# Multi-Task Learning for Speech Attribute Detection of Children's Speech

*Mostafa Shahin, Beena Ahmed, Julien Epps*

School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, Australia

mostafa_shahin@ieee.org, beena.ahmed@unsw.edu.au, j.epps@unsw.edu.au

## Abstract

Speech attributes, including manners and places of articulation, provide a detailed description of sound production directly related to the speech articulators. An accurate modeling of such attributes enriches automatic pronunciation assessment applications by providing informative feedback to the user. In this paper we propose a DNN based classification model structure to automatically detect the absence or existence of 25 specific attributes in English children's speech. In the models, we utilised multi-task learning (MTL) with frame-level phoneme classification as an auxiliary task and a discriminative additive (DA) task for highly confusable phonemes to improve the detection of attributes. We compared the performance of the 25 DNN-MTL and DNN-MTL-DA attribute detection models across all the phonemes to base DNN models in two different children corpora to determine the impact of MTL and the DA task. We also compared the performance of our DNN-MTL-DA across different children's age groups. Our attribute models achieved detection accuracies ranging from ~80% to ~91%, with the best detection accuracy for *Nasal* and *retroflex* and worst for *Tense* and *affricates*.

**Index Terms**: speech attributes, children speech, multi-task learning

## 1. Introduction

Unlike acoustic modeling for automatic speech recognition (ASR) applications where an abstract model that can handle all variations of the same word is desirable, pronunciation assessment applications such as L2 learning, speech therapy and language proficiency tests, require the accurate detection of any deviation from standard pronunciation.

Speech attributes, such as manners and places of articulations, provide a low-level description of sound production in terms of which articulators are involved and how these articulators move to produce a specific sound. Any alteration in these attributes causes a pronunciation error. Therefore, an accurate modeling of these attributes can pave the way to a fully automated and interactive pronunciation assessment application where the learner receives formative and diagnostic automatic feedback not only about the existence of incorrect pronunciation but also how the error is made. Furthermore, modeling speech attributes can be performed solely using standard pronunciation datasets which are abundantly available compared to non-standard datasets. Additionally, speech attributes are common among most spoken languages enabling modelling with speech corpora from multiple languages.

Speech attributes have been successfully utilized in various domains. In the context of ASR, Lee et al. proposed a bottom-up speech recognition approach based on a bank of speech attribute detectors [1]. In [2] authors used speech attribute-based model to rescore a word-lattice generated by Gaussian mixture model (GMM) and deep neural network (DNN) based ASR system. [3] introduced an i-vector representation of manners and places of articulation attributes for the detection of foreign accent speech in Finnish and English languages. The system achieved a 15% error reduction when compared to a spectrum-based technique.

In the pronunciation assessment domain, [4] used scores derived from speech attribute models to assess the pronunciation quality of L2 Mandarin learners. The system outperforms phoneme-based goodness of pronunciation technique (GOP) by reducing the equal error rate by ~9% relative. In [5], authors proposed an anomaly detection-based pronunciation verification system by training phoneme-specific one-class support-vector machine (OCSVM) model using speech attribute features. However, most previous work has focused on speech attributes in adult speech. Children's speech poses an additional challenge due to its high inter- and intra-speaker variability compared to adult speech.

In this paper we propose a deep learning approach to accurately detect 25 speech attributes including manners and places of articulations in English children's speech. In this approach we used 25 separate binary output DNN-based models to detect the absence or existence of each attribute trained two public speech corpora including speech from children aged 5 to 15 years old. In the models, we used phoneme classification as a secondary task in a multi-task structure (DNN-MTL) and additional discriminative tasks (DNN-MTL-DA) between confusable phonemes to improve the performance of the speech attribute detectors. We conducted experiments to validate the performance of the DNN-MTL models in detecting all 25 attributes across all phonemes and over different age groups.

The rest of the paper is organized as follow. The method and the speech corpora used are explained in Section 2. Section 3 represents the experiments and the analysis of the results. Finally, conclusions are drawn in Section 4.

## 2. Method

### 2.1. Speech Corpora

Two publicly available speech corpora were utilized in this work, the Oregon Graduate Institute (OGI) kids' speech corpus [6] and Colorado University (CU) Kid's prompted, read and summarized speech corpus [7, 8].

