A Deep Learning Algorithm for Objective Assessment of Hypernasality in Children with Cleft Palate

Vikram C. Mathad,

Department of Speech & Hearing Sciences, Arizona State University, Tempe, AZ-85281

Nancy Scherer,

Department of Speech & Hearing Sciences, Arizona State University, Tempe, AZ-85281

Kathy Chapman,

Department of Communication Sciences and Disorders, University of Utah, Salt Lake City, UT-84112

Julie M. Liss,

Department of Speech & Hearing Sciences, Arizona State University, Tempe, AZ-85281

Visar Berisha

College of Health Solutions, and School of Electrical, Computer, & Energy Engineering, Arizona State University, Tempe, AZ-85281

Abstract

Objectives: Evaluation of hypernasality requires extensive perceptual training by clinicians and extending this training on a large scale internationally is untenable; this compounds the health disparities that already exist among children with cleft. In this work, we present the objective hypernasality measure (OHM), a speech-based algorithm that automatically measures hypernasality in speech, and validate it relative to a group of trained clinicians.

Methods: We trained a deep neural network (DNN) on approximately 100 hours of a publicly-available healthy speech corpus to detect the presence of nasal acoustic cues generated through the production of nasal consonants and nasalized phonemes in speech. Importantly, this model does not require any clinical data for training. The posterior probabilities of the deep learning model were aggregated at the sentence and speaker-levels to compute the OHM.

Results: The results showed that the OHM was significantly correlated with perceptual hypernasality ratings from the Americleft database (*r=0.797, p <0.001*) and the New Mexico Cleft Palate Center (NMCPC) database (*r=0.713, p<0.001*). In addition, we evaluated the relationship between the OHM and articulation errors; the sensitivity of the OHM in detecting the presence of very mild hypernasality; and established the internal reliability of the metric. Further, the performance of the OHM was compared with a DNN regression algorithm directly trained on the hypernasal speech samples.

Significance: The results indicate that the OHM is able to measure the severity of hypernasality on par with Americleft-trained clinicians on this dataset.

vchikkaw@asu.edu .

Author Manuscript

Keywords

Cleft palate; clinical speech analysis; deep neural networks; hypernasality; speech assessment; vocal biomarkers

I. INTRODUCTION

Cleft palate (CP), with or without cleft lip, is a craniofacial anomaly and the most common birth disorder, with 1 in every 700 live births presenting with craniofacial clefts [1]. In healthy craniofacial development, the bilateral bony palatal shelves fuse horizontally at midline to create the roof of the mouth (hard palate) and provide points of muscular attachment for the soft palate (velum). These velar muscles, along with those in the upper pharynx, allow for modulation of the opening between the oral and nasal cavities (velopharyngeal port) during respiration, swallowing, and speaking. The failure of the palatal shelves to fuse at midline during embryological development (cleft) means there is no intact hard or soft palate and no separation between the oral and nasal cavities. The primary intervention involves surgical repair of the palatal cleft to produce anatomical closure and to create the ability to modulate the velopharyngeal port aperture. When velopharyngeal dysfunction (VPD) persists post primary palate surgery [2], different surgeries are available based on individual surgeon preference. In the presence of VPD, the velopharyngeal port fails to close off the nasal tract completely during speech production for non-nasal sounds, and air and acoustic energy escape through the nasal cavity, resulting in reduced speech intelligibility. Twenty to thirty percent of children with clefts require a secondary surgery to rectify VPD [3] for the exclusive purpose of improving speech outcomes.

The inability to achieve adequate velopharyngeal closure during speech results in the percept of hypernasality, characterized by excessive nasal resonance due to passage of the vibrating column of air through the nasal cavity[1]. The perception of hypernasality in speech, secondary to VPD, is considered a primary outcome measure in CP as it drives decisions related to secondary surgery, speech therapy, and is an important determinant of long-term educational and social outcomes [4], [5]. As a result, it is considered a primary outcome by the American Cleft Palate-Craniofacial Association and the Cleft Palate Committee of the International Association of Logopedics and Phoniatrics [6].

Instrumental methods such as videoendoscopy, vide-ofluorscopy, nasometry, magnetic resonance imaging, and aerodynamic techniques can be used to assess VPD and aerodynamics [7]. Most of these instruments require special training and expensive equipment. Nasometry is an exception but it only shows a moderate correlation with perceptual impressions of hypernasality [8] and it is not part of the standard battery in the Americleft rating protocol [9]. Instead, clinicians rely on their perception of hypernasality to assess VPD. Perception of hypernasality is a complex task that requires the clinician to infer, from the acoustic signal, the ratios of resonances across the pharyngeal, oral, and

---

[1] https://acpa-cpf.org/speech-samples/

nasal cavities. The clinician then maps the perceived ratios to equal-interval or visual-analog scales of hypernasality. However, this percept is vulnerable to other co-modulating variables such as the words being spoken, the quality and loudness of the voice, audible turbulence and escape of air through the nose (nasal emission), the severity of other speech errors, and the idiosyncratic shape of an individual's resonating cavities [10], [11]. This results in a highly nonlinear mapping between the percept and the actual acoustic nasal resonance and considerable inter-rater and intra-rater variability in the assessment. Fundamentally, this limits the reliability and validity of the ratings obtained from untrained clinicians [12]. The Americleft Speech Project was developed to address this dilemma by facilitating inter-center collaborations for speech outcomes research [9]. The first step included the development of a standardized protocol and calibration of craniofacial speech-language pathologists (SLPs) on perceptual ratings of hypernasality. Over the study period, recalibration was required to maintain high levels of inter-rater reliability. To date, only a small number of clinicians have participated in this program. While such programs may be able to be carried in the United States (US) and other developed countries, it is less tenable in countries without any or few craniofacial trained SLPs.

In this paper, we present the objective hypernasality measure (OHM) to assess hypernasality in CP speech and show that it tracks with the clinical perception of Americleft-trained SLPs. When clinicians make judgements of hypernasality, they focus on specific acoustic cues that are hallmarks of hypernasal speech. Similarly, we design an automatic assessment tool based on deep learning and demonstrate that the learned acoustic features from speech correlate with the clinical ratings of hypernasality.

A. Related work

The development of a speech technology-based system involves the extraction of acoustic features which reflect abnormal nasal resonance present in the hypernasal speech. Spectral measures, such as addition of an extra nasal formant around 250 Hz, increased spectral flatness, reduced first formant amplitude, voice low-to-high tone ratio, and vowel space area have previously shown a correlation with the perceived hypernasality [13]–[19]. Acoustic features in combination with the machine learning algorithms have been used to develop automatic hypernasality assessment systems. Mel-frequency cepstral coefficients (MFCCs), jitter, shimmer, vowel space area, wavelet transform-based features have been used to train classifiers (e.g. support vector machine (SVM), Gaussian mixture model (GMM)) that detect hypernasal speech [17], [18], [20], [21]. Recently, convolutional neural networks and recurrent neural networks have also been used for the same purpose [22], [23].

Most of these automatic algorithms for detection of hypernasality were developed in a binary classification setting, i.e., healthy vs. hypernasal speech [17], [18], [20]–[23]. This is inconsistent with clinical practice, where clinicians require more fine-grained information (e.g. evaluation of hypernasality on a scale that ranges from normal to severe) for decision making. For example, a secondary surgery may only be required for treating moderate-severe hypernasal cases [24]. There is a limited number of multi-class classification [25]–[27] and regression-based approaches for predicting hypernasality severity [28], [29]. These approaches rely on analysis of sustained phonations or segmented phonemes from

utterances. This is limiting in two ways: (1) Sustained phonations don't capture phonetic context and don't provide a reliable percept of hypernasality. That is why clinicians prefer to use connected speech for reliable estimation of hypernasality severity [30]. (2) For approaches that rely on connected speech, the phonetic segmentation was achieved either by manual marking or forced-alignment using orthographic transcriptions. However, orthographic transcription is time consuming and the forced alignment procedure may be prone to errors for children's speech [31].

Hypernasality estimation based on supervised machine learning requires large, labeled speech corpora. Most of the existing speech-based hypernasality evaluation methods use speech samples and corresponding perceptual ratings to train machine learning models, including *k*-nearest neighborhood classifier [27], GMMs [26], SVMs [20], [21], [25], [32], and deep neural networks (DNNs) [22], [23], [33]. The performance of these systems critically depends on the availability of clinical hypernasal speech databases that include speech samples from patients and corresponding hypernasality ratings from trained SLPs. However, development of a large hypernasal speech database is difficult in practice due to the limited availability of patients' speech and the associated SLP clinical ratings. As a result, the models run the risk of overfitting to a particular database and rating scale.

