat rhinoceros porcupine wolf	/be/ /kku/ /kha/ /ḍga/ /mrū/ /ga/ /mu/ /ḷḷu/ /hã/ /di/ /to:/ /ḷa/
4.	Birds	crow duck kingfisher woodpecker	/ka:/ /ge/ /ba:/ /tu/ ko:/ /ḷi/ /mĩ̄/ /cu/ /ḷḷi/ /ma/ /ra/ /ku/ /ṭi/ /ga/
5.	Flowers	jasmine kanakambara hibiscus rose	/ma/ /lli/ /ge/ /ka/ /na/ /kã:/ /ba/ /ra/ /da:/ /sa/ /va:/ /ḷa/ /gu/ /la:/ /bi/
6.	Numbers	nine seven two	/õ/ /ba/ /ttu/ /e:/ /ḷu/ /e/ /ra/ /ḍu/

TABLE II: Words according to syllables

No. of Syllables	Words
2	6
3	8
4	4
5	3

IV. DataSet and Annotation

For our prototype, we recorded audio samples of 21 Kannada words, spoken slowly and clearly by a native Kannada female speaker. These utterances were used as templates (canonical pronunciations) against which student utterances were evaluated. Since we target beginners, the words we chose are nouns for common objects. These were chosen from first and second grade Kannada textbooks. We chose six categories, with three to four words in each category. Table I lists the categories and the words used, together with their pronunciations in Kannada.

The number of syllables in these words ranges from 2 to 5, and Table II lists the number of words according to syllables. The selected 21 words cover 56 different syllables in the Kannada language, as listed in Table III. In addition to being simple, the words were selected so that learners could practice the pronunciations of all basic vowels in Kannada (/a/, /i/, /u/, /e/ and /o/, including short and long forms). We excluded the dipthongs /ai/ and /ou/ for this prototype. We also ensured that at least one consonant from each of the following groups was included: velar (/ka/), palatal (/ca/), retroflex (/ṭa/), dental (/ta/) and labial (/pa/). We also included the most commonly used nasal sounds (/na/ and /ma/). To keep the prototype simple, the only aspirated consonant we considered is /kha/. The most common mispronunciation in Kannada by non-native speakers is the unstructured consonant ḷ. Thus, we have included different syllable pronunciation forms for ḷ namely, 'ḷa', 'ḷi', 'ḷe' and 'ḷu' in nine words in the dataset across all the categories.

TABLE III: List of Syllables

Sl. No.	Syllables	Sl. No.	Syllables
1	/se:/	29	/ra/
2	/bu/	30	/du/
3	/ka:/	31	/ki/
4	/ge/	32	/tta/
5	/baa/	33	ḷe
6	/tu/	34	/a/
7	/ko:/	35	/na:/
8	/ḷi/	36	/nas/
9	/da:/	37	/mu/
10	/sa/	38	/ḷḷu/
11	/va:/	39	/hã/
12	/ḷa/	40	/di/
13	/ma/	41	/ne:/
14	/lli/	42	/kha/
15	/ka/	43	/ḍga/
16	/na/	44	/mrū/
17	/kã:/	45	/ga/
18	/ba/	46	/gu/
19	/ra/	47	/la:/
20	/mī̃/	48	/bi/
21	/cu/	49	/ḷi/
22	/ḷḷi/	50	/to:/
23	/õ/	51	/ḷa/
24	/ba/	52	/ma/
25	/ttu/	53	/ku/
26	/e:/	54	/ṭi/
27	/ḷu/	55	/hã/
28	/e/	56	/di/

Recordings for the 21 Kannada words were obtained from 19 students aged between 18 and 25 years (both native and non-native Kannada speakers). A total of 1169 audio samples were recorded using our Android application with JBL C100SI, Sony MDR-ex155 Noise cancellation earphones in an environment with minimal background noise. The length of each audio sample is approximately 5 seconds and includes minute variations in pronunciation, which helps us obtain a robust model to detect mispronunciations. These audio samples were given to a Kannada teacher who rated them on a 5-point Likert scale for clear pronunciations. The audio samples were also processed to remove regions containing silence for meaningful analysis.

V. System Design

Our system consists of an Android front-end that connects to a pronunciation evaluation and mispronunciation detection framework deployed on the cloud as shown in Fig. 1. Categories of words are displayed on the screen of the mobile once the user logs in with his or her ID. A sample of the screen in shown in Fig. 2.

