JASA ARTICLE

Phone duration modeling for speaker age estimation in children

Prashanth Gurunath Shivakumar,[1,a] Somer Bishop,[2] Catherine Lord,[3] and Shrikanth Narayanan[1]
[1]*Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, California 90089, USA*
[2]*Department of Psychiatry, University of California, San Francisco, California 94143, USA*
[3]*Semel Institute of Neuroscience and Human Behavior, University of California, Los Angeles, California 90095, USA*

ABSTRACT:
Automatic inference of paralinguistic information from speech, such as age, is an important area of research with many technological applications. Speaker age estimation can help with age-appropriate curation of information content and personalized interactive experiences. However, automatic speaker age estimation in children is challenging due to the paucity of speech data representing the developmental spectrum, and the large signal variability including within a given age group. Most prior approaches in child speaker age estimation adopt methods directly drawn from research on adult speech. In this paper, we propose a novel technique that exploits temporal variability present in children's speech for estimation of children's age. We focus on phone durations as biomarker of children's age. Phone duration distributions are derived by forced-aligning children's speech with transcripts. Regression models are trained to predict speaker age among children studying in kindergarten up to grade 10. Experiments on two children's speech datasets are used to demonstrate the robustness and portability of proposed features over multiple domains of varying signal conditions. Phonemes contributing most to estimation of children speaker age are analyzed and presented. Experimental results suggest phone durations contain important development-related information of children. The proposed features are also suited for application under low data scenarios.
 https://doi.org/10.1121/10.0015198
(Received 27 February 2022; revised 21 October 2022; accepted 29 October 2022; published online 15 November 2022)
[Editor: John H. L. Hansen] Pages: 3000–3009

I. INTRODUCTION

Speech contains important paralinguistic information including speaker's age, gender, emotions, and other behavior constructs (Schuller *et al.*, 2013). Inference of age and gender from children's speech can help better tailor conversational interfaces such as education and learning platforms, entertainment, interactive gaming, tutoring, and social networking for different age-gender demographics. Speech-based biomarkers are also increasingly used in supporting health applications (Bone *et al.*, 2017b), including developmental disorders (Bone *et al.*, 2017a).

Knowledge of an individual's age is important meta-data for several applications. Automatic recognition of speaker age is especially valuable when speech is the only form of data available. Age information can facilitate adaptive information sharing and better personalization, including ensuring privacy and security, thereby enhancing the experiences supported by the speech technology applications. Arguably, child-centred applications can benefit by using paralinguistic information for safeguarding children and enforcing age appropriate content and interactive experiences.

Most of the past research in speaker age estimation is based on adult speech. Earlier research involved training classifiers on statistical functionals of speech descriptors such as loudness, pitch, jitter, shimmer, and mel-frequency cepstral coefficients (MFCC) (Schuller *et al.*, 2010, 2013). Gaussian mixture models (GMM) based systems trained on MFCCs have been a popular choice for speaker age prediction (Li *et al.*, 2013). Maximum *a posteriori* adaptation, discriminative training using maximum mutual information have been shown to be successful additions to GMMs (Kockmann *et al.*, 2010; Li *et al.*, 2013). Kockmann *et al.* (2010) proposed joint factor analysis with a GMM back-end for age classification. Later, i-vector with total variability modeling trained on MFCC features significantly advanced the performance of age regression achieving 7.6 years of mean absolute error (MAE) (Bahari *et al.*, 2014). Supervised i-vectors further improved the performance by decreasing the MAE by a relative 2.4% (Shivakumar *et al.*, 2014a). Within class covariance normalization was found to be useful both in the case of i-vector and supervised i-vector (Grzybowska and Kacprzak, 2016; Shivakumar *et al.*, 2014a). Cosine distance scoring is typically used for classification with i-vectors and support vector regression in the case of age regression tasks. Fedorova *et al.* (2015) reported improvements using i-vectors by adopting shallow artificial neural networks as the backend for regression.

More recently, deep neural networks have been employed for speaker age estimation. In Sadjadi *et al.* (2016), the hybrid acoustic deep neural network–hidden Markov model (DNN-HMM) from an automatic speech

[a]Electronic mail: pgurunat@usc.edu

https://doi.org/10.1121/10.0015198

recognition (ASR) system is used to extract phonetically aware senone posterior i-vector, instead of the typical GMM-universal background model (UBM). In Mallouh *et al.* (2018), bottleneck features are extracted from a hybrid DNN-HMM phone recognition system and subsequently used to train the i-vector, yielding better performance. End-to-end deep neural network architectures have also been explored for age estimation (Ghahremani *et al.*, 2018; Qawaqneh *et al.*, 2017; Zazo *et al.*, 2018). One such system, popularly termed as x-vectors, comprises several layers of time delay neural network followed by a pooling layer that computes mean and standard deviation over time. The statistics are concatenated and propagated through several feed forward layers to finally output softmax distribution over predefined, binned, age categories. With large amounts of training data or data augmentation, x-vectors have been shown to generally outperform i-vectors for age estimation (Ghahremani *et al.*, 2018). Recurrent and convolutional neural network architectures have also been explored (Sánchez-Hevia *et al.*, 2019; Zazo *et al.*, 2018).

Although there has been interest in automatic recognition of paralinguistic information from speaker data, there has been considerably less research focused on children's speech where there is significant age-dependent developmental variability (Lee *et al.*, 1999, 2014). In addition to variability due to anatomical changes, other factors related to neuro-developmental differences are reflected in the speech of children as they grow. Most of the past work involving children treat them as a broad sub-population group and perform classification across broad age groups such as children, youth, adult, and senior adults (Grzybowska and Kacprzak, 2016; Kockmann *et al.*, 2010; Li *et al.*, 2013; Qawaqneh *et al.*, 2017; Schuller *et al.*, 2010). Bocklet *et al.* (2008) proposed GMM supervectors (GMM-UBM) and support vector machines (SVM) for classification and regression of children's age. Mirhassani *et al.* (2014) proposed fuzzy based strategy to aggregate the output of multiple classifiers, each trained using MFCC features pertaining to vowels. Extreme learning machine and SVM was used for classification among children of 6 age classes (7 to 12 years). In Safavi *et al.* (2014) and Safavi *et al.* (2018), children ranging from age 4 to 14 years are categorized into three groups based on their age and classification is performed demonstrating the performance advantage of the i-vector system trained on MFCC features and linear discriminant analysis (LDA) against GMM-UBM, GMM-SVM systems. In Grzybowska and Kacprzak (2016), age regression is performed on the sub-population of children, however, the mean absolute error and the correlation in the case of children was found to be poor. Sarma *et al.* (2020) performed age and gender classification in a multitask setup among children between 4 to 14 years using deep neural network with TDNN-LSTM architecture trained on raw speech waveform. The OGI Kids corpus was employed, and data augmentation was performed using amplitude and speed perturbation to increase the training data for DNN. The results showed gender related information was not useful in prediction of children's age.