Table 1. *The distribution of the speech corpora.*

| | OGI | | | CU | | |
|---|---|---|---|---|---|---|
| | **Train** | **Valid** | **Test** | **Train** | **Valid** | **Test** |
| Speakers | 794 | 162 | 162 | 716 | 107 | 94 |
| Segments | ~42k | ~8.5k | ~8.5k | ~66k | ~9k | ~8.3k |
| Hours | 59.4 | 7.7 | 7.6 | 59 | 7.3 | 7.6 |

The OGI corpus contains 11 age groups from kindergarten to grade 10 while the CU corpus consists of 6 age groups from kindergarten to grade 5. Each dataset was split into three subsets for training, validation and testing as demonstrated in Table 1. The data was split to ensure that each subset included all age ranges.

In our experiments we used each speech corpus separately to show the robustness of speech attribute detection over different domains. The OGI corpus includes recordings from across a wider age range while the CU corpus has a larger vocabulary and more spontaneous speech.

### 2.2. Speech Attributes Detectors

We adopted 25 speech attributes representing mainly the manners and places of articulations in addition to other phonetic characteristics such as voiceless, roundness, tenseness, etc. Table 2 shows the list of speech attributes along with their associated phonemes in ARPAbet format [9].

Table 2. *List of speech attributes.*

| | Attribute | Phonemes |
|---|---|---|
| Manners | Vowel | iy ih eh ey ae aa aw ah ao oy ow uh uw er |
| | Semivowel | y w |
| | Fricative | jh ch s sh z zh f th v dh hh |
| | Nasal | m n ng |
| | Stop | b d g p t k |
| | Approximant | w y l r |
| | Affricates | ch jh |
| Places | Coronal | d l n s t z |
| | High | ch ih iy jh sh uh uw y ow g k ng |
| | Dental | dh th |
| | Glottal | hh |
| | Labial | b f m p v w |
| | Low | aa ae aw ay oy |
| | Mid | ah eh ey ow |
| | Retroflex | er r |
| | Velar | g k ng |
| Others | Anterior | b d dh f l m n p s t th v z w |
| | Back | ay aa ah ao aw ow oy uh uw g k |
| | Continuant | aa ae ah ao aw ay dh eh er r ey l f ih iy oy ow s sh th uh uw v w y z |
| | Round | aw ow uw ao uh v y oy r w |
| | Tense | aa ae ao aw ay ey iy ow oy uw ch s sh f th p t k hh |
| | Voiced | aa ae ah aw ay ao b d dh eh er ey g ih iy jh l m n ng ow oy r uh uw v w y z |
| | Liquids | l r |
| | Monophthong | ao aa iy uw eh ih uh ah ae |
| | Diphthong | ey ay ow aw oy |

For each attribute, a binary fully connected DNN model was trained to classify each frame as +ve or -ve when the underlying attribute was detected or not respectively. We further used a frame-level phoneme classification auxiliary task to improve the generalization of the main attribute detection task in a MTL scenario. In some specific attributes, additional task/s were added to force the network to discriminate between pairs of highly confused phonemes better that were described by attributes in opposite classes (+ve/-ve). The confused phonemes were determined from a phoneme confusion matrix computed using a frame-level phoneme classification model. The MTL model architecture is depicted in Figure 1.

The cross-entropy function was used to compute the loss in the attribute output as well as outputs of all other tasks. The total network loss was computed as follows:

$$L_A(\theta) = -\sum_i \log P(y_A^{(i)}/x^{(i)};\theta) \quad (1)$$

$$L_P(\theta) = -\sum_i \log P(y_P^{(i)}/x^{(i)};\theta) \quad (2)$$

$$L_D(\theta) = -\sum_i log\, P(y_D^{(i)}/x^{(i)};\theta) \quad (3)$$

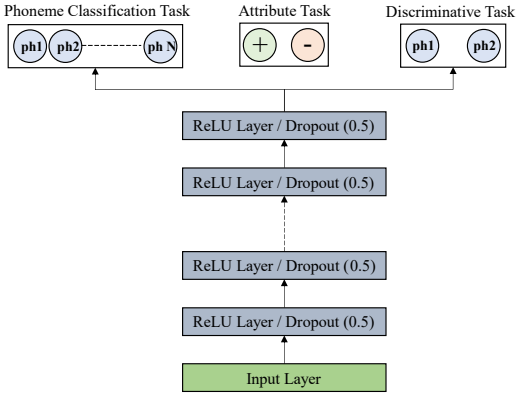


Figure 1: *Multi-Task learning model architecture.*

Where $L_A, L_P$ and $L_D$ are the losses computed from the attribute, phoneme classification and discriminative tasks respectively, while $y_A^{(i)}, y_P^{(i)}$ and $y_D^{(i)}$ are labels of sample $x^{(i)}$ for the three outputs. The total loss is a weighted sum of the three losses as follow:

$$L(\theta) = L_A(\theta) + \lambda_P L_P(\theta) + \lambda_D L_D(\theta) \quad (4)$$

Where $\lambda_P$ and $\lambda_D$ are the weights of the losses at the phoneme classification and the discriminative tasks, respectively.

From each frame, 26 filter banks were extracted along with their delta and acceleration components. N consecutive frames were spliced together to form the final feature vector fed to the DNN model. The number of layers, the number of units per layer, initial learning rate and the N spliced frames were empirically determined. The rectified linear unit (ReLU) was used as an activation for the hidden units while the softmax used for the output units of all tasks. The training ran for a maximum of 50 epochs with early stopping by monitoring the validation loss of the attribute detection task. Moreover, the dropout regularization of 0.5 was used to alleviate the overfitting effect and the update of the parameter was optimized using the Adam technique.

In the training of each speech attribute model, frames from phonemes belonging to the underlying attribute were labeled as +ve while all other phonemes were labeled as -ve. To avoid the model being biased, we balanced the data by selecting an equal number of samples in each class. We ensured that samples were selected from all phonemes of +ve and -ve groups. The time boundaries of the phonemes were obtained by forced alignment using a children's ASR system [10].

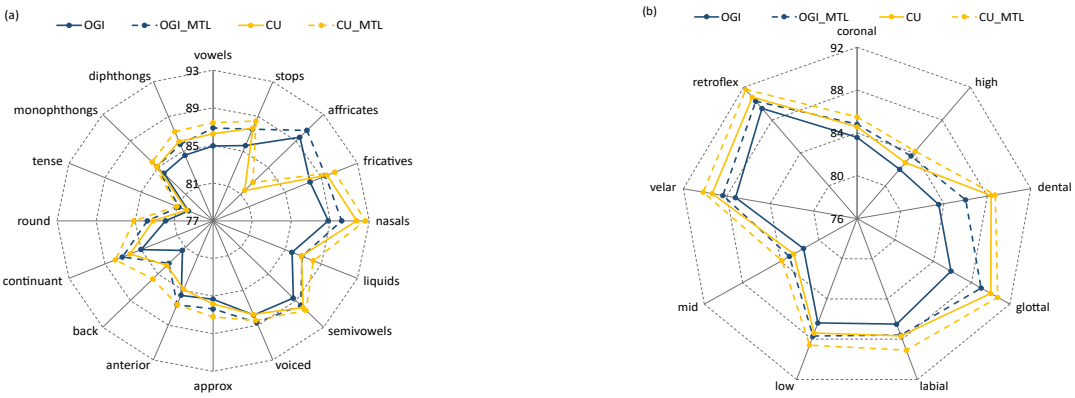


Figure 2: *Detection accuracy of (a) manners of articulations and other phonetic attributes and (b) places of articulation attributes for OGI and CU datasets.*

## 3. Results

The baseline DNN model consists of one output layer for attribute detection, and 5 hidden layers with 1024 ReLU hidden units each. Features from 5 frames were spliced to form the input feature vector. This baseline architecture is similar to the one used in [11] for adult speech attribute detections of the Wall Street Journal (WSJ) speech corpus [12]. Table 3 shows the classification accuracy of the baseline DNN models when trained and tested against OGI speech corpora and CU speech corpora. The last column shows the classification accuracy of a similar DNN model applied to the adult WSJ speech corpus obtained from [11].