TABLE IV: Dataset Summary

Human Rating	Number of Samples
1	40
2	50
3	90
4	243
5	746
Total	1169

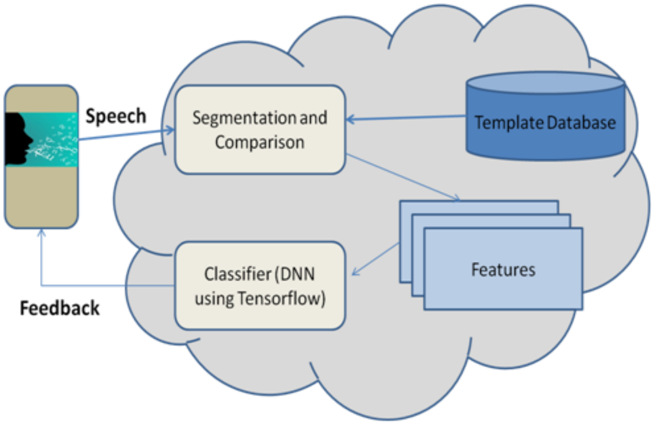

Speech
Segmentation and Comparison
Template Database
Features
Classifier (DNN using Tensorflow)
Feedback

Fig. 1: Comparison Framework

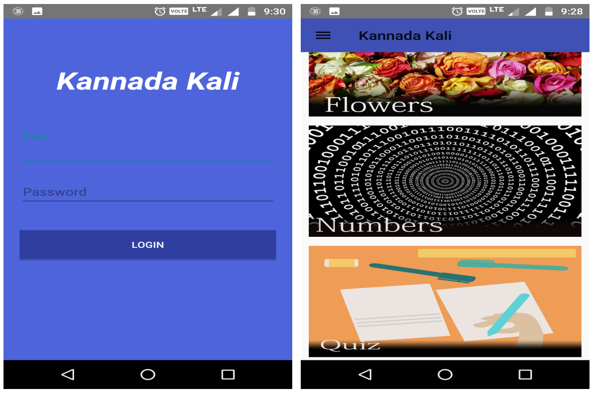


Fig. 2: Android App Screen Shot

A. Pronunciation Rating

Once the user chooses one of the categories, the object corresponding to the Kannada word in the category is displayed as shown in Fig. 3. The user can listen to the template pronunciation by pressing the speaker icon on the bottom left of the screen. He or she can then speak by pressing on the microphone icon on the bottom right of the screen. Once the user releases the button, the audio sample of the user is sent to the back-end on the cloud for evaluation. The rating value is received from the framework and displayed on the mobile screen.

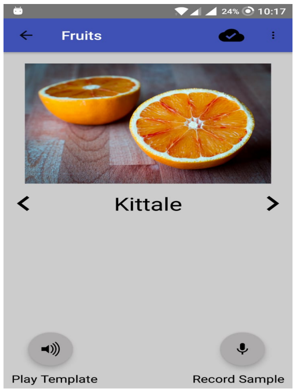


Fig. 3: Word Learning

We use a template-based approach of pronunciation evaluation of spoken Kannada words. We compare the template and the test audio after breaking them into segments for finer granularity. We first perform a spectrogram analysis of the template audio using a self-similarity matrix (SSM) [24] for segmenting the audio. The segments obtained using our algorithm approximately corresponds to syllables in Kannada. After segmentation, we compare the audio segments by implementing the DTW algorithm on the MFCC features of the segments. We compute 14 different values [24] from the DTW cost matrix as listed in Table V. These values represent the differences in the template and the test audio segments.

We train a feed-forward neural network with two hidden layers as shown in Fig. 4. The input to the neural net is the 14 computed values for each segment. The rating assigned by a human teacher for each of the audio samples in the training set is used to train the expected output of the neural net. The rating estimated by the neural net is sent to the Android front-end and displayed on the user screen, as shown in Fig. 5.