There are additional challenges involved in handling and modeling children's speech which complicate the process of automatic age estimation in children. Collection of children's speech data are relatively more expensive. The data scarcity of child speech resources poses additional challenges for data-driven statistical modeling. Typical data augmentation techniques such as speech rate, pitch perturbations which are effective tools in children's speech recognition, may prove less helpful in the case of age estimation task. This is because children's speech is characterized by age-dependent shifts in overall spectral content and formant frequencies (Potamianos and Narayanan, 2003). The inclusion of adult speech data is less likely to aid in the performance improvement of children's age estimation because of the wide mismatch of spectral parameters between children and adult speech. Moreover, Skoog Waller *et al.* (2015) in a study conducted on adults showed that speech rate influences human perception of speaker's age. A higher speech rate is associated with lower age estimates and vice versa. Error in age estimation is also linked to misclassification of gender (Barreda and Assmann, 2018). For example, perception of gender as male portrayed tendencies toward lower age estimates (Barreda and Assmann, 2018). These observations make commonly used data augmentation techniques such as speech rate and pitch perturbations unsuitable for the task of automatic age estimation.

From a speech modeling perspective, child speech is relatively more complex with high signal variability due to the developmental changes along various aspects including structural (e.g., vocal tract anatomy), motoric (e.g., speech related movements), cognitive (e.g., linguistic knowledge), and social (e.g., affect expressions) (Lee *et al.*, 1999, 2014). High within-speaker variability is also observed across all ages through adulthood (Lee *et al.*, 1999). Substantial variation in growth rates of children results in substantial variation of vocal tract structure for children of the same age and for a specific child at different ages (Vorperian *et al.*, 2009), which further complicates modeling of children's age from speech. High inter-speaker variability observed across age groups (Gerosa *et al.*, 2009; Lee *et al.*, 1999) poses further difficulties in estimating efficient within-age class boundaries. It is well documented in the literature that children's speech recognition is significantly less accurate (Gerosa *et al.*, 2009; Potamianos and Narayanan, 2003; Shivakumar and Georgiou, 2020; Shivakumar and Narayanan, 2022; Shivakumar *et al.*, 2014b) underscoring the modeling difficulties associated with children's speech.

From a psycho-acoustics perspective, the perception of children's age is particularly distinct for the following reasons. Humans tend to incorporate assumptions about a child speaker's gender in estimating child's age (Barreda and Assmann, 2018). In general, speaker gender inference is relatively poor in the case of younger children compared to adult speakers since speech of both female and male children is characterized with higher F0 values and high overlap between distributions of F0 and formant frequencies (Barreda and Assmann, 2021), which potentially manifests

as errors in age estimation. Human listeners were found to persistently underestimate age for older girls (Barreda and Assmann, 2018). These trends pose further challenges in speaker age estimation in children.

Our work is motivated from the investigations of variations of temporal parameters in children across age categories (Lee *et al.*, 1999, 2014). Lee *et al.* (1999) found phoneme durations to be associated with speech development in children. In this study, we propose novel features derived from phone durations for the task of age estimation from children's speech. Manual transcriptions are forced-aligned with speech data to obtain phone duration distributions and are subsequently used to train regression models. Although there have been a few past works that incorporate durations in terms of pauses and overall length of utterances in speaker age prediction, our work is distinct in its explicit modeling of phone duration in children's speech as a biomarker for speaker age estimation. To the best of our knowledge, this work is a unique attempt at modeling speaker age information purely based on temporal variations. Our study is one of the very few targeting low data scenarios which are commonplace considering the availability of age-labeled children's speech.

The rest of the paper is organized as follows: Section II A describes the proposed phone-duration features and Sec. II B presents the regression model. Section III describes the speech databases employed in our study. The experimental setup is described in Sec. IV. The results are presented and discussed in Sec. V. Finally, the study conclusions are provided in Sec. VI and possible future directions discussed.

II. PHONE DURATION FEATURES

Duration is an important descriptor of speech signals. They can convey a variety of information ranging from low level descriptors such as speaking rate and pauses in speech to more abstract information about cognitive process, emotion, and conversational dynamics. Lee *et al.* (1999) studied the crucial role of variability of duration in children's speech. Analysis of durations of ten vowels and fricative portion of /s/ established significant effect of age on duration descriptors. Younger children especially of age 5 and 6 years exhibited significantly longer mean vowel durations compared to older children, with age-dependent duration values reaching a minimum around the age of 15 years. Increased intra and inter speaker variation in duration was observed across age groups. Both inter and intra speaker variation patterns reduce with increasing age and approach adult levels for children of 13 years and higher. It was found that younger children tend to exaggerate long vowels including /i/, /æ/, /ɑ/, and /ɝ/. The authors also reported that the effect of gender on the duration was not significant.

Studies have established correlation patterns between mean durations of predefined set of syllables and children speaker's true (chronological) age as well as human perceived age (Barreda and Assmann, 2018). Phrase and word durations, as well as the inter-word pause durations, decrease with increase in age (Singh *et al.*, 2007). Phonological studies in children link phoneme durations to developing speech articulatory and neuro-motor timing control in growing children (Kent, 1976; Kent and Forner, 1980; Smith, 1978). Lee *et al.* (1999) also found significant correlation between child age and sentence duration.

Psycholinguistic studies have also found that speech duration is related to cognition in children, i.e., speech patterns reveal children take longer time to express utterances with higher cognitive demand (Dillon, 1983). Cognitive processes such as selection, retrieval and planning also reflect in temporal speech pause patterns of children (Esposito *et al.*, 2004).

Moreover, children's speech is associated with increased mispronunciations, disfluencies, frequent pauses, non-vocal verbalization (Potamianos and Narayanan, 2003; Potamianos *et al.*, 1997). Children's speech is characterized by repetitions and revisions which are reflective of language development (Gallagher, 1977). Phone duration distributions can implicitly encode such speech characteristics found in children and have the ability to capture several linguistic supra-segmental properties including speaking rate and stress markers. In summary, phone durations carry critical information about various aspects of child development.

A. Proposed phone-duration features

Motivated by the findings in prior literature, in this work we propose features explicitly engineered to model phone duration distribution in child speech to determine a speaker's age. First, the speech data are forced-aligned to the manual transcriptions. Subsequently, the temporal occupancy distribution for the following set of phones are computed:

- Position dependent phones: to capture temporal patterns of phones depending on their position in the word [beginning (B), intermediate (I), or end (E)] or in isolation.
- Position-independent phones: obtained from aggregated statistics of position-independent phones.
- Lexical stress marked phones: vowels carrying either no stress, primary stress, or secondary stress.
- Silence phones: to model the pauses and speaking traits.
- Special phones such as spoken noise: to model and capture hesitancy, disfluencies, and filled pauses.
- Global distributions: set of all non-silence phones, consonants, and vowels.