Table 3. *Classification accuracy of different speech attribute detectors trained and tested using children's speech corpora (CU and OGI) and adult speech corpus (WSJ).*

| Attribute | OGI | CU | WSJ [11] |
|---|---|---|---|
| **anterior** | 85.6 | 84.9 | 92.5 |
| **approximant** | 85.3 | 85.8 | 96.4 |
| **back** | 81.5 | 83.7 | 93.1 |
| **continuant** | 85.0 | 86.2 | 93.5 |
| **coronal** | 83.5 | 84.6 | 92.4 |
| **dental** | 83.6 | 88.4 | 99.0 |
| **fricative** | 87.7 | 89.6 | 96.2 |
| **glottal** | 85.8 | 90.0 | 99.7 |
| **high** | 82.0 | 82.8 | 95.0 |
| **labial** | 86.5 | 87.7 | 96.9 |
| **low** | 86.4 | 87.3 | 96.9 |
| **mid** | 81.6 | 82.6 | 93.8 |
| **nasal** | 88.8 | 91.7 | 97.7 |
| **retroflex** | 89.4 | 90.8 | 98.5 |
| **round** | 81.9 | 83.1 | 94.9 |
| **stop** | 85.6 | 87.6 | 95.7 |
| **tense** | 80.7 | 81.0 | 90.6 |
| **velar** | 87.2 | 89.3 | 98.7 |
| **voiced** | 87.8 | 87.7 | 95.3 |
| **vowel** | 85.0 | 86.2 | 92.8 |
| **Average** | 85±2.5 | 86.6±2.9 | 95.5±2.5 |

A separate model was trained for the detection of each attribute. Approximately 15 hours of speech from the WSJ corpus was used to training the adult speech attribute detection models [11]. Despite training the CU and OGI speech attribute detection models with almost 4 times as much training data, the classification accuracies of both children's datasets over all the speech attributes were much lower than the corresponding accuracies achieved using models for the adult WSJ dataset as shown in Table 3. The *tense* attribute was shown to be the most difficult attribute to be automatically detected for both adult and children. The most accurate attributes in adult are *glottal* followed by *dental* and *velar* while in children, *nasal* achieved the highest detection accuracy followed by *retroflex* and *fricative*. These results demonstrate the increased difficulties of detecting speech attributes in children's speech compared to adult speech.

Figure 2 depicts the results on both OGI and CU datasets using the baseline DNN model and the proposed DNN-MTL model. The best parameters of the MTL model were 5 hidden layers with 2048 hidden units each, 0.001 initial learning rate, and 5 spliced windows.

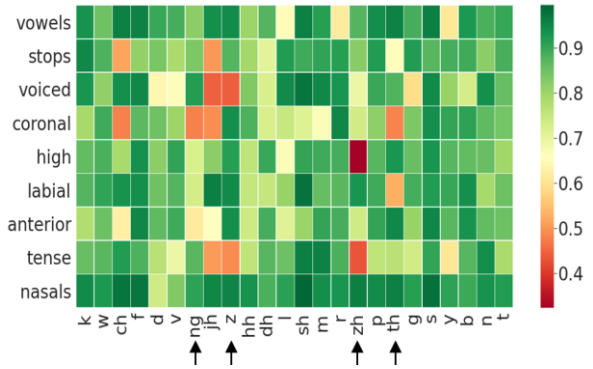


Figure 3: *Attribute detection accuracy over consonant phonemes for CU dataset. /zh/ has very low accuracies in tense and high attributes and high accuracies in nasals and labial. Similar behavior for /ng/ in coronal compared to voiced, and /th/ in labial and tense compared to vowel.*

Overall, results with the CU dataset are better than with OGI. Although both models were trained with almost the same amount of data, the CU dataset has more variations in vocabulary and more continuous speech compared to OGI. As shown in the figure, the detection accuracy ranges from 80% to almost 91% for both datasets. *Nasal*, *retroflex* and *semivowel* attributes achieved the highest detection accuracy of around 91% followed by *glottal*, *fricatives* and *velar* with nearly 90% correct detection rate. *Tense* has the lowest detection accuracy of ~80% followed by *affricates*, mid and high of ~82%.