B. The proposed approach

While large-scale databases of CP speech are untenable, healthy speech provides us with clues as to the acoustic manifestation of hypernasality. For example, the voiced sounds /m/ and /n/, and the sounds that precede and follow them, require opening of the velopharyngeal port to shunt the vibrating column of air through the nasal cavity. Thus, /m/ and /n/ are nasal consonants (NC). Because the velum is a relatively sluggish articulator in comparison with the tongue and lips, the velopharyngeal port opens and closes more slowly, creating nasalization of vowels adjacent to the NCs, or nasalized vowels (NV). For example, the vowel /æ/ is nasalized in the word "man".

This is in contrast to production of the oral consonants (OC), which involves closure of the velopharyngeal port to impound oral air pressure that creates a burst upon release of the articulatory closure (plosive). The voiced stop consonants, /b/ and /d/, and unvoiced stop consonants, /p/ and /t/, share the exact same places of articulation as /m/ and /n/, respectively, but are completely orally produced. This means that vowels (oral vowels, OV) adjacent to these OCs are also not nasalized, as in the /æ/ in the word "cat". Since, the effect of nasalization is evident in healthy speech, the acoustic manifestation of the velopharyngeal port opening can be modelled using a healthy speech corpus. Compared to CP speech, there are a large number of healthy speech corpora available in the public domain. In our algorithm, we make use of a publicly-available healthy speech corpus and train a nasality feature extraction model using only healthy speech. This results in an objective measure of hypernasality that can be computed frame-by-frame and aggregated at the level of an utterance or speaker.

An overview of the proposed algorithm for estimating the OHM is described in Fig. 1. To learn the acoustic manifestation of the velopharyngeal port opening, we utilize 100 hours of speech from a publicly-available Librispeech corpus to train a DNN model that classifies

among NC, OC, NV, and OV. Training the DNN to classify among these classes forces it to learn the acoustic manifestation of an open velopharyngeal port. As a result, we refer to this DNN as the nasality model and use the four DNN posteriors of this model as "features" for assessing the presence of nasalization in CP speech.

The input children's speech was pre-processed and we extracted the four posterior features using the pre-trained DNN nasality model. These features were combined to derive the OHM for each speech utterance. The details of the algorithm can be found in the Methods section. We established the construct validity of the OHM through several experiments using cleft speech samples and gold-standard clinical ratings from the Americleft project; then we evaluated the external validity of the model using data from the New Mexico cleft palate center (NMCPC) database.

II. DATABASES

The details of the healthy speech corpus and the two CP speech databases are described below.

A. Healthy speech corpus

One hundred hours of healthy speech from the Librispeech database (*train-clean-100*) was used to train the DNN [34]. The database contains English read speech samples recorded from 251 healthy adult speakers (125 male and 126 female). In addition to the speech samples, the database also contains orthographic transcriptions for each read sentence. A separate test set (*test-clean*) consisting of 5.4 hours of speech was used as a validation set.

B. Americleft database

The Americleft database was collected as a part of the Americleft Speech Project at the University of Utah. The database consists of 60 children with CP (37 boys and 23 girls) of average age 6.276 ± 0.676 years. The control group consisted of 10 typically developing children (6 boys and 4 girls) with typically-developing speech characteristics (as determined by an SLP) having an average age of 5.912 ± 0.593 years. The recorded stimuli was comprised of 24 sentences containing different target consonants [9]. The Americleft database was used with an approval from Institutional Review Board (IRB) with IRB ID: STUDY00008224 and written consent was obtained from all participants.

The hypernasality of the recorded speech samples was perceptually evaluated by 4 SLPs from the Americleft speech outcomes group (ASOG) according to a standardized protocol [9]. The speaker-level hypernasality was rated on the Americleft Speech Protocol scale on a 5-point scale (0-normal, 1-borderline, 2-mild, 3-moderate, 4-severe) [9]. The Pearson correlation coefficient was computed between different pairs of raters to evaluate the inter-rater reliability. The average inter-rater correlation coefficient was found to be 0.797 ± 0.079. The ratings for all 4 SLPs were averaged to obtain a single "ground-truth" rating per speaker [35]. The histogram in Fig. 2(a) shows the distribution of hypernasality ratings for the 70 speakers from the Americleft database. It is clear from the figure that the database is skewed towards the normal ('0') end of the scale. We balance the original Americleft data by randomly removing a subset of speakers rated with normal nasality. The balanced

Americleft database, Americleft(38), is comprised of 38 speakers and the histogram of the ground-truth ratings is shown in Fig. 2(b). The average inter-rater correlation for Americleft(38) was 0.776 ± 0.068.

In addition to hypernasality, the Americleft samples were evaluated for consonant production errors and audible nasal emission (ANE). To assess consonant production errors, the Americleft sentences were phonetically transcribed by the four Americleft raters using the International Phonetic Alphabet. Based on the transcriptions, the consonant production errors are grouped into 4 groups as per the Americleft speech protocol [9]. These groups are anterior oral (dentalization, lateralization, palatalization), posterior oral (double articulation, backed to velar/uvular), non-oral (pharyngeal articulation, glottal articulation, nasal fricatives, double articulation), and passive (weak and/or nasalized consonants, and nasal realization of plosives) articulation errors. The consonant production errors were evaluated for 24 Americleft sentences, and the percentage of anterior oral, posterior oral, non-oral, and passive articulation errors are computed at a speaker-level. The ANE was rated on a 3-point rating scale (0-normal, 1-occasionally/seldom noted, and 2-frequently noted) by 4 SLPs and the average inter-rater correlation for Americleft(38) was found equal to be 0.694 ± 0.164.

C. The NMCPC database

The NMCPC database is described in [32]. The database is comprised of speech samples from 10 controls (8 boys and 2 girls) and 41 children with CP (22 boys and 19 girls) with an average age of 9.2±3.3 years. All participants spoke American English. Each child was asked to repeat a random subset of sentences selected from a larger set of 76 sentences. The number of sentences per participant ranged from 7 to 69. The recorded samples were perceptually evaluated by 5 listeners on a continuous scale ranging from 0 to 3, where 0 stands for normal and 3 for severe hypernasality. The raters were not speech language pathologists, but they were speech processing experts. Initially, a few NMCPC samples were evaluated by expert SLPs and used as reference exemplars by the speech processing experts. By listening to these examples, the researchers were self-trained to evaluate the complete NCMPC database. The average inter-rater correlation of this database was 0.872 ± 0.078. The ratings of 5 raters were averaged to obtain a single "ground-truth" rating per speaker [35]. A histogram of average ratings is shown in Fig. 2(c). This database is nicely balanced across the different hypernasality levels. The NMCPC database is a publicly available database that can be acquired upon request to Dr. Luis Cuadros, New Mexico Cleft Palate Center.

III. METHODS

A. DNN Nasality Model

Below we describe the development of the DNN nasality model, including its architecture and the training procedure.

DNN model architecture: The architecture of the DNN nasality model is shown in Fig. 3. The model has an input layer with 39-nodes, corresponding to the 39-dimensional MFCCs

input speech feature. The model is comprised of 3-hidden layers, where each layer has 1024 hidden neurons with rectified linear unit (ReLU) activation. The output layer consists of 4 softmax nodes, each interpreted as a posterior probability corresponding to NC, OC, NV, and OV classes.

Training the DNN: First, the 100 hours of healthy speech recordings from the Librispeech corpus and the corresponding orthographic transcriptions were passed through the Montreal forced-aligner [36] to align the speech acoustics to the transcript at the phoneme level. The segmented phonemes were grouped into the four classes of interest: nasal consonants (NC), oral consonants (OC), nasalized vowels (NV), and oral vowels (OV). The NC group was formed by combining across nasal consonants (/n/, /m/, and /ng/); the OC group was formed by combining across oral consonants, including plosives (/b/, /d/, /g/, /p/, /t/, /k/), fricatives (/z/, /v/, /s/, / ʃ /, /f/, /h/, /θ/, /ð/), affricates (/tʃ/, /dʒ/), glides and liquids (/l/, /ɹ/, /w/). The NV group was formed by combining thirty percent of the vowel segments that follow and precede a nasal consonant. The OV group was formed by combining across the remaining vowels segments, which were not surrounded by nasal consonants. An example grouping of phonemes in a healthy speech sample is illustrated in Fig. 4. The speech waveform corresponding to the phrase "no one who had ever seen" and its spectrogram are shown in Fig. 4(a) and (b), respectively. Fig. 4(a) shows the time boundaries for each phoneme in ARPABET form. Based on the velopharyngeal activity, the phonemes are grouped into NC, OC, NV, and OV categories. In the example shown in Fig. 4, the nasal consonant (/n/) and the vowels (/o/, /ɐ/, /i/) surrounding it are grouped into NC and NV classes, respectively. The oral consonants (/w/, /h/, /d/, /v/, /s/) and the vowels (/u/, /æ/, /ɚ/) adjacent to them are grouped as OC and OV classes, respectively.