Finally, several statistical functionals are computed from the duration distributions for each phone, i.e., eight distribution descriptors, namely, *mean, variance, minimum, maximum, skewness, kurtosis, entropy*, and *mean absolute deviation*.

Figure 1 shows an example of duration distribution for the position dependent phoneme /l/(E), where "(E)" indicates the occurrence of the phone /l/ at the end of the word. The figure shows histograms comparing phone duration distributions for three different corpora: an adult speech

https://doi.org/10.1121/10.0015198

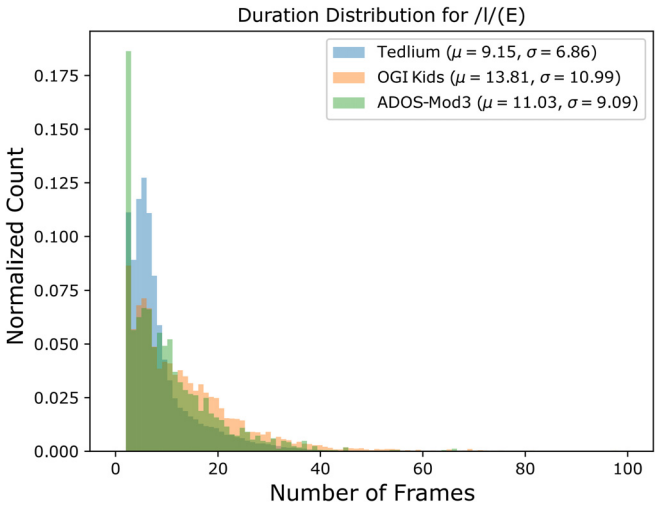


FIG. 1. (Color online) Phone duration distribution for phoneme /l/ (end position)—Adult (TEDLIUM) vs children (OGI Kids & ADOS-Mod3).

(Tedlium) and two child speech (OGI Kids and ADOS-Mod3; see Sec. III for descriptions) datasets. It is evident from the figure, that adult speech is associated with significantly shorter durations compared to children's speech. Children's phone durations have typically higher means and standard deviations. Such developmental trends with phoneme durations are well known among children of different age categories. Figure 2 presents the phone duration distributions of phoneme /t/(I) [(I) represents intermediate position of phoneme in words] for each age group ranging from kindergarten to children studying in 10th grade.

B. Proposed age regression model

The proposed regression model architecture is shown in Fig. 3. The architecture is based on stacking two layers of regressors. An ensemble of individual estimators, handling one phoneme each, are trained independently on the eight previously described feature descriptors to predict the speaker's age. A final, meta regressor that operates on the outputs of the individual estimators is used to obtain the final age prediction. The final estimator is trained on the predictions of individual estimators using cross-validation. In this work, we employ two regression models, a support vector regressor and a random-forest based AdaBoost regressor. The meta regressor model is of the same class as the base individual estimator. The choice of regression models is based on the following factors: (i) the amount of training data available; smaller amounts of data make DNN based models unsuitable, (ii) support vector based models are a popular choice in prior literature due to their performance,

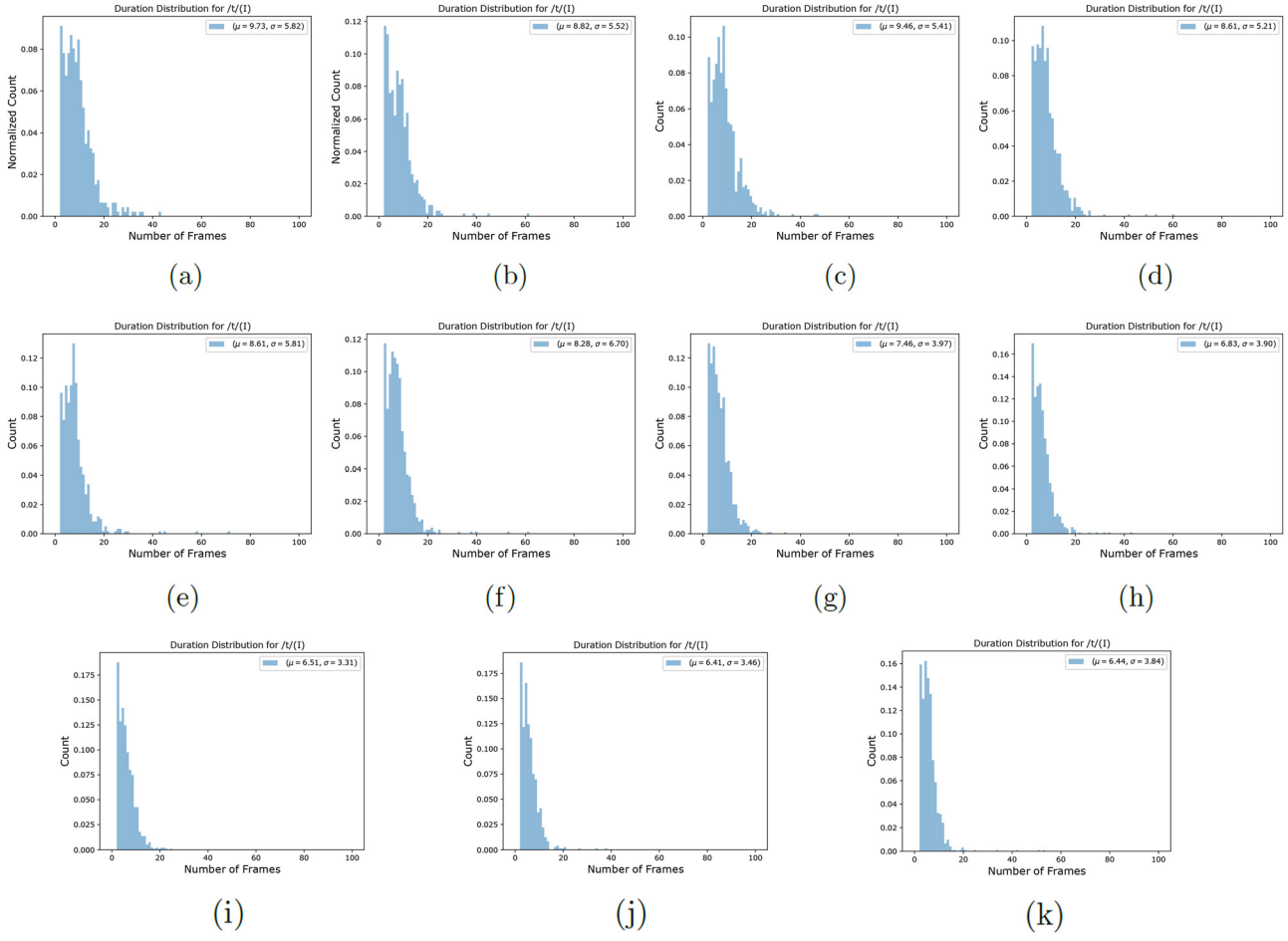


FIG. 2. (Color online) Phone duration distribution for phone /t/(I) for different ages.