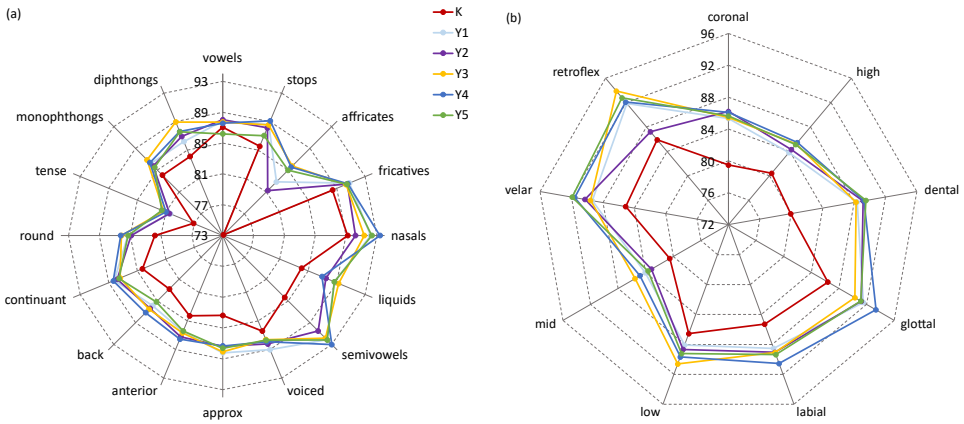


Figure 4: *Distribution of detection accuracy of (a) manners of articulations and other phonetic attributes and (b) places of articulation attributes of CU dataset over different age group from Kindergarten(K) to year 5 (Y5).*

It is obvious from the figure that using the phoneme classification auxiliary task improved the performance of all speech attributes in both CU and OGI datasets with an absolute increase in the accuracy between 0.5% to as high as ~3%. The detection error of the OGI glottal attribute decreased from 14% to around 11% while the CU back attribute detection error decreased from ~16% to ~14%.

We further broke down the results of each attribute by all +ve and -ve phonemes for the CU dataset. Figure 3 shows a heatmap of a subset of this breakdown focusing on the consonants. It is noticeable from the figure that the performance is inconsistence over different phonemes. For instance, /zh/ suffers from low classification accuracy of ~40% and ~35% in *tense* and *high* attributes respectively while *nasals* and *labial* obtained ~97% and ~93% respectively. To better understand this behaviour, we computed the frame-level phoneme confusion matrix using a phoneme classification DNN model trained on the same CU dataset.

Table 4 lists the most confused phonemes for consonants that obtained low attribute classification accuracy as indicated in Figure 3. We concluded that when two highly confused phonemes are in opposite sides of specific attributes (i.e one +ve phoneme and the other one -ve phoneme), the accuracy of one or both degraded significantly. For example, /ng/ has a high confusion with /n/, in *labial* both /n/ and /ng/ are positive samples and hence they achieved a high accuracy > 90%. On the other hand, in *coronal* and *anterior*, where /ng/ is -ve and /n/ is +ve, the accuracy of /ng/ degraded to ~49% and ~62%, respectively. The same behaviour was present with /th/ and /f/ in *labial* and *coronal* and /z/ and /s/ in *tense* and *voiced*.

Table 4. *Phoneme confusion*

| Phoneme | Confused phonemes (misclassification rate) |
|---|---|
| /zh/ | /sh/ (37%), /jh/ (7%) |
| /z/ | /s/ (38%), /t/ (2%) |
| /ch/ | /sh/ (23%), /t/ (22%) |
| /ng/ | /n/ (34%), m (4%) |
| /jh/ | /t/ (17%), ch (10%), sh (10%) |
| /th/ | /s/ (19%), f (15%), t (7%) |

To improve the performance of the model in these specific phonemes, we investigated the use of a discriminative task in the learning of the attribute detection model to guide the model, so it discriminates better between highly confused phonemes. In the training of the anterior detector, we added an additional output to discriminate between /ng/ and /n/ achieving ~8% absolute increase in the *anterior* attribute detection of /ng/ frames from ~62% to ~70%. The same architecture was used in training the *labial* detector; adding a /th/ and /f/ discrimination output, also improved the /th/ accuracy from ~66% to ~71%.