The input speech to the DNN was sampled at a 16 kHz sampling rate, and short-time processed using a 20 ms Hamming window with 10 ms overlap. From each frame, 13 dimensional MFCCs, velocity ( ) and acceleration (   ) coefficients were computed. This 39-dimensional feature vector was the input to the DNN nasality model; the label for each 20 ms frame corresponded to the category to which that frame belonged to. The classifier was trained to classify between the four phoneme categories described above. The error between the predicted and ground truth labels was computed using a categorical cross-entropy loss function. The ADAM optimizer was used to estimate the optimum parameters of the network. The network was trained for 25 epochs with a learning rate of 0.001. The MFCC features were computed using the Librosa package in Python and the DNN was implemented using the Keras 2.2.4 toolkit with a TensorFlow 1.13.1 backend.

DNN posteriors as nasality features: For the given input speech frame, the DNN results in 4 posterior probabilities corresponding to the NC, OC, NV, and OV classes. As described in Fig. 5, increased values of $P(NC)$ and $P(NV)$ indicate the presence of nasalization in consonants and vowels, respectively. Hence, we consider these posteriors as the nasality features and use them to compute an objective measure of hypernasality.

Evaluating the OHM: Evaluating the OHM requires the pre-trained DNN nasality model described above. The DNN was trained on adult speech, however we aim to use it to evaluate hypernasality in children's speech. To compensate for the acoustic mismatch

between children and adult speech, we used the pitch modification algorithm proposed by [37]. The pitch modification pre-processing step lowers the pitch and speaking rate of children per the details in the paper. This same approach was used to improve the performance of speech recognition algorithms trained on adult speech and evaluated on children speech [37].

The pitch-modified speech signal was resampled at 16 kHz and short-time processed using a 20 ms Hamming window with 10 ms overlap. The frame-size and frame-shift were consistent with the parameters used during DNN nasality model training. As before, a 39-dimensional MFCC feature vector was computed and provided as input to the pre-trained DNN nasality model. The 4 DNN posterior probabilities were computed for each frame of children's speech. The DNN posteriors obtained for the pre-processed children's speech were used to compute the OHM. Let, x_i be the feature vector corresponding to the MFCC input features for the ith frame; the DNN outputs probabilities corresponding to NC, OC, NV, and OV classes for that frame, i.e., $P(NC|x_i)$, $P(OC|x_i)$, $P(NV|x_i)$, and $P(OV|x_i)$, respectively. Then the objective hypernasality measure $OHM(x_i)$ for i^{th} frame is computed as

$$OHM(x_i) = max\left(log\left(\frac{P(NC \mid x_i)}{P(OC \mid x_i)}\right), log\left(\frac{P(NV \mid x_i)}{P(OV \mid x_i)}\right)\right). \tag{1}$$

In equation 1, we compute the ratios of posterior probabilities of nasal to oral consonants and nasalized to oral vowels. The frame-level OHM, $OHM(x_i)$, is built by logarithmically transforming each ratio and taking the maximum across the two.

For clarity, it is useful to demonstrate the OHM with two examples. The speech waveform, frame-wise DNN nasality posterior features, and the OHM contours from a control sample and a CP sample are plotted in Fig. 5(a)–(h). The target sentence is "buy baby a bib". Since the sentence does not contain any nasal consonants, no nasal cues are expected in the speech from the control group; this is consistent with panels (b) and (c) where $P(OC) > P(NC)$ and $P(OV) > P(NV)$ for the speech from the control group. For the case of hypernasal speech from the CP group in panels (f) and (g), we see that $P(NC) > P(OC)$ and $P(NV) > P(OV)$. Although the target text does not contain any nasal consonants, large $P(NC)$ and $P(NV)$ values indicate the presence of abnormal nasal resonances in CP speech indicative of hypernasality. As expected, the OHM measure obtained by combining $P(NC)$, $P(OC)$, $P(NV)$, and $P(OV)$ indicates relatively higher values for hypernasal speech (Fig. 5(h) than for healthy speech (Fig. 5(d)).

The frame-level OHM scores ($OHM(x_i)$) were averaged over all frames of a given utterance to obtain sentence-level OHM scores. Similarly, sentence-level OHM scores were averaged over all utterances spoken by the same speaker to obtain speaker-level OHM scores.

IV. EXPERIMENTS AND RESULTS

The experiments were conducted using the Americleft database to evaluate the sentence and speaker-level performance of the OHM, the robustness of OHM to co-existing speech production errors, the sensitivity of OHM, and the internal reliability of OHM. The external

validity of the OHM was evaluated using the NMCPC database. Further, a comparison of the OHM with a representative supervised learning method based on DNN regression was also conducted. The details of the experiments and the results are presented in the following subsections.

A. Validation of sentence-level OHM scores

The frame-level results in Fig. 5 are averaged over the entire duration of the utterance to generate a sentence-level OHM score. The correlation between the sentence-level OHM and the speaker-level perceptual ratings, the ground-truth rating obtained from Americleft-trained SLPs, was evaluated using Pearson's correlation coefficient (r). In Table I we list the sentences from the Americleft database, grouped by target consonant category; for each sentence, we also list the correlation between the sentence-level OHM and the perceptual hypernasality level. As expected, oral sentences, i.e., the sentences containing oral consonants (plosives, fricatives, affricates, liquids, and glides) showed a high correlation with the perceptual ratings; whereas the OHM calculated from nasal sentences shows a low correlation. This makes sense as the OHM demonstrates a ceiling effect for nasal sentences since it is expected that they are nasalized.

B. Validation of speaker-level OHM scores

We compute the speaker-level OHM scores by averaging across all oral sentences produced by a speaker. The average Pearson correlation between the speaker-level OHM with each rater's perceptual ratings *and* the average inter-rater Pearson correlation are shown in Table II.

For finer-grained analysis, we compare the OHM with the ground-truth rating obtained by averaging the clinical ratings from the 4 raters. Fig. 6 shows the relationship between the speaker-level OHM and perceptual ratings for the Americleft database. The OHM shows a significant correlation ($r = 0.797$, $p < 0.001$) with the ground truth perceptual ratings. A scatter plot of sample-level data is shown in Fig. 6.

C. Robustness to co-existing speech production errors

In addition to hypernasality, CP speech is also characterized by the presence of consonant production errors and ANE; therefore, we also analyzed the relationship of OHM with these co-existing speech production errors.

The bar plot in Fig. 7 shows the correlation of the perceptual hypernasality ratings and OHM with respect to 4 categories of consonant production errors, namely, anterior oral, posterior oral, non-oral, and passive errors. The perceptual hypernasality showed a moderate correlation ($r = 0.460$, $p < 0.05$) with respect to passive errors. Passive errors include nasalized consonants, which carry nasal cues, therefore, it is expected that the presence of passive errors increases with increased hypernasality severity. In fact, the perception of nasalized consonants was considered an important criterion in developing the hypernasality rating scale in [30]. Nasal resonances are not evident in anterior, posterior and non-oral errors. Hence, the perceptual hypernasality ratings showed a low correlation with the

Americleft ratings for anterior ($r = 0.166$, $p = 0.201$), posterior ($r = 0.151$, $p = 0.211$) and non-oral errors ($r = 0.169$, $p = 0.191$).

Similar to the perceptual hypernasality ratings, the OHM scores also showed a moderate correlation ($r = 0.478$, $p < 0.05$) with passive errors and a very low correlation with the posterior ($r = -0.097$, $p = 0.559$) and anterior errors ($r = -0.060$, $p = 0.718$). However, when compared to perceptual ratings, the OHM showed a relatively lower correlation for posterior and anterior and higher for passive errors. The OHM showed a slightly higher correlation for non-oral errors ($r = 0.268$, $p = 0.102$).

The perceptual ANE ratings showed a low correlation ($r = 0.393$, $p < 0.05$) with the perceptual hypernasality ratings. The OHM scores also resulted in a low correlation with the perceptual ANE ratings ($r = 0.287$, $p = 0.081$). These results provide additional evidence that the OHM captures hypernasality as it correlates with the passive errors carrying nasal cues and is largely robust to co-existing anterior and posterior consonant production errors.

D. Evaluating the sensitivity of the OHM

In our analysis, we considered a balanced set of 38 speakers in the Americleft database to evaluate the sentence-level and speaker-level OHM scores relative to the perceptual ratings. Additionally, we analyzed the OHM for the remaining 32 'CP speakers rated as normal' (no hypernasality) and compared them with the control group. Here, the 'CP rated as normal' corresponds to speakers whose SLP hypernasality rating was considered normal. Fig. 8 shows the range of the OHM for the two groups. The OHM scores of speakers with CP rated '0' were greater than that of controls. A *t*-test reveals a statistically significant difference between the two groups ($t = -2.899$, $p < 0.05$). This result may indicate the presence of very mild hypernasality in the CP group not detected by the SLPs.