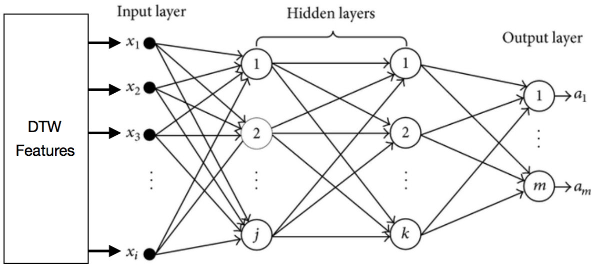


Fig. 4: Feed Forward Neural Network

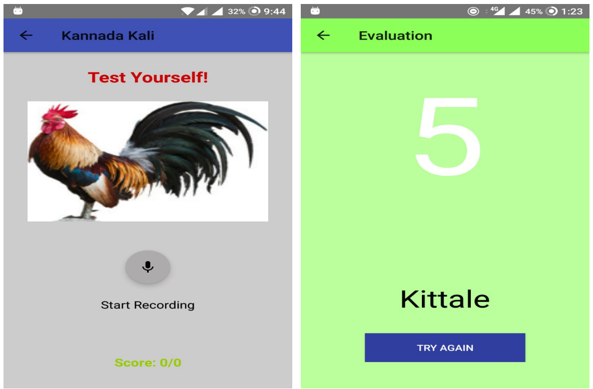


Fig. 5: Pronunciation Evaluation

B. Mispronunciation Detection

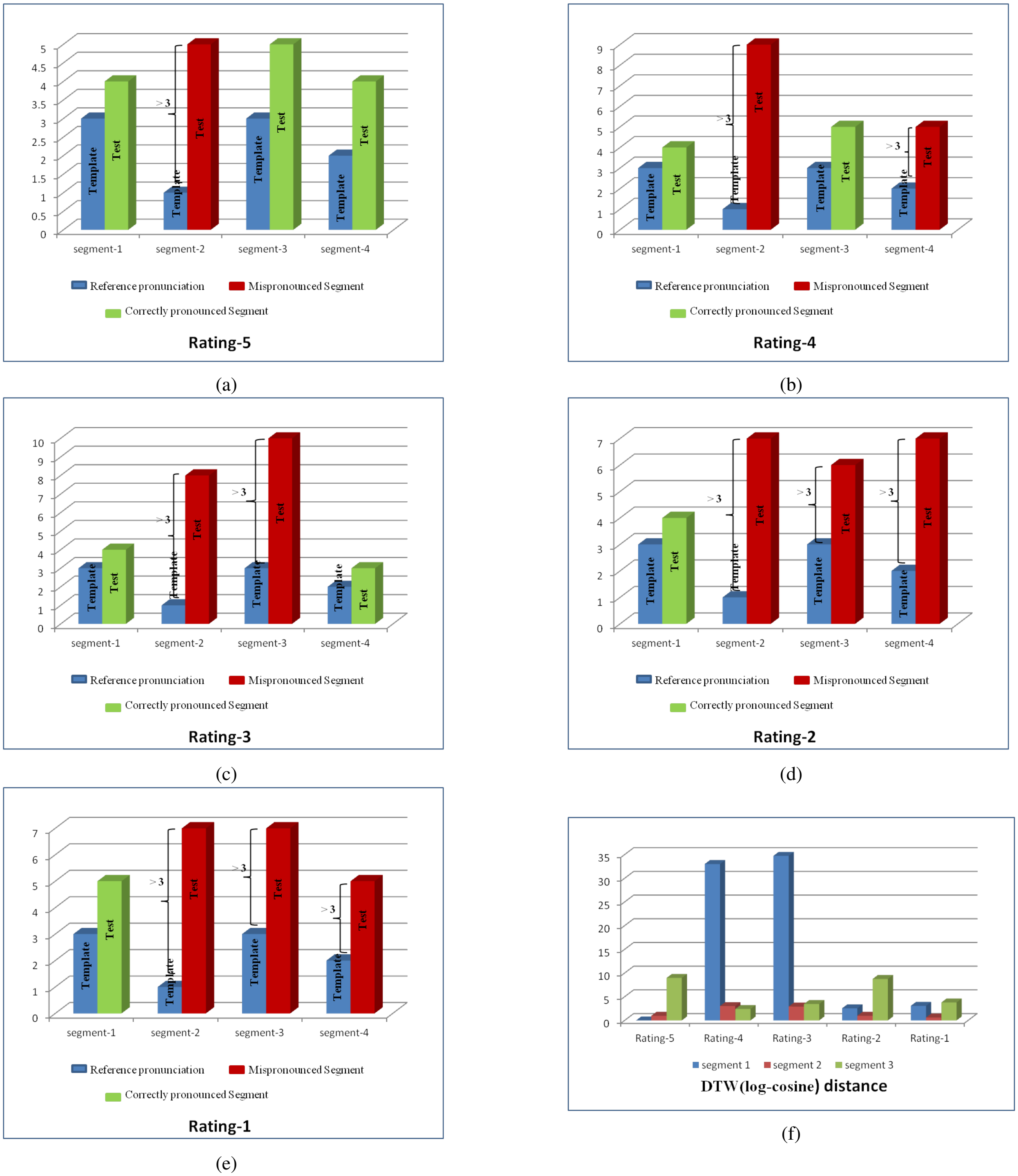
We train a Self Organizing Map (SOM) on the MFCC vector obtained for each segment of the all audio samples, using samples in the training set with rating 5 as a reference. We train different SOMs for different speech segments of rating-5 audio. The predictions of the SOM on the test audio segments are then compared for deviations against the reference using an appropriate distance measure. We determine the number of deviations for the winner nodes predicted by the SOM based on an empirically determined distance threshold. An audio segment is identified as mispronounced if it has more deviations from the reference than this threshold. To provide the user with feedback, we display the picture and the corresponding audio segments. We also play the mispronounced segment followed by its template version (canonical pronunciation) to emphasize the location of mispronunciation.