Figure 4 represents a further analysis of the results over 6 age groups from the CU dataset from Kindergarten (K) to Year 5 (Y5). Except for *vowels*, Kindergarten children show a low detection accuracy for almost all other attributes with the highest degradation notice in *affricates*. This can be explained by the rapid change in the articulators between younger and older children and the development of speech sounds. Generally, *vowels* are among the first phonemes to be mastered by children at very young ages. On the contrary, a pronunciation error of the *affricate* sounds known as de-affrication is common in children till 5 years old. This is where the child tends to replace affricates like /ch/ and /jh/ with fricative or stop like /sh/ or /d/.

## 4. Conclusions

In this work, 25 binary DNN-based models were trained to detect different speech attributes, including manners and places of articulation, in children's English speech. Two publicly available datasets were employed, namely CU and OGI corpora, for the training and evaluation of the models.

The speech attribute models achieved detection accuracies ranging from ~80% up to 91% where *nasal*, *retroflex* and semivowels obtained the highest accuracy. A frame-level phoneme classification auxiliary task was further explored to improve the generalization of the speech attribute models attaining ~20% average reduction in the detection error.

A phoneme-level breakdown of the detection error shows a significant increase in the error rate in highly confused phonemes when belonging to different classes of specific attribute. An addition discriminative task was added to the model architecture to alleviate this effect by discriminating between confused phonemes. When applied to /th/ and /f/ in the detection of *labial* attribute and to /ng/ and /n/ in the detection of anterior, the method achieved a 21% and 14% reduction in the classification error of /ng/ and /th/ respectively.

A further age group performance analysis was conducted demonstrating an obvious degradation in the detection accuracy at the young age group (Kindergarten) for most of the speech attributes except for *vowels* where the performance was consistent amongst all age groups.

## 5. References


[1] C.-H. Lee, and S. M. Siniscalchi, “An information-extraction approach to speech processing: Analysis, detection, verification, and recognition,” *Proceedings of the IEEE,* vol. 101, no. 5, pp. 1089-1115, 2013.

[2] I.-F. Chen, S. M. Siniscalchi, and C.-H. Lee, "Attribute based lattice rescoring in spontaneous speech recognition." pp. 3325-3329.

[3] H. Behravan, V. Hautamäki, S. M. Siniscalchi, T. Kinnunen, and C. Lee, “i-Vector Modeling of Speech Attributes for Automatic Foreign Accent Recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 24, no. 1, pp. 29-41, 2016.

[4] W. Li, S. M. Siniscalchi, N. F. Chen, and C. Lee, "Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling." pp. 6135-6139.

[5] M. Shahin, and B. Ahmed, “Anomaly detection based pronunciation verification approach using speech attribute features,” *Speech Communication,* vol. 111, pp. 29-43, 2019.

[6] K. Shobaki, J.-P. Hosom, and R. A. Cole, "The OGI kids' speech corpus and recognizers."

[7] R. Cole, P. Hosom, and B. Pellom, *University of colorado prompted and read childrens speech corpus*, Center for Spoken Language Research, University of Colorado, Boulder, 2006.

[8] R. Cole, and B. Pellom, *University of colorado read and summarized story corpus*, Center for Spoken Language Research, University of Colorado, Boulder, 2006.

[9] A. Klautau. "ARPABET and the TIMIT alphabet," https://web.archive.org/web/20160603180727/http://www.laps.ufpa.br/aldebaro/papers/ak_arpabet01.pdf.

[10] M. Shahin, R. Lu, J. Epps, and B. Ahmed, "UNSW System Description for the Shared Task on Automatic Speech Recognition for Non-Native Children’s Speech."

[11] D. Yu, S. M. Siniscalchi, L. Deng, and C.-H. Lee, "Boosting attribute and phone estimation accuracies with deep neural networks for detection-based speech recognition." pp. 4169-4172.

[12] D. B. Paul, and J. Baker, "The design for the Wall Street Journal-based CSR corpus."