E. Assessing the internal reliability of the OHM

To evaluate the internal reliability of OHM scores, we grouped the 20 oral sentences from the Americleft database into set-1 and set-2, where set-1 contains the first 10 sentences and the remaining 10 formed set-2. We computed speaker-level OHM scores for set-1 and set-2 data. Fig. 9 shows the scatter plot of OHM scores for set-1 vs. set-2. The OHM scores of set-1 are significantly correlated ($r = 0.898$, $p < 0.001$) with that of set-2.

F. Assessing the external validity of the OHM

We also evaluated the performance of the OHM on the NMCPC database in order to evaluate how well the OHM generalizes to data collected in other studies. The hypernasality level of the NMCPC speech samples was evaluated by 5 raters. The correlation of the OHM with respect to each rater and the inter-rater correlation are shown in Table III. The average correlation of the OHM vs. individual raters was found equal to $r = 0.695$, $p < 0.001$ whereas the inter-rater correlation was $r = 0.872$, $p < 0.001$.

The scatter plot of the speaker-level OHM vs. the average of the 5 clinical ratings is shown in Fig. 10. The OHM showed a significant correlation ($r = 0.713$, $p < 0.001$) with the average hypernasality ratings provided by the SLPs.

G. Comparison with a fully supervised approach

Conventionally, automatic evaluation of hypernasality involves the supervised training of models like SVMs [20], [21], artificial neural networks [23], and recurrent neural networks [22] for the binary classification of healthy and hypernasal speech samples. In all of these existing approaches, the supervised training of models was carried out by using a perceptually labeled CP speech database. In the proposed approach, the OHM was computed using the nasality DNN, whose parameters were estimated by using only the healthy speech samples. Most of the existing deep learning-based implementations [22], [23] were developed for binary classification from segmented vowels. We found only one DNN-based method in the literature [33] which predicts hypernasality severity from connected speech samples. In this work [33], the DNN was trained directly on a labeled CP speech database. To compare the OHM with conventional supervised approaches, we implemented a DNN regressor, which was directly trained on the MFCC features extracted from the speech samples and the perceptual ratings of the Americleft database.

We used only oral sentences to train and test the DNN regressor using leave-one-speaker-out (LOSO) cross-validation. The sample size of the Americleft database (20 oral sentences x 38 speakers=760) is very small to train a DNN. To address this issue, we used data augmentation using (a) addition of noise: white, babble, and factory noise with 5, 10, 15, and 20 dB SNRs, (b) speaking rate modification using the factors 0.8, 0.9, 1.1, and 1.2, (c) vocal tract length perturbation (VTLP) using the perturbation factors 0.9, 0.95, 1.05, and 1.1 [38]. After the data augmentation, the sample-size of the database was increased from 760 to 9120 sentence-level recordings.

The 39-dimensional MFCC features extracted using 20 ms Hamming windowed speech frames with a shift of 10 ms were fed to a feed-forward DNN regressor. The architecture details of DNN regression are as follows: 39 input nodes, three hidden layers with 512 neurons with ReLU activation, and 1 output node with a linear activation function. The error between the predicted outputs and ground truth labels was estimated using the mean squared error (MSE) loss function. The ADAM optimizer was used to estimate the optimum parameters of the network. The network was trained for 25 epochs, with a learning rate of 0.001.

The MFCCs were computed at the frame-level, but the ground truth was available at the speaker-level. During training we assigned the speaker-level hypernasality ratings to every frame-level feature vector belonging to that particular speaker. In the testing phase, speaker-wise averaging of DNN outputs was carried out to obtain a single score per speaker. The performance of the DNN regressor was evaluated using the LOSO cross-validation criteria. In LOSO cross-validation, the samples of all speakers except one speaker were used to train DNN and the remaining one's sample was considered for the testing. Note that the augmented samples were used only during the training phase, while for testing we only used the original samples.

The correlation coefficient computed between predicted scores and the perceptual ratings is shown in Table IV. The correlation was found to be statistically significant ($r = 0.524$, $p < 0.05$), but it is well below that of the OHM. The DNN regressor trained on the Americleft

samples was also used to evaluate hypernasality using the NMCPC samples and the results are presented in Table IV. The predicted scores for the NMCPC database showed a weak correlation ($r = 0.301$, $p = 0.032$) with the perceptual ratings. These results indicate possible overfitting. The OHM scores were more strongly correlated for both databases and the results empirically show that the OHM was robust to a variety of recording conditions, sentence contexts, and perceptual rating scales.

V. DISCUSSION

In this work, we introduced an objective measure of hypernasality based on a DNN nasality model trained on healthy speakers with no clinical labels. First, we modeled the acoustic cues related to an open velum (nasalized) from healthy speech by training a DNN classifier to classify between NC, OC, NV, and OV classes. This pre-trained DNN on healthy speech samples was used to characterize the presence of abnormal nasal resonances in the speech of children with CP. The OHM was computed at the level of a speech frame and aggregated into sentence-level and speaker-level scores.

Comparison with existing hypernasality evaluation approaches:

Most of the existing approaches for the automatic hypernasality evaluation in children with CP were developed to classify between normal and hypernasal speech [20], [23] or multi-class classification (normal, mild, moderate, and severe) [25], [32]. Since these supervised machine learning models were directly trained on CP speech samples and perceptual hypernasality ratings, their performance critically depends on the availability of labeled hypernasality speech databases. To compare to the OHM, we trained a regressor directly on the speech samples and ratings from the Americleft database. Even though we increased the database size via data augmentation, the DNN regressor's performance was well below that of the OHM. The availability of a perceptually labeled database became a practical limitation in clinical speech research due to the limited number of subjects and Americleft SLPs. The OHM bypasses the need for clinical labels as the scores are estimated from a DNN trained on a publicly available large healthy speech corpus.

We recognize that this is only a proxy for hypernasality as the manifestation of nasalization in healthy speech does not perfectly translate into the same pattern as clinical hypernasality (which refers to the perception of excessive nasal resonance during vowels and voiced sounds). Nevertheless, our results indicate that this seems to be a more robust measure of hypernasality when compared to methods based on supervised learning.

Another important advantage of OHM is that the approach does not require phonetic segmentation. Most of the automated hypernasality evaluation methods rely on the phonetic segmentation, where segmentation is performed either by manual labeling or forced-alignment using orthographic transcriptions. The OHM is computed directly on connected speech samples and this makes the evaluation process faster as manual labeling and speech transcription are time-consuming processes.

The role of the speech stimuli in estimating hypernasality:

The choice of stimuli or target sentence plays an important role in the assessment of clinical hypernasality. In the case of healthy speakers, nasal resonance is inherently present during the production of nasal sentences and completely absent in the case of oral sentences. The presence of non-negligible nasal resonance during the production of oral sentences is considered to be abnormal, which indicates the presence of hypernasality [30]. Sentence-wise correlation values listed in Table I revealed that the proposed measure yields a good correlation for the oral sentences and poor correlation for nasal sentences. Since, for both CP subjects and controls, the acoustic energy passes through the nasal tract while producing nasal sentence, it is difficult to evaluate hypernasality using these sentences. These results indicate that to reliably compute the speaker-level OHM, the target speech samples should only contain oral consonants. These results closely match with the perceptual assessment guidelines mentioned in [30], where only sentences with oral consonants were suggested for assessing hypernasality.

High-pressure and low-pressure consonants play an important role in clinical evaluation of hypernasality [24], [30]. Pressure-sensitive or high-pressure consonants (plosives: /t/, /k/, /k/, /b/, /d/, /g/, fricatives: /h/, /f/, /ʃ/, /z/, /θ/, /ð/, and affricates: /tʃ/, /dʒ/) require generation adequate intra-oral pressure, which is developed by the closure or significant constriction of oral tract and closure of the velopharyngeal port. Whereas, low-pressure consonants (glides: /y/, /w/ and liquids: /ɹ/, /l/) do not require high intra-oral pressure. Since the loss of airflow in patients with CP and VPD affects the ability to build up and maintain intra-oral air pressure, the production of high pressure consonants is severely affected. The substitution of nasal consonants for target high pressure consonants is most commonly reported in speakers with CP because of the escape of air through the incompletely closed VP port . Therefore, in clinical settings, target speech containing high-pressure oral consonants is highly recommended for the perceptual assessment of hypernasality [24], [30]. Our results are consistent with this as the OHM shows higher correlation for sentences containing high-pressure consonants (plosives, fricatives, and affricates) than those with low-pressure consonants.