VI. Results

We consider the teacher rating as a gold standard for evaluating our system. The dataset was divided into training and test sets, with 80% of the audio samples for each of the rating levels comprising the training set and remaining 20% as the test set. We consider both Binary (2 Class) and Multiple (5 Class) classification. For multi-class classification, the classes correspond directly to the human rating on the 5-point Likert scale. For binary classification, the samples with human rating 1 to 3 form class 0 and the remaining samples (rated 4 or 5) form class 1.

Classifier accuracy is obtained by 10-fold cross validation on the dataset taken. Our results are listed in Table VI. Binary classification gives a better accuracy because this grouping results in more data and better balance. Table VII lists the precision, recall and F-score for binary and multi-class classification [25].

Fig. 6 shows the segment deviations for a three syllable word "/ki/ /tta/ /ḷe/", which is divided into four segments. Fig. 6(a) depicts the segment deviations for a test sample that has ha human rating of 5. We obtain a deviation count of 4 for segment-2 which is above the threshold of 3 deviations for the reference audio. Hence, this segment is denoted in red



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 6: Segment Deviations

TABLE V: Computations on DTW Cost Matrix

S. No.	Feature	Computation
1	RATIO_MAX_SEG	max segment length ÷ pathlen
2	AVG_PATH_SCORE	pathscore ÷ pathlen
3	AVG_DIAG_SCORE	diagscore ÷ diaglen
4	DIFF_AVG_DIAG_PATH_SCORE	avg diag score - avg path score
5	RATIO_AVG_DIAG_PATH_SCORE	avg diag score ÷ avg path score
6	MEAN_SCORE	MFCC mean score for a segment
7	RATIO_ABS_DURATION	max((width ÷ height),(height ÷ width))
8	DIFF_REL_DURATION	abs(relative width - relative height)
9	RATIO_REL_DURATION	max((relative width ÷ relative height),(relative height ÷ relative width)
10	DIFF_MEAN_SCORE	mean score - mean ref scores
11	DIFF_AVG_PATH_REF_SCORE	avg path score - avg path ref scores
12	DIFF_AVG_DIAG_REF_SCORE	avg diag score - avg path ref scores
13	DIFF_AVG_MFCC	avgTemplateMfcc - avgTestMfcc (complete audio)
14	RATIO_AVG_MFCC	avgTemplateMfcc ÷ avgTestMfcc (complete audio)

TABLE VI: Classification Accuracy

Classification	Cross-validation Accuracy	Train Accuracy	Test Accuracy
Binary	92%	93%	86%
5-Class	75%	74%	68%

TABLE VII: Precision, Recall and F-score

Classification	Precision	Recall	F-score
Binary	0.84	0.89	0.86
5-Class	0.63	0.87	0.73

to indicate a mispronunciation. Similarly, Fig 6(b) has two mispronounced segments indicated in red for a sample with human rating 4 (deviation counts of 8 for segment-2 and 3 for segment-4). Fig. 6(c) also has two deviations beyond the count of 3, but the deviation count is 7 for segment-2 and segment-3 which is more than the total deviation count for the audio sample. This is in agreement with the human rating. From Fig. 6(d) and Fig. 6(e), we observe that there are three segments that have a deviation count greater than 3. We note that the deviation counts of Rating-2 samples are more than the deviations for Rating-1 samples. This may be due to the fact that pronunciations in both the audio samples are poor, and human rating is subjective. Hence, using an unsupervised algorithm such as SOM will help evaluate the pronunciations in an objective manner. Fig. 6(f) shows the DTW distance values obtained for the corresponding audio samples using log-cosine distance. Here the distances for Rating-4 and Rating-3 audio are more than Rating-2 and Rating-1 audio samples, as expected. Hence, SOM seems to be more effective is determining mispronunciations than DTW distances.