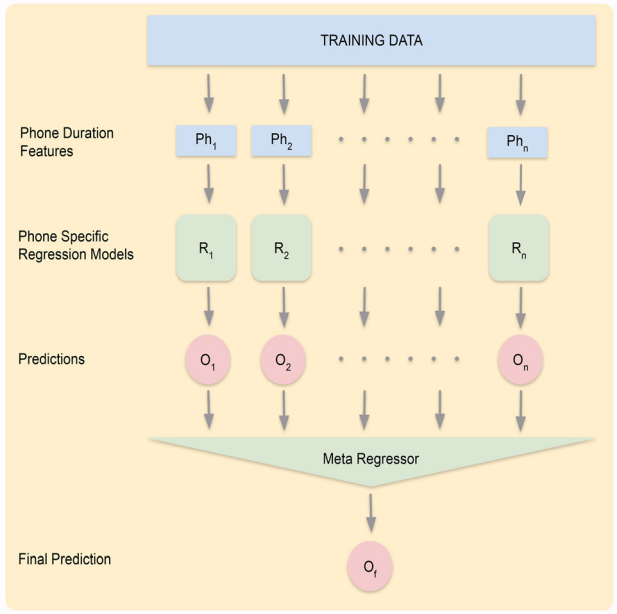


FIG. 3. (Color online) Proposed age regression model architecture.

and (iii) the decision tree based model is used for inferring feature importance.

The proposed stacking ensemble learning offers certain advantages over a single regression model. Ensemble learning helps in achieving low bias and low variance in final predictions. The stacked estimators also help handle high feature dimension (2912 features, i.e., 364 phonemes, 8 features each) efficiently in contrast to typical dimension reduction alternatives. The stacked architecture for duration modeling helps in alleviating over-fitting issues since the final estimator is trained on the cross-validated predictions of the base estimators. It enables implicit feature selection among different phonemes, since the meta classifier operates on top of outputs of base estimators pertaining to each phoneme. It also has the added advantage to enable easier assessment of the contribution of information provided by phone duration distributions for age estimation.

III. CHILD SPEECH DATABASES

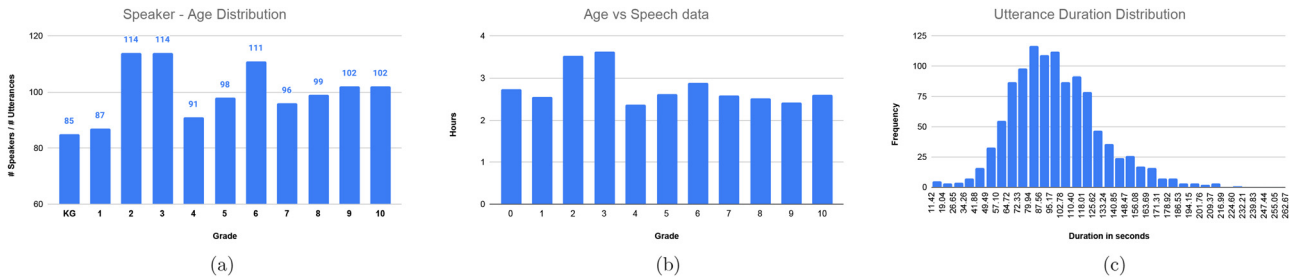
In this study, we perform experiments on two child speech corpora to assess the transferability and robustness to different domains and acoustic conditions.

A. OGI Kids speech corpus

We employ OGI Kids speech corpus (Shobaki *et al.*, 2000) as the primary database for our experiments due to its wide age distribution demographics among children. We make use of the spontaneous speech subset of the corpus comprising adult interviewers asking a series of questions and eliciting a spontaneous response from children. The corpus consists of 1100 distinct children speakers with ages ranging from children studying in kindergarten to 10th grade. Since chronological age in terms of months or years is not available, we use the grade level as a proxy for chronological age for this data. It includes a total of approximately 30.5 h of speech recorded using a head-mounted microphone. For every speaker there is a single recording. The mean duration for each speaker is approximately 100 s. The speaker age distribution, amount of speech data per age and utterance duration distribution statistics are presented in Fig. 4.

B. Autism diagnostic observation schedule—Module 3 (ADOS-Mod3) data

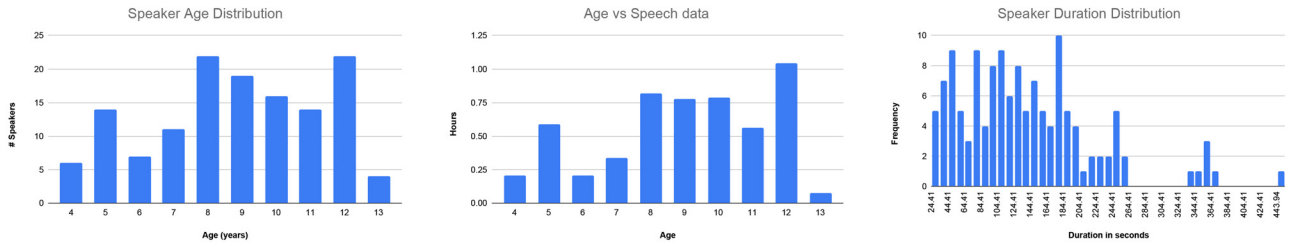
The ADOS-Mod3 corpus (Lord *et al.*, 2000) comprises child-adult dyadic conversations involving semi-structured, standardized assessment of communication and social interactions. The children in the corpus are diagnosed with autism spectrum disorder (ASD), attention deficit hyperactivity disorder (ADHD) and various other developmental disorders including language disorder. The speech sessions were collected at two different locations including the University of Michigan Autism and Communication Disorder Center and the Cincinnati Children's Medical Center. A single distant microphone was used to record the speech data. We make use of only the speech data from children for speaker age estimation and omit the adult speech data from the modeling. The corpus consists of 179 children, out of which we consider a subset of 135 children for whom we had generated good quality automatic phonetic alignments (see Sec. IV for details regarding alignments) for the age regression analysis of this study. The age of children range from 43 to 158 months (4 to 13 years). The corpus contains a total of 5.4 h of manually transcribed speech. Each speaker has a mean duration of approximately 144 s. The speaker age distribution, amount of speech data per age and speaker duration distribution are presented in Fig. 5.