Co-existing speech production errors in assessment of hypernasality:

Articulation errors in speakers with CP are broadly divided into active and passive errors. In the case of passive articulation errors, the speaker tries to retain the place of articulation of the target phoneme but the air escapes through the velopharyngeal port. Hence, the consonant is perceived to be weak or nasalized. In the case of severe VPD, the target consonant is completely replaced by a nasal consonant (/t/, /d/→/n/ and /p/, /b/→/m/) [30].

Active articulation errors can be further divided into different subgroups. In the case of non-oral (also known as compensatory errors), a speaker attempts to compensate for the effect of reduced oral air pressure by shifting the location of articulatory constriction. The shift in the place of articulation may be within the oral or non-oral cavity. In many cases, the presence of VPD leads to a shift towards glottal and pharyngeal regions [30]. Although the consonant production errors and ANE are produced as a consequence of VPD, their contribution in the perception of hypernasality is different. The passive errors contain nasal

cues and hence, their presence contributes to the perception of hypernasality. In contrast, the anterior, posterior, non-oral substitutions (glottal and/or pharyngeal substitutions), and ANE do not carry nasal cues; therefore their contribution is relatively low to the perception of hypernasality. However, the presence of active errors can create variability in perceptual evaluation as it is difficult to perceptually decouple co-existing errors from the presence of hypernasality. As shown in Fig. 7, the OHM scores do not have this bias for anterior and posterior oral errors. However, the OHM showed some relationship with non-oral errors and ANE. This result provides evidence that OHM scores capture the nasal cues present in the speech signal and is robust to most co-existing active errors in CP speech.

Differences in OHM performance on Americleft and NM-CPC databases:

For the Americleft database, the inter-rater reliabilities are on average a little higher than the OHM-torater reliabilities, but they are within 1 standard deviation of each other. For the NMCPC database, inter-rater reliabilities are significantly greater than the OHM-to-rater reliabilities. Also, the OHM scores showed relatively higher correlation with the averaged perceptual ratings ($r = 0.797$), when compared to the NMCPC database ($r = 0.713$). The reasons for this difference in the OHM's performance are multifold. The NMCPC database is not balanced in terms of the number of sentences per speaker. This imbalanced nature of the NMCPC database affects the estimation of speaker-level OHM, where the speaker-level OHM was computed by averaging the sentence-level scores. The average signal-tonoise ratio (SNR) of the Americleft samples was found equal to 23.88 ± 9.10 dB. Whereas the speech samples of NMCPC database are noisier with an SNR of 15.95 ± 8.10 dB. The raters of the Americleft database were trained under the Americleft speech project and followed the same protocol [9]. In the Americleft speech project, a group of SLPs with extensive expertise in the evaluation and treatment of children with CP developed a standard hypernasality rating scale and protocols for speech evaluation. The OHM showed good agreement with ratings that followed these guidelines, but a lower agreement with the NMCPC raters who were not trained in the Americleft standard protocols. Moreover, the NMCPC raters were speech researchers and not SLPs with clinical expertise in the domain of CP speech assessment. However, it is interesting that the NMCPC raters have a much higher inter-rater reliability than the Americleft ratings, well above what has been reported in the literature previously [9]. We posit that this is because the raters discussed their ratings while evaluating the speech samples, reaching consensus in some cases. Another possible reason for reduced inter-rater reliability in the Americleft database may be the use of a different rating scale. The NMCPC speech samples were rated on a 4-point rating scale, whereas a 5-point rating scale was used in the Americleft database. This may impact the inter-rater reliability.

VI. SUMMARY AND FUTURE WORK

In this work, we introduced an objective measure of hypernasality based on a DNN nasality model trained on healthy speakers with no clinical labels. First, we modeled the acoustic cues related to an open velum (nasalized) from healthy speech by training a DNN classifier to classify among NC, OC, NV, and OV classes. This pre-trained DNN on healthy speech samples was used to characterize the presence of abnormal nasal resonances in the

speech of children with CP. The OHM was computed at the level of a speech frame and aggregated into sentence-level and speaker-level scores. After optimizing the algorithm, a user-friendly interface will be developed for in-clinic use. The nasality DNN was trained for the acoustic difference between nasal and oral sounds and there were no language specific features used while training. Hence, we expect that the model is not sensitive to different languages. However, the robustness of the system to different languages needs to be evaluated prospectively.

In the current implementation of the OHM, the model is completely tuned and implemented using only the healthy speech database. Future work can focus on further refinement of the implementation by using a corpus of cleft speech to tune model parameters, or perhaps to change it to a supervised model by using linear regression across different sentences to produce speaker-level hypernasality scores. The present work uses a pitch modification algorithm to compensate for the difference between adult and children's speech. However, the pitch modification algorithm does not address all possible differences between adult and children's speech. For example, differences in co-articulation or duration are not accounted for. Hence, future work could focus on better speaker adaptation methods such as identity vector (i-vector) and transfer learning approaches. Another limitation of the current work is that the algorithm assumes the presence of information related to nasality over the entire duration of utterance and simply averages the frame-level OHM scores over an entire utterance. However, depending on the severity level, the hypernasality information may be distributed unevenly over different phonemes. Therefore, instead of a simple averaging operation, recurrent neural networks and attention models can be used to capture unevenly distributed nasality information.

Furthermore, hypernasality is not only specific to CP speech. It is also present in speech from individuals with neurological disorders, such as Huntington's and Parkinson's disease. Therefore, future work can focus on extending the validation of this model for evaluating hypernasality in dysarthric speech.

ACKNOWLEDGEMENTS

REFERENCES

[1]. Mossey PA, Catilla EE et al., "Global registry and database on craniofacial anomalies: Report of a WHO registry meeting on craniofacial anomalies," 2003.

[2]. Kummer AW, Cleft palate & craniofacial anomalies: Effects on speech and resonance. Nelson Education, 2013.

[3]. Hardin-Jones MA and Jones DL, "Speech production of preschoolers with cleft palate," The Cleft Palate-Craniofacial Journal, vol. 42, no. 1, pp. 7–13, 2005. [PubMed: 15643919]

[4]. Chapman KL, "The relationship between early reading skills and speech and language performance in young children with cleft lip and palate," The Cleft Palate-Craniofacial Journal, vol. 48, no. 3, pp. 301–311, 2011. [PubMed: 20815721]

[5]. Bell JC, Raynes-Greenow C, Turner R, Bower C, Dodson A, Nicholls W, and Nassar N, "School performance for children with cleft lip and palate: A population-based study," Child: Care, Health and Development, vol. 43, no. 2, pp. 222–231, 2017.

[6]. "Parameters for evaluation and treatment of patients with cleft lip/palate or other craniofacial differences," The Cleft Palate-Craniofacial Journal, vol. 55, no. 1, pp. 137–156, 2018. [Online]. Available: 10.1177/1055665617739564 [PubMed: 34162066]

[7]. Bettens K, Wuyts FL, and Van Lierde KM, "Instrumental assessment of velopharyngeal function and resonance: A review," Journal of Communication Disorders, vol. 52, pp. 170–183, 2014. [PubMed: 24909583]

[8]. Watterson T, "The use of the nasometer and interpretation of nasalance scores," Perspectives of the ASHA Special Interest Groups, vol. 5, no. 1, pp. 155–163, 2020.

[9]. Chapman KL, Baylis A, Trost-Cardamone J, Cordero KN, Dixon A, Dobbelsteyn C, Thurmes A, Wilson K, Harding-Bell A, Sweeney T et al. , "The americleft speech project: a training and reliability study," The Cleft Palate-Craniofacial Journal, vol. 53, no. 1, pp. 93–108, 2016. [PubMed: 25531738]

[10]. Baylis A, Chapman K, and Whitehill TL, "Validity and reliability of visual analog scaling for assessment of hypernasality and audible nasal emission in children with repaired cleft palate," The Cleft Palate-Craniofacial Journal, vol. 52, no. 6, pp. 660–670, 2015. [PubMed: 25322442]

[11]. Yamashita RP, Borg E, Granqvist S, and Lohmander A, "Reliability of hypernasality rating: Comparison of 3 different methods for perceptual assessment," The Cleft Palate-Craniofacial Journal, vol. 55, no. 8, pp. 1060–1071, 2018. [PubMed: 29634363]

[12]. Sweeney T and Sell D, "Relationship between perceptual ratings of nasality and nasometry in children/adolescents with cleft palate and/or velopharyngeal dysfunction," International Journal of Language & Communication Disorders, vol. 43, no. 3, pp. 265–282, 2008. [PubMed: 17852526]

[13]. Kataoka R, Michi K-I, Okabe K, Miura T, and Yoshida H, "Spectral properties and quantitative evaluation of hypernasality in vowels," The Cleft Palate-Craniofacial Journal, vol. 33, no. 1, pp. 43–50, 1996. [PubMed: 8849858]

[14]. Kent RD, Liss J, and Philips B, "Acoustic analysis of velopharyngeal dysfunction in speech," Bzoch K. Communicative Disorders: related to cleft lip and palate. Boston: Little Brown, pp. 258–70, 1989.