VII. Conclusions and future work

In this paper we have proposed a mobile application to assist L2 speakers learn pronunciations in Kannada using an evaluation framework deployed on the cloud. Our initial results show that pronunciation evaluation can be achieved through segmentation, deep learning for classification and semi-supervised learning (an unsupervised learning tool trained on only canonical pronunciations) for mispronunciation detection with minimal reference audio. This eliminates the need for a speech recognition system that requires significant amounts of data to achieve good accuracy. We believe that our approach can be extended to other languages, which is part of our ongoing efforts.

Mispronunciation detection has also been employed for pronunciation training in children [26], [27]. Our classification framework for children of grades 1 and 2 for the same set of 21 Kannada words resulted in a classification accuracy of 58% for binary and 38% for multi-class approximately. In future, we intend to generalize the framework to all age groups and genders by preprocessing the audio samples. The feedback mechanism can also be improved by emphasizing mispronounced audio segments [28], [29].

Our dataset is freely available for download at `github.com/anandankit95/Kannada-Kali`. With the necessary annotation (e.g., through crowd sourcing), this dataset can also be used for voice identification and accent recognition. This application can be further extended to evaluate spoken sentences for paragraphs, to evaluate fluency in pronunciations. Finally, this application can be a base framework for language learning with the inclusion of language specific grammar (specified as language model).

References

[1] C. Bhat, K. L. Srinivas, and P. Rao, "Pronunciation scoring for Indian English learners using a phone recognition system," *Proceedings of the First International Conference on Intelligent Interactive Technologies and Multimedia - IITM '10*, pp. 135–139, 2010. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1963564.1963587

[2] X. Li, J. Chen, M. Yao, D. Shen, and F. Lin, "English sentence pronunciation evaluation using rhythm and intonation," *2014 2nd International Conference on Systems and Informatics, ICSAI 2014*, no. Icsai, pp. 366–371, 2015.

[3] V. Laborde, T. Pellegrini, L. Fontan, J. Mauclair, H. Sahraoui, and J. Farinas, “Pronunciation assessment of Japanese learners of French with GOP scores and phonetic information,” *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 08-12-Sept, pp. 2686–2690, 2016.

[4] W. Hu, Y. Qian, and F. K. Soong, “A new DNN-based high quality pronunciation evaluation for computer-aided language learning (call),” *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, no. August, pp. 1886–1890, 2013.

[5] X. G. Li, S. M. Li, L. R. Jiang, and S. B. Zhang, “A multi-parameter objective evaluation system for English sentence pronunciation,” *Proceedings of the 2013 6th International Congress on Image and Signal Processing, CISP 2013*, vol. 3, no. Cisp, pp. 1292–1297, 2013.

[6] J. Lin, Y. Xie, and J. Zhang, “Automatic pronunciation evaluation of non-native Mandarin tone by using multi-level confidence measures,” *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 08-12-Sept, pp. 2666–2670, 2016.

[7] K. Kyriakopoulos, K. M. Knill, and M. J. F. Gales, “A deep learning approach to assessing non-native pronunciation of English using phone distances,” no. September, pp. 1626–1630, 2018.

[8] A. Al Hindi, M. Alsulaiman, G. Muhammad, and S. Al-Kahtani, “Automatic pronunciation error detection of nonnative Arabic Speech,” *Proceedings of IEEE/ACS International Conference on Computer Systems and Applications, AICCSA*, vol. 2014, pp. 190–197, 2014.

[9] X. Qian, H. Meng, and F. Soong, “A two-pass framework of mispronunciation detection & diagnosis for computer-aided pronunciation training,” *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2015*, no. December, pp. 384–387, 2016.

[10] A. Lee, Y. Zhang, and J. Glass, “Mispronunciation detection via dynamic time warping on deep believe network-based posteriorgrams,” *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8227–8231, 2013.

[11] W. Hu, Y. Qian, and F. K. Soong, “A DNN-based acoustic modeling of tonal language and its application to Mandarin pronunciation training,” *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 3206–3210, 2014.