(a) (b) (c)

FIG. 4. (Color online) OGI kids corpus statistics.

https://doi.org/10.1121/10.0015198



(a) (b) (c)

FIG. 5. (Color online) ADOS-Mod3 data statistics.

Several factors associated with this dataset add additional complexity to the task of speaker age estimation. First, the differences in the neuro-developmental condition due to ASD, ADHD, and other developmental disorders possibly reflected in the speech adds to complexity of the age estimation from speech (Oller *et al.*, 2010). Second, these speech data are recorded under far-field conditions with a single distant microphone, contributing to greater acoustic signal variability. Third, the data analyzed are from two different locations with different room and channel characteristics adding to the complexity of speech modeling. Finally, the corpus consists of significantly less data (17%) compared to the OGI speech corpus. The above challenges help us evaluate the robustness of the proposed phone duration model.

IV. EXPERIMENTAL SETUP

In this section, we provide the details of our experimental setup. For forced-alignment we employ the KALDI speech recognition toolkit (Povey *et al.*, 2011). The feature pipeline consists of extracting 13-dimensional mel-filter cepstral coefficients (MFCC) using a window size of 25 ms and a shift of 10 ms. Linear discriminant analysis (LDA) transformation is applied to the MFCC features by considering left and right context of three frames. Furthermore, the maximum likelihood linear transform is applied on top of LDA features. Finally, feature-space maximum likelihood linear regression (fMLLR) based speaker-adaptive training is used to train a Gaussian mixture model–hidden Markov model based acoustic model. The resulting acoustic model is used for forced-alignment to obtain phoneme level alignments. Later, the statistics are accumulated for each phoneme under consideration and their eight functional descriptors, namely, *mean, variance, minimum, maximum, skewness, kurtosis, entropy*, and *mean absolute deviation* are computed.

Two separate acoustic models are trained for forced alignments, one for the OGI Kids corpus and the other for the ADOS-Mod3 corpus. Since both the OGI Kids corpus and ADOS-Mod3 corpus are fairly small, we include additional speech data for constructing better acoustic models and thereby obtaining better quality alignments. In the case of the OGI Kids corpus, we employ an acoustic model trained on the My Science Tutor (MyST) corpus. The MyST corpus (Ward *et al.*, 2011) comprises 198 h of conversational children's speech involving children studying in grades 3, 4, and 5. The speech is recorded under low noise and close talk conditions similar to OGI Kids. The acoustic model is identical to the one in Shivakumar and Narayanan (2022), and achieves a WER of 30.4% on the OGI Kids corpus. We perform forced alignment on a leave-one-speaker-out basis using the OGI Kids corpus to ensure that the test speaker is unseen both during alignment and regression evaluation. The acoustic model employed for ADOS-Mod3 is trained by including 44 children (excluding speaker subjects employed for age estimation) and adult speech of clinicians conducting the autism diagnostic assessment. The addition of adult speech is known to yield better quality of acoustic models under low data scenarios (Shivakumar and Georgiou, 2020). We do not include the MyST data since we believe the addition of adult speech data under similar recording conditions, i.e., far-field, high reverberation environment in the case of ADOS-Mod3 is more beneficial. Performance in the case of AdosMod3 is much worse yielding a WER of approximately 70%. The total number of phones in the OGI Kids corpus is 364, whereas the number of phones in the case of the ADOS-Mod3 corpus is restricted to 185 phones (excluding lexical stress markers) to better handle the smaller size of the training corpus.

For the age estimation task, we directly perform regression to predict the (reference) school grade level of children (used as proxy for age) in the case of the OGI Kids corpus. In the case of ADOS-Mod3, the age of children were converted from months to years. The performances are hence roughly comparable between the two models. In this work, we experiment with two regression models, i.e., support vector machine regressor (SVR) and the decision tree based random forest AdaBoost regressor. The choice of SVR is due to its popularity and proven effectiveness for age estimation in prior works. The decision tree based AdaBoost model offers the benefit of providing feature importance which helps us analyze the contributions of phonemes and their discriminative power in age estimation. Given the small size of the speech corpora, we perform leave-one-speaker-out (LOSO) cross-validation. The hyper-parameter tuning of the regression models are handled implicitly through nested cross-validation. For performance evaluation, we report mean absolute error (MAE), proportion of explained variance (R^2 score) and Pearson correlation.

TABLE I. Results: Children speaker age estimation. Mean and median baseline have similar MAE.

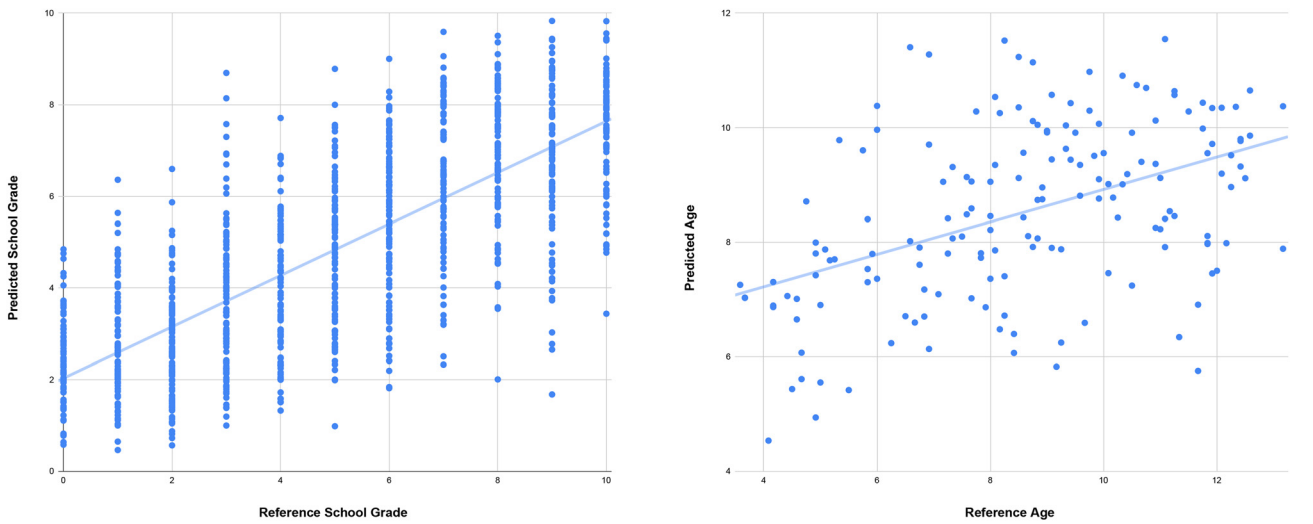
Database	Model	MAE	R^2 score	Correlation
OGI Kids	Baseline	2.69	0.0	0.0
	SVR (RBF)	1.62	0.58	0.76
	AdaBoost	1.82	0.48	0.71
ADOS-Mod3	Baseline	2.07	0.0	0.0
	SVR (RBF)	1.79	0.24	0.49
	AdaBoost	1.74	0.29	0.54