[15]. Vikram CM, Adiga N, and Prasanna SRM, "Spectral enhancement of cleft lip and palate speech." in Interspeech, 2016, pp. 117–121.

[16]. Vijayalakshmi P, Reddy MR, and O'Shaughnessy D, "Acoustic analysis and detection of hypernasality using a group delay function," IEEE Transactions on Biomedical Engineering, vol. 54, no. 4, pp. 621– 629, 2007. [PubMed: 17405369]

[17]. Lee G-S, Wang C-P, and Fu S, "Evaluation of hypernasality in vowels using voice low tone to high tone ratio," The Cleft Palate-Craniofacial Journal, vol. 46, no. 1, pp. 47–52, 2009. [PubMed: 19115797]

[18]. Lee G-S, Wang C-P, Yang CC, and Kuo TB, "Voice low tone to high tone ratio: A potential quantitative index for vowel [a:] and its nasalization," IEEE Transactions on Biomedical Engineering, vol. 53, no. 7, pp. 1437–1439, 2006. [PubMed: 16830950]

[19]. Nikitha K, Kalita S, Vikram C, Pushpavathi M, and Prasanna SM, "Hypernasality severity analysis in cleft lip and palate speech using vowel space area," in Interspeech, 2017, pp. 1829–1833.

[20]. Golabbakhsh M, Abnavi F, Kadkhodaei Elyaderani M, Derakhshandeh F, Khanlar F, Rong P, and Kuehn DP, "Automatic identification of hypernasality in normal and cleft lip and palate patients with acoustic analysis of speech," The Journal of The Acoustical Society of America, vol. 141, no. 2, pp. 929–935, 2017. [PubMed: 28253654]

[21]. Orozco-Arroyave JR, Vargas-Bonilla J, Arias-Londono JD,˜ Murillo-Rendon S, Castellanos-Dom´´ınguez G, and Garces J, "Nonlin-´ ear dynamics for hypernasality detection in spanish vowels and words," Cognitive Computation, vol. 5, no. 4, pp. 448–457, 2013.

[22]. Wang X, Yang S, Tang M, Yin H, Huang H, and He L, "Hypernasalitynet: Deep recurrent neural network for automatic hypernasality detection," International Journal of Medical Informatics, vol. 129, pp. 1–12, 2019. [PubMed: 31445242]

[23]. Wang X, Tang M, Yang S, Yin H, Huang H, and He L, "Automatic hypernasality detection in cleft palate speech using CNN," Circuits, Systems, and Signal Processing, pp. 1–27, 2019.

[24]. Kummer AW and Lee L, "Evaluation and treatment of resonance disorders," Language, Speech, and Hearing Services in Schools, vol. 27, no. 3, pp. 271–281, 1996.

[25]. Dubey AK, Prasanna SM, and Dandapat S, "Detection and assessment of hypernasality in repaired cleft palate speech using vocal tract and residual features," The Journal of the Acoustical Society of America, vol. 146, no. 6, pp. 4211–4223, 2019. [PubMed: 31893680]

[26]. He L, Zhang J, Liu Q, Yin H, and Lech M, "Automatic evaluation of hypernasality and consonant misarticulation in cleft palate speech," IEEE Signal Processing Letters, vol. 21, no. 10, pp. 1298–1301, 2014.

[27]. He L, Zhang J, Liu Q, Yin H, Lech M, and Huang Y, "Automatic evaluation of hypernasality based on a cleft palate speech database," Journal of Medical Systems, vol. 39, no. 5, p. 61, 2015. [PubMed: 25814462]

[28]. Mathad VC, Chapman K, Liss J, Scherer N, and Berisha V, "Deep learning based prediction of hypernasality for clinical applications," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 6554–6558.

[29]. Saxon M, Liss J, and Berisha V, "Objective measures of plosive nasalization in hypernasal speech," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, Brighton, United Kingdom, pp. 6520–6524.

[30]. Henningsson G, Kuehn DP, Sell D, Sweeney T, Trost-Cardamone JE, and Whitehill TL, "Universal parameters for reporting speech outcomes in individuals with cleft palate," The Cleft Palate-Craniofacial Journal, vol. 45, no. 1, pp. 1–17, 2008. [PubMed: 18215095]

[31]. Knowles T, Clayards M, and Sonderegger M, "Examining factors influencing the viability of automatic acoustic analysis of child speech," Journal of Speech, Language, and Hearing Research, vol. 61, no. 10, pp. 2487–2501, 2018.

[32]. Javid MH, Gurugubelli K, and Vuppala AK, "Single frequency filter bank based long-term average spectra for hypernasality detection and assessment in cleft lip and palate speech," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 6754–6758.

[33]. Vikram CM, Tripathi A, Kalita S, and Prasanna SRM, "Estimation of hypernasality scores from cleft lip and palate speech." in Interspeech, 2018, Hyderabad, India, pp. 1701–1705.

[34]. Panayotov V, Chen G, Povey D, and Khudanpur S, "Librispeech: An ASR corpus based on public domain audio books," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, South Brisbane, Australia, pp. 5206–5210.

[35]. Grimm M, Kroschel K, Mower E, and Narayanan S, "Primitives-based evaluation and estimation of emotions in speech," Speech Communication, vol. 49, no. 10–11, pp. 787–800, 2007.

[36]. McAuliffe M, Socolof M, Mihuc S, Wagner M, and Sonderegger M, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in Interspeech, 2017, pp. 498–502.

[37]. Shahnawazuddin S, Adiga N, and Kathania HK, "Effect of prosody modification on children's ASR," IEEE Signal Processing Letters, vol. 24, no. 11, pp. 1749–1753, 2017.