[12] W. Hu, Y. Qian, F. K. Soong, and Y. Wang, “Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers,” *Speech Communication*, vol. 67, pp. 154–166, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2014.12.008

[13] Y. C. Hsu, M. H. Yang, H. T. Hung, and B. Chen, “Mispronunciation detection leveraging maximum performance criterion training of acoustic models and decision functions,” *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 08-12-Sept, no. 2, pp. 2646–2650, 2016.

[14] R. Duan, T. Kawahara, M. Dantsuji, and J. Zhang, “EFFECTIVE ARTICULATORY MODELING FOR PRONUNCIATION ERROR DETECTION OF L2 LEARNER WITHOUT NON-NATIVE TRAINING DATA Academic Center for Computing and Media Studies , Kyoto University , Japan School of Information Science , Beijing Language and Culture Univer,” pp. 5815–5819, 2017.

[15] V. Arora, A. Lahiri, and H. Reetz, “Phonological feature based mispronunciation detection and diagnosis using multi-Task DNNs and Active Learning,” *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2017-Augus, pp. 1432–1436, 2017.

[16] K. Li, H. Meng, T. Chinese, H. Kong, and H. K. Sar, “Mispronunciation Detection and Diagnosis in L2 English Speech Using Multi - Distribution Deep Neural Networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 1, pp. 193–207, 2017.

[17] W. Li, N. F. Chen, S. M. Siniscalchi, and C. H. Lee, “Improving mispronunciation detection for non-native learners with multisource information and LSTM-based deep models,” *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2017-Augus, pp. 2759–2763, 2017.

[18] N. F. Chen and H. Li, “Computer-assisted pronunciation training: From pronunciation scoring towards spoken language learning,” *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2016*, 2017.

[19] A. Lee and J. Glass, “Mispronunciation Detection without Nonnative Training Data,” vol. 1, pp. 643–647, 2015.

[20] A. Lee, N. F. Chen, and J. Glass, “PERSONALIZED MISPRONUNCIATION DETECTION AND DIAGNOSIS,” *Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6145–6149, 2016.

[21] A. Lee and J. Glass, “A Comparison Based Approach to Mispronunciation Detection,” pp. 1–92, 2012.

[22] P. Liu, K.-w. Yuen, W.-k. Leung, and H. M. Meng, “mENUNCIATE : DEVELOPMENT OF A COMPUTER-AIDED PRONUNCIATION TRAINING SYSTEM ON A CROSS-PLATFORM FRAMEWORK FOR MOBILE , SPEECH-ENABLED APPLICATION DEVELOPMENT Pengfei Liu , Ka-Wa Yuen , Wai-Kim Leung and Helen Meng Department of Systems Engineering and Engi,” *Chinese Spoken Language Processing (ISCSLP)*, pp. 170–173, 2012.

[23] T. Kohonen, *Self Organizing Maps_Kohonen*, 3rd ed. Springer, 2001.

[24] A. Lee and J. Glass, “Pronunciation Assessment via a Comparison-based System,” *SLaTE*, pp. 122–126, 2013.

[25] M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” *Information Processing and Management*, vol. 45, no. 4, pp. 427–437, 2009. [Online]. Available: http://dx.doi.org/10.1016/j.ipm.2009.03.002

[26] R. Tong, N. F. Chen, and B. Ma, “Multi-Task Learning for Mispronunciation Detection on Singapore Children ’ s Mandarin Speech,” *Interspeech*, pp. 2193–2197, 2017.

[27] J. Proença, C. Lopes, M. Tjalve, A. Stolcke, S. Candeias, and F. Perdigão, “Detection of mispronunciations and disfluencies in children reading aloud,” *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2017-Augus, pp. 1437–1441, 2017.

[28] S. Zhao, J., Yuan, H., Leung, W. K., Meng, H., Liu, J., & Xia, “AUDIOVISUAL SYNTHESIS OF EXAGGERATED SPEECH FOR CORRECTIVE FEEDBACK IN COMPUTER-ASSISTED PRONUNCIATION TRAINING University of Chinese Academy of Sciences , Beijing 100190 , China TNList , Department of Electronic Engineering , Tsinghua University , China,” *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[29] L. Ning, Y., Wu, Z., Jia, J., Meng, F., Meng, H., & Cai, “HMM-BASED EMPHATIC SPEECH SYNTHESIS FOR CORRECTIVE FEEDBACK IN COMPUTER-AIDED PRONUNCIATION TRAINING Tsinghua-CUHK Joint Research Center for Media Sciences , Technologies and Systems , Shenzhen Key Laboratory of Information Science and Technology , Gradua,” *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 4934–4938, 2015.