V. RESULTS

Table I presents the results of children speaker age estimation on the OGI Kids and ADOS-Mod3 data sets through regression. In this work, we employ trivial baseline models that predict the mean and median of the age of the speakers. Both the mean and median baseline model achieve MAE of 2.69 and 2.07 on OGI Kids and Ados-Mod3 corpus, respectively. The proposed phone duration model achieves a mean absolute error of 1.62 years and a correlation of 0.76 with SVR on the OGI Kids corpus. The results are significantly better than the baseline system based on mean age prediction. Moreover, it is notable that the correlation results are comparable to correlation between human perceived speaker age and the chronological age which is estimated to be approximately 0.7 (Barreda and Assmann, 2018). We observe that the SVR model outperforms the AdaBoost model.

The results on the ADOS-Mod3 data are slightly worse compared to OGI Kids. This is not surprising due to three factors. (i) Neuro-developmental disorders can impact cognitive development (Joseph *et al.*, 2002; Karalunas *et al.*, 2018), and be reflected in language skills and speech production differences (Long *et al.*, 2011; Sperdin and Schaer, 2016); this in turn can lead to differences in speaker age perception and predictions from the acoustic speech signal. (ii) The ADOS-Mod3 corpus (5.4 h) has significantly less data compared to the OGI Kids data set (30.5 h). (iii) ADOS-Mod3 corpus is noisier due to far field recording conditions, which negatively affects quality of alignments. However, the estimation results are significantly better than the mean baseline. In the case of ADOS-Mod3, we observe that the AdaBoost regression model outperforms the SVR. The results on ADOS-Mod3 corpus further attests to the robustness of the proposed phone duration modeling. Figures 6(a) and 6(b) illustrate scatter plots of chronological age vs the predicted speaker age on the OGI Kids and ADOS-Mod3 corpora, respectively. Note, the scatter plots for OGI Kids have age represented in terms of the school grade of children, whereas the ADOS-Mod3 has age in terms of months. Overall, the results suggest that the proposed phone duration features convey valuable developmental information in children's speech. This can be leveraged in turn to robustly estimate speaker age in children by modeling the temporal variation via phone duration distributions.

Next, we perform additional analysis to assess the factor of age on the performance of the system. Figure 7 plots the mean absolute error in each age category of the OGI Kids corpus using the SVR model. From the results, we observe that the error is low for children ages ranging from 2nd grade to 8th grade categories, reaching minimum for the 2nd grade group. However, the error increases sharply for younger (kindergarten and 1st grade) and older children in the corpus (9th–10th grade). One possible explanation for the observed trend is as follows: (i) in the case of younger children studying in kindergarten and 1st grade, although both the inter and intra age variations are expected to be high, the intra-age variation dominates, thereby resulting in higher error, (ii) in the case of children between 2nd and 8th grades, the inter-age variation dominates resulting in lower error rates, and (iii) in case of older children (9th and 10th grade) the inter age variations are significantly less which results in relatively low performance.



(a) (b)

FIG. 6. (Color online) Age regression scatter plot.

https://doi.org/10.1121/10.0015198

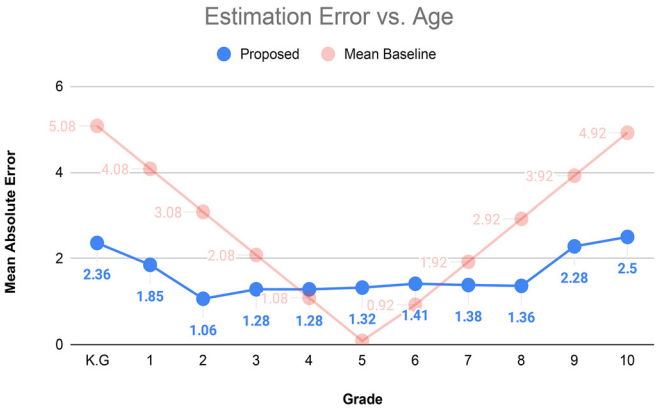


FIG. 7. (Color online) Mean absolute error (MAE) in estimation for different age categories—OGI Kids Corpus.

We also perform feature importance analysis to gain insights on the contributions of each phoneme toward children speaker age estimation. We derive impurity based feature importance that are accessible from tree-based algorithms, in our case the random forest based AdaBoost model trained on the OGI Kids corpus. The importance measures are computed based on total reduction of the optimization criterion, often referred to as the Gini importance. We compute the feature importance only on the final, meta estimator that operates on the output of the phoneme specific base estimators. This allows us to assess the contribution of input features at the phoneme level. Figure 8 shows the bar plot of the top-20 most contributing phonemes computed on the OGI Kids corpus. Higher values translate to more importance. The following observations are made with our experimental setup:

(1) /!/ (aggregated duration of *all* position dependent non-silent phones) is the most important feature for speaker age estimation. We believe this finding derives from the observations that the overall sentence duration decreases with increase in age (Lee *et al.*, 1999). Another reason can be the increased repetitions, false-starts and filled pauses (e.g., um, uh) that are characteristic of children's speech especially at younger age (Potamianos and Narayanan, 2003).
(2) Durations of position dependent phones /t/(I), /h/(B), and /k/(I) is a discriminating factor of speaker age in children. Moreover, we find that these position dependent phones are consonants. In conjunction to the observations made in Fig. 2, specifically for phoneme /t/(I), there is evidence that younger children spend more time during utterance of consonants occurring in the beginning and middle of a word. The duration decreases with age of children.
(3) /-/ (silence) durations can intrinsically capture inter-word pauses, speaking rate, hesitations, and disfluencies which are known to correlate with children's age (Potamianos and Narayanan, 2003).
(4) The above five phoneme durations [/!/, /t/(I), /h/(B), /k/(I), /-/] are more than twice as important as other phonemes in the speech acoustic inventory.
(5) The appearance of position dependent phones among the top-20 indicates that (differential) duration of phones appearing in different parts of a word carry discriminating information on children growth. This finding is of interest given that most of the prior studies have only considered position-independent phonemes (Gerosa *et al.*, 2006; Lee *et al.*, 1999; Potamianos and Narayanan, 2003).