[38]. Lo T, Chao F-A, Weng S, and Chen B, "The NTNU system at the interspeech 2020 non-native children's speech ASR challenge," arXiv preprint arXiv:2005.08433, 2020.
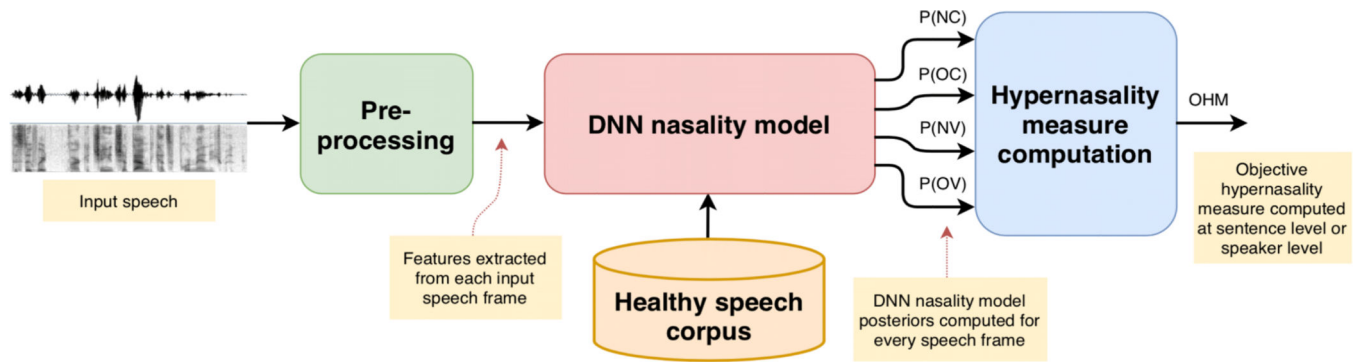
Fig. 1.
Overview of the proposed approach for the hypernasality prediction. First, the input children's speech is pre-processed and passed through the pre-trained DNN nasality model. The DNN model posteriors are combined to form the objective hypernasality measure.
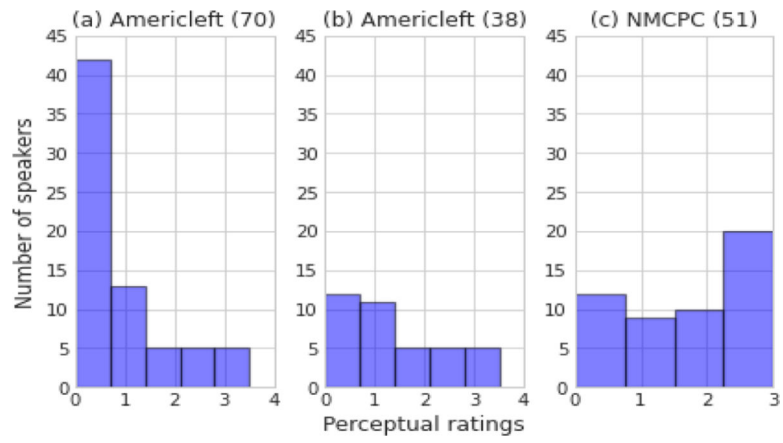
Fig. 2.
The distribution of the number of speakers over different ground-truth hypernasality levels. Histograms of weighted averaged ratings for (a) Americleft (70 speakers), (b) balanced Americleft (38 speakers) and (c) NMCLP (51 speakers) databases.
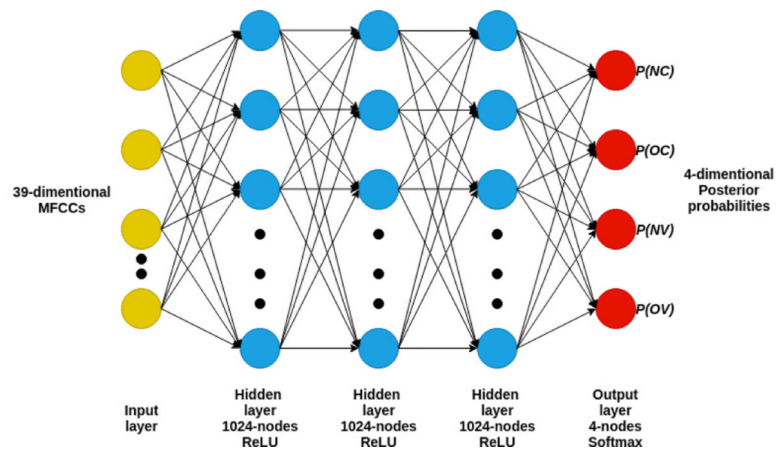
Fig. 3.
The architecture of the DNN nasality model. The model is a feed forward neural network consisting of a 39-dimensional input layer, three hidden layers with 1024 hidden neurons in each layer, and a 4-dimensional softmax output layer. The output layer yields posterior probabilities corresponding to nasal consonants (NC), oral consonants (OC), nasalized vowels (NV), and oral vowels (OV).
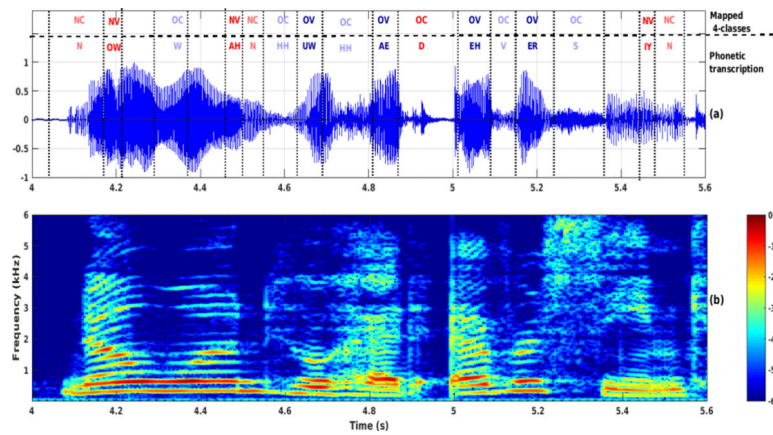
Fig. 4.

An illustration of the phoneme mapping procedure used for DNN training. (a) The speech waveform corresponding to the text 'no one who had ever seen' and (b) its spectrogram. Overlaid on the waveform, we show the transcription in English ARPABET format and the mappings to the four classes of interest, i,e. nasal consonant (NC), oral consonant (OC), nasalized vowel (NC), and oral vowel (OV). ARPABET to IPA mapping: N→/n/, OW→/o/, W→/w/, AH→/ɐ/, HH→/h/, UW→/u/, AE→/æ/, D→/d/, EH→/e/, V→/v/, ER→/ɚ/, S→/s/,IY→/i/.
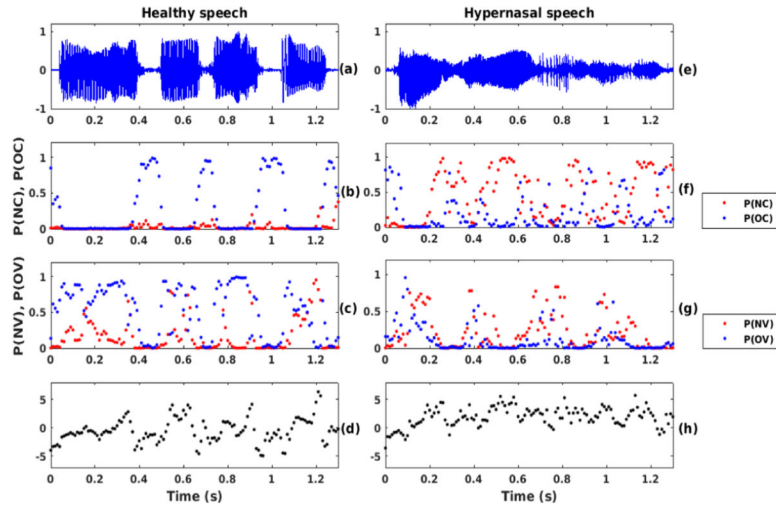
Fig. 5.
The four frame-wise DNN posteriors and corresponding frame-wise OHM for healthy and hypernasal speech: (a) the waveform of a control speech sample ("buy baby a bib"), (b) $P(NC)$ and $P(OC)$, (c) $P(NV)$ and $P(OV)$, and (d) the OHM for the target sentence produced by a child from the control group; (e) the waveform of a CP speech sample ("buy baby a bib"), (f) $P(NC)$ and $P(OC)$, (g) $P(NV)$ and $P(OV)$, and (h) OHM for target oral sentence produced by a participant with CP.
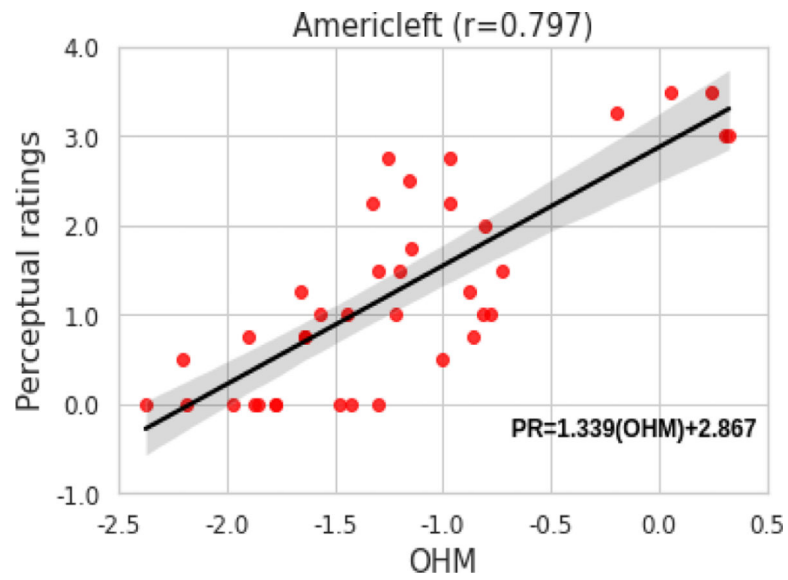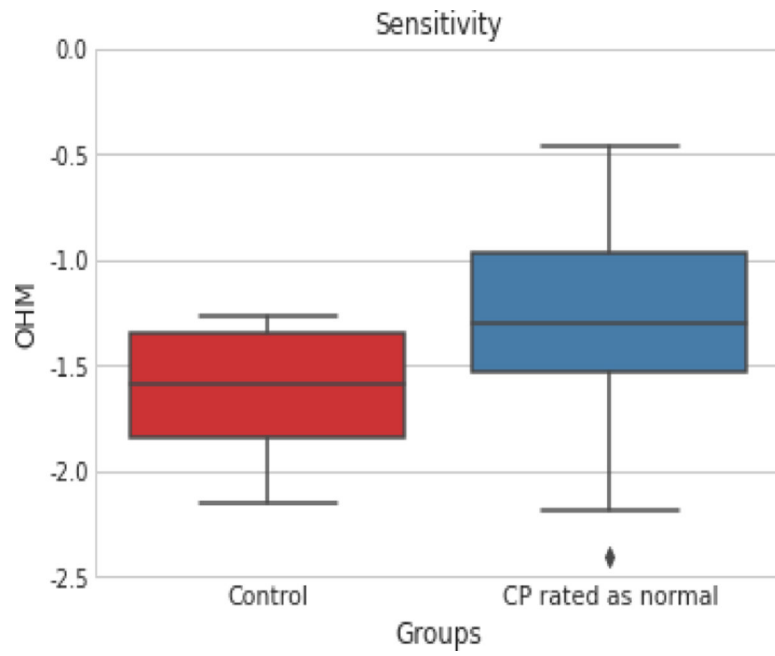
Fig. 6.
A scatter plot of speaker-level OHM scores vs. ground-truth perceptual ratings for the Americleft database. The linear regression equation for mapping between the OHM and perceptual ratings (PR) is $PR = 1.339(OHM) + 2.867$.
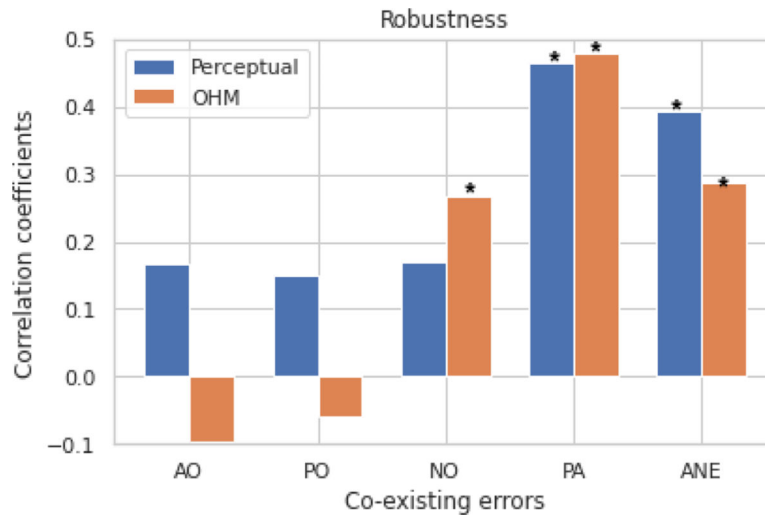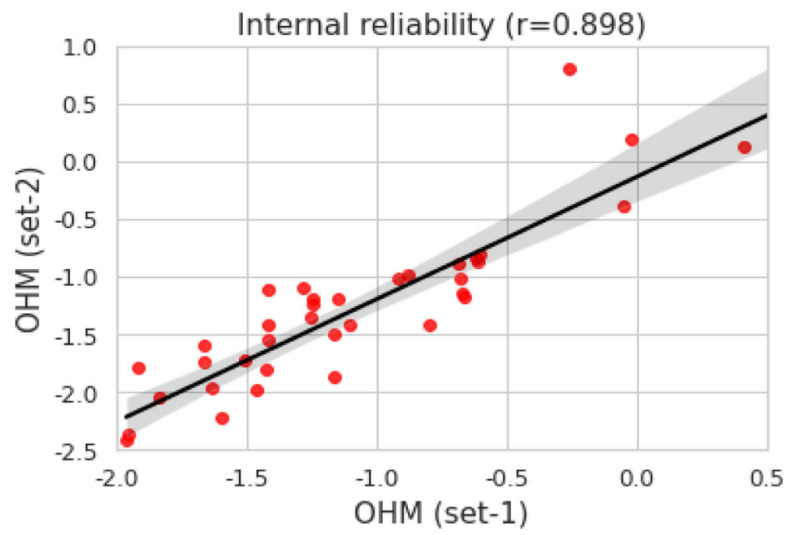
Fig. 7.
Correlation of OHM scores and perceptual hypernasality ratings with co-existing speech production errors. AO-Anterior oral, PO-Posterior oral, NO-Non oral, PA-Passive articulation, and ANE-Audible nasal emission ($*p < 0.05$).
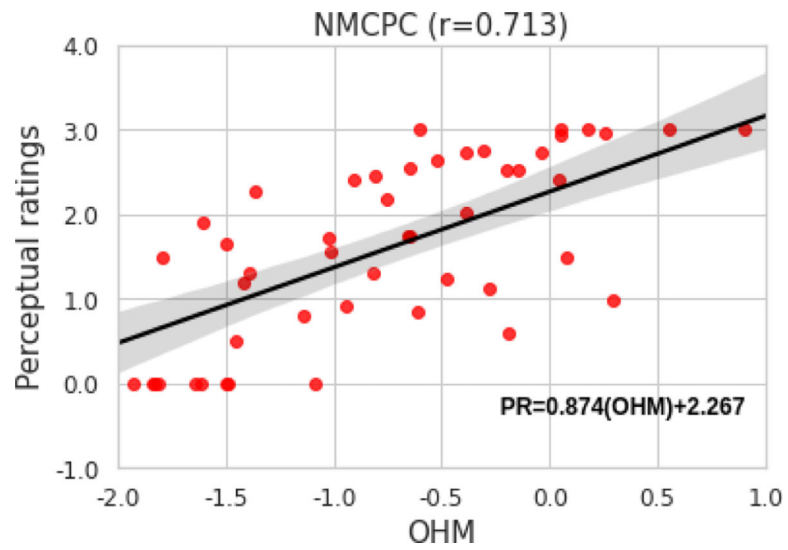
Fig. 8.
OHM scores for the control group and for children with CP rated as having normal nasality in the Americleft database.

Fig. 9.

Analysis of internal reliability. A scatter plot of speaker-level OHM scores computed for two independent sets of sentences.

Fig. 10.
A scatter plot of speaker-level OHM scores vs. ground-truth perceptual ratings for the NMCPC database. The linear regression equation for mapping between OHM and perceptual ratings (PR) is $PR = 0.874(OHM) + 2.267$.

TABLE I

LIST OF SENTENCES IN THE AMERICLEFT DATABASE, TARGET CONSONANTS, AND THE SENTENCE-LEVEL CORRELATION VALUES TO THE OHM.

Sl. No	Target	Sentence	r	p-value	Category	r (Category)	p-value (Category)
1	/p/	Puppy will pull a rope	0.414	< 0.05	Plosives	0.703	< 0.001
2	/b/	Buy baby a bib	0.568	< 0.001			
3	/t/	Your turtle ate a hat	0.586	< 0.001			
4	/d/	Do it today for dad	0.658	< 0.001			
5	/k/	A cookie or a cake	0.568	< 0.001			
6	/g/	Give aggie a hug	0.651	< 0.001			
7	/f/	A fly fell off a leaf	0.715	< 0.001	Fricatives	0.793	< 0.001
8	/v/	I love every view	0.716	< 0.001			
9	/θ/	Thirty-two teeth	0.616	< 0.001			
10	/ð/	The other feather	0.630	< 0.001			
11	/s/	Sissy saw sally race	0.691	< 0.001			
12	/z/	Zoey has roses	0.471	< 0.05			
13	/ʃ/	She washed a dish	0.724	< 0.001			
14	/s/-cluster	I spy a starry sky	0.540	< 0.001			
20	/h/	Hurry ahead harry	0.663	< 0.001			
15	/tʃ/	Watch a choo-choo	0.715	< 0.001	Affricates	0.752	< 0.001
16	/dʒ/	George saw gigi	0.698	< 0.001			
17	/l/	Laura will yell	0.487	< 0.05	Liquids	0.564	< 0.001
18	/ɹ/	Ray will arrive early	0.507	< 0.05			
19	/w/	We were away	0.640	< 0.001	Glides	0.640	< 0.001
21	/m/	Mom and amy are home	0.069	0.682	Nasals	0.108	0.518
22	/n/	Anna knew no one	−0.126	0.452			
23	/ŋ/	We are hanging on	0.269	0.103			
24	/m/, /n/, /ŋ/	We ran a long mile	0.162	0.331			

TABLE II

A COMPARISON OF THE AVERAGE PAIRWISE CORRELATION BETWEEN AMERICLEFT-TRAINED RATERS AND THE AVERAGE CORRELATION BETWEEN THE OHM AND EACH RATER.

	Mean±std.
Inter-rater correlation	0.776±0.068
OHM-rater correlation	0.733±0.051

TABLE III

A COMPARISON OF THE AVERAGE PAIRWISE CORRELATION BETWEEN NMCPC RATERS AND THE AVERAGE CORRELATION BETWEEN THE OHM AND EACH RATER.

	Mean±std.
Inter-rater correlation	0.872±0.078
OHM-rater correlation	0.695±0.016

TABLE IV

COMPARISON BETWEEN OHM AND DNN REGRESSOR.

Database	Approach	
	OHM	**DNN Regressor**
Americleft	$r = 0.797$, $p < 0.001$	$r = 0.524$, $p < 0.05$
NMCPC	$r = 0.713$, $p < 0.001$	$r = 0.301$, $p = 0.032$