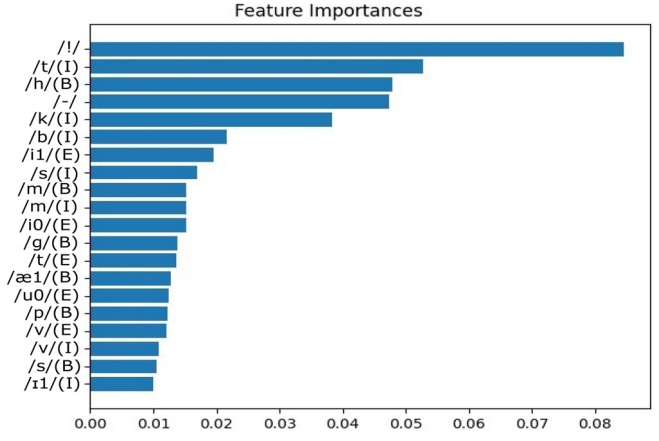


FIG. 8. (Color online) Phoneme-wise feature importance—OGI Kids Corpus.

VI. CONCLUSION AND FUTURE WORK

In this work, we investigated features solely based on phone durations in speech (i.e., acoustic realizations of phonemes) for the task of speaker age estimation from children's speech. Phoneme occupancy distributions are derived by force-aligning manual transcripts with the speech signal. Statistical functionals describing the distributions are extracted for each phone which constitute the features for the age estimation. A double layer stacking regressor architecture is employed with a meta estimator operating on top of multiple base estimators, each trained on statistical functional features corresponding to each phoneme. The results suggest that phone durations contain critical developmental information helpful in predicting the age of children speakers. The results indicate that age of children can be effectively predicted by using just temporal information obtained from the speech signal. The best performing phone duration model yields a mean absolute error of 1.62 and a correlation of 0.76. The estimation of age from children's speech is associated with higher error among young and older children, while yielding minimum estimation error in children within the 2nd grade to 8th grade age groups. We find that aggregated phone durations of non-silence phones are the most important feature. Among the other phonemes, /t/(I), /h/(B), and /k/(I) play an important role. We also find that inter-speech silence duration also plays an important role in predicting child speaker age. Subsequent experiments on additional speech corpora, ADOS-Mod3, comprising speech data from children with ASD/ADHD diagnosis,

further underscores the robustness of the phone duration features. Relatively higher estimation errors observed with ADOS-Mod3 may imply that the cognitive developmental age as established from speech may reflect developmental differences that may be due to cognitive and other factors related to neuro-diversity.

In the future, we plan to combine the phone duration features along with other speech based features including spectral features such as MFCC and voice quality features such as jitter and shimmer to explore complementary information to duration for improved age estimation. Future work should also consider scenarios when manual speech transcripts are unavailable and thus alignments derived from an ASR is the only option, i.e., exploring the effect of automated transcriptions on the performance of speaker age estimation. This work also sets the stage for future work that helps systematically disentangle the various underlying developmental factors reflected in speech duration (anatomical, neuro cognitive, motoric, and social elements in speech communication).

Bahari, M. H., McLaren, M., Van hamme, H., and van Leeuwen, D. A (**2014**). "Speaker age estimation using i-vectors," Eng. Appl. Artif. Intell. **34**, 99–108.

Barreda, S., and Assmann, P. F. (**2018**). "Modeling the perception of children's age from speech acoustics," J. Acoust. Soc. Am. **143**(5), EL361–EL366.

Barreda, S., and Assmann, P. F. (**2021**). "Perception of gender in children's voices," J. Acoust. Soc. Am. **150**(5), 3949–3963.

Bocklet, T., Maier, A., and Nöth, E. (**2008**). "Age determination of children in preschool and primary school age with GMM-based supervectors and support vector machines/regression," in *International Conference on Text, Speech and Dialogue* (Springer, Berlin), pp. 253–260.

Bone, D., Chaspari, T., and Narayanan, S. (**2017a**). "Behavioral signal processing and autism: Learning from multimodal behavioral signals," in *Autism Imaging Devices* (CRC Press, Boca Raton, FL), pp. 335–360.

Bone, D., Lee, C.-C., Chaspari, T., Gibson, J., and Narayanan, S. (**2017b**). "Signal processing and machine learning for mental health research and clinical applications," IEEE Sign. Process. Mag. **34**(5), 196.

Dillon, J. (**1983**). "Cognitive complexity and duration of classroom speech," Instrum. Sci. **12**(1), 59–66.

Esposito, A., Marinaro, M., and Palombo, G. (**2004**). "Children speech pauses as markers of different discourse structures and utterance information content," in *Proceedings of the International Conference: From Sound to Sense*, Vol. 50, pp. 10–13.

Fedorova, A., Glembek, O., Kinnunen, T., and Matějka, P. (**2015**). "Exploring ANN back-ends for i-vector based speaker age estimation," in *Sixteenth Annual Conference of the International Speech Communication Association*.

Gallagher, T. M. (**1977**). "Revision behaviors in the speech of normal children developing language," J. Speech Hear. Res. **20**(2), 303–318.

Gerosa, M., Giuliani, D., Narayanan, S., and Potamianos, A. (**2009**). "A review of ASR technologies for children's speech," in *Proceedings of the 2nd Workshop on Child, Computer and Interaction*, pp. 1–8.

Gerosa, M., Lee, S., Giuliani, D., and Narayanan, S. (**2006**). "Analyzing children's speech: An acoustic study of consonants and consonant-vowel transition," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, IEEE*, Vol. 1.

Ghahremani, P., Nidadavolu, P. S., Chen, N., Villalba, J., Povey, D., Khudanpur, S., and Dehak, N. (**2018**). "End-to-end deep neural network age estimation," in *Interspeech*, pp. 277–281.

Grzybowska, J., and Kacprzak, S. (**2016**). "Speaker age classification and regression using i-vectors," in *INTERSPEECH*, pp. 1402–1406.

Joseph, R. M., Tager-Flusberg, H., and Lord, C. (**2002**). "Cognitive profiles and social-communicative functioning in children with autism spectrum disorder," J. Child Psychol. Psychiat. **43**(6), 807–821.

Karalunas, S. L., Hawkey, E., Gustafsson, H., Miller, M., Langhorst, M., Cordova, M., Fair, D., and Nigg, J. T. (**2018**). "Overlapping and distinct cognitive impairments in attention-deficit/hyperactivity and autism spectrum disorder without intellectual disability," J. Abnorm. Child Psychol. **46**(8), 1705–1716.

Kent, R. D. (**1976**). "Anatomical and neuromuscular maturation of the speech mechanism: Evidence from acoustic studies," J. Speech Hear. Res. **19**(3), 421–447.

Kent, R. D., and Forner, L. L. (**1980**). "Speech segment durations in sentence recitations by children and adults," J. Phon. **8**(2), 157–168.

Kockmann, M., Burget, L., and Černockỳ, J. (**2010**). "Brno University of Technology system for Interspeech 2010 paralinguistic challenge," in *Eleventh Annual Conference of the International Speech Communication Association*.

Lee, S., Potamianos, A., and Narayanan, S. (**1999**). "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," J. Acoust. Soc. Am. **105**(3), 1455–1468.

Lee, S., Potamianos, A., and Narayanan, S. (**2014**). "Developmental acoustic study of American English diphthongs," J. Acoust. Soc. Am. **136**(4), 1880–1894.

Li, M., Han, K. J., and Narayanan, S. (**2013**). "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," Comput. Speech Lang. **27**(1), 151–167.

Long, C., Gurka, M. J., and Blackman, J. (**2011**). "Cognitive skills of young children with and without autism spectrum disorder using the BSID-III," Autism Res. Treat. **2011**, 759289.

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., Pickles, A., and Rutter, M. (**2000**). "The autism diagnostic observation schedule–generic: A standard measure of social and communication deficits associated with the spectrum of autism," J. Autism Dev. Disorders **30**(3), 205–223.

Mallouh, A. A., Qawaqneh, Z., and Barkana, B. D. (**2018**). "New transformed features generated by deep bottleneck extractor and a GMM–UBM classifier for speaker age and gender classification," Neural Comput. Applic. **30**(8), 2581–2593.

Mirhassani, S. M., Zourmand, A., and Ting, H.-N. (**2014**). "Age estimation based on children's voice: A fuzzy-based decision fusion strategy," Sci. World J. **2014**, 534064.

Oller, D. K., Niyogi, P., Gray, S., Richards, J. A., Gilkerson, J., Xu, D., Yapanel, U., and Warren, S. F. (**2010**). "Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development," Proc. Natl. Acad. Sci. U.S.A. **107**(30), 13354–13359.

Potamianos, A., and Narayanan, S. (**2003**). "Robust recognition of children's speech," IEEE Trans. Speech Audio Process. **11**(6), 603–616.

Potamianos, A., Narayanan, S., and Lee, S. (**1997**). "Automatic speech recognition for children," in *Fifth European Conference on Speech Communication and Technology*.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G., and Veselý, K. (**2011**). "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, IEEE Signal Processing Society.

Qawaqneh, Z., Mallouh, A. A., and Barkana, B. D. (**2017**). "DNN-based models for speaker age and gender classification," in *International Conference on Bio-Inspired Systems and Signal Processing*, Vol. 5, pp. 106–111.

Sadjadi, S. O., Ganapathy, S., and Pelecanos, J. W. (**2016**). "Speaker age estimation on conversational telephone speech using senone posterior based i-vectors," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 5040–5044.

Safavi, S., Russell, M., and Jančovič, P. (**2014**). "Identification of age-group from children's speech by computers and humans," in *Fifteenth Annual Conference of the International Speech Communication Association*.

Safavi, S., Russell, M., and Jančovič, P. (**2018**). "Automatic speaker, age-group and gender identification from children's speech," Comput. Speech Lang. **50**, 141–156.

Sánchez-Hevia, H. A., Gil-Pita, R., Utrilla-Manso, M., and Rosa-Zurera, M. (**2019**). "Convolutional-recurrent neural network for age and gender prediction from speech," in *2019 Signal Processing Symposium (SPSympo)*, IEEE, pp. 242–245.

Sarma, M., Sarma, K. K., and Goel, N. K. (**2020**). "Children's age and gender recognition from raw speech waveform using DNN," in *Advances in Intelligent Computing and Communication* (Springer, Berlin), pp. 1–9.
Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., and Narayanan, S. S. (**2010**). "The interspeech 2010 paralinguistic challenge," in *Eleventh Annual Conference of the International Speech Communication Association*.
Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., and Narayanan, S. S. (**2013**). "Paralinguistics in speech and language–state-of-the-art and the challenge," Comput. Speech Lang. **27**(1), 4–39.
Shivakumar, P. G., and Georgiou, P. (**2020**). "Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations," Comput. Speech Lang. **63**, 101077.
Shivakumar, P. G., Li, M., Dhandhania, V., and Narayanan, S. S. (**2014a**). "Simplified and supervised i-vector modeling for speaker age regression," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 4833–4837.
Shivakumar, P. G., and Narayanan, S. (**2022**). "End-to-end neural systems for automatic children speech recognition: An empirical study," Comput. Speech Lang. **72**, 101289.
Shivakumar, P. G., Potamianos, A., Lee, S., and Narayanan, S. S. (**2014b**). "Improving speech recognition for children using acoustic adaptation and pronunciation modeling," in *Wocci*, pp. 15–19.
Shobaki, K., Hosom, J.-P., and Cole, R. A. (**2000**). "The OGI Kids' speech corpus and recognizers," in *Sixth International Conference on Spoken Language Processing*.
Singh, L., Shantisudha, P., and Singh, N. C. (**2007**). "Developmental patterns of speech production in children," Appl. Acoust. **68**(3), 260–269.
Skoog Waller, S., Eriksson, M., and Sörqvist, P. (**2015**). "Can you hear my age? Influences of speech rate and speech spontaneity on estimation of speaker age," Front. Psychol. **6**, 978.
Smith, B. L. (**1978**). "Temporal aspects of English speech production: A developmental perspective," J. Phon. **6**(1), 37–67.
Sperdin, H. F., and Schaer, M. (**2016**). "Aberrant development of speech processing in young children with autism: New insights from neuroimaging biomarkers," Front. Neurosci. **10**, 393.
Vorperian, H. K., Wang, S., Chung, M. K., Schimek, E. M., Durtschi, R. B., Kent, R. D., Ziegert, A. J., and Gentry, L. R. (**2009**). "Anatomic development of the oral and pharyngeal portions of the vocal tract: An imaging study," J. Acoust. Soc. Am. **125**(3), 1666–1678.
Ward, W., Cole, R., Bolanos, D., Buchenroth-Martin, C., Svirsky, E., Vuuren, S. V., Weston, T., Zheng, J., and Becker, L. (**2011**). "My science tutor: A conversational multimedia virtual tutor for elementary school science," ACM Trans. Speech Language Process. (TSLP) **7**(4), 1–29.
Zazo, R., Nidadavolu, P. S., Chen, N., Gonzalez-Rodriguez, J., and Dehak, N. (**2018**). "Age estimation in short speech utterances based on lstm recurrent neural networks," IEEE Access **6**, 22524–22530.