An attention model for hypernasality prediction in children with cleft palate

Vikram C. Mathad[1], Nancy Scherer[1], Kathy Chapman[3], Julie Liss[1], and Visar Berisha[1,2]

[1]Department of Speech & Hearing Sciences

[2]School of Electrical, Computer, & Energy Engineering, Arizona State University, Tempe, AZ, USA

[3]Department of Communication Sciences and Disorders, University of Utah, Salt Lake City, UT, USA

Paper ID: 3812

Introduction

- Hypernasality refers to excessive air leakage through the nasal cavity during the production of voiced sounds
- Individuals with cleft palate (CP) exhibit hypernasality due to inadequate velopharyngeal closure
- Perception of hypernasality in speech is a primary outcome measure
- Evaluation of hypernasality is used to help determine the need for secondary cleft surgeries
- The perceptual evaluation: Gold standard
  - Requires trained speech language pathologists
  - Inter and intra-rater variability

Speech processing-based approaches

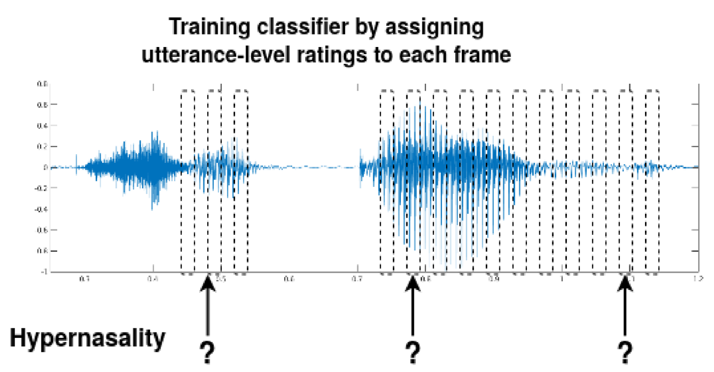

Figure: Existing approaches assign utterance-level labels to each frame. Is this reasonable?

- Conventional speech processing-based approaches for hypernasality prediction
  - Compute frame-level features like MFCCs
  - Training classifiers on utterance-level/speaker-level labels e.g support vector machine (SVM), deep neural network (DNN), etc.
- Hypernasality cues may not be present in each and every frame
- Training classifiers by assigning utterance-level ratings to each frame may not be appropriate

Attention BLSTM

- Attention-based recurrent neural network regressor maps frame-level features into utterance-level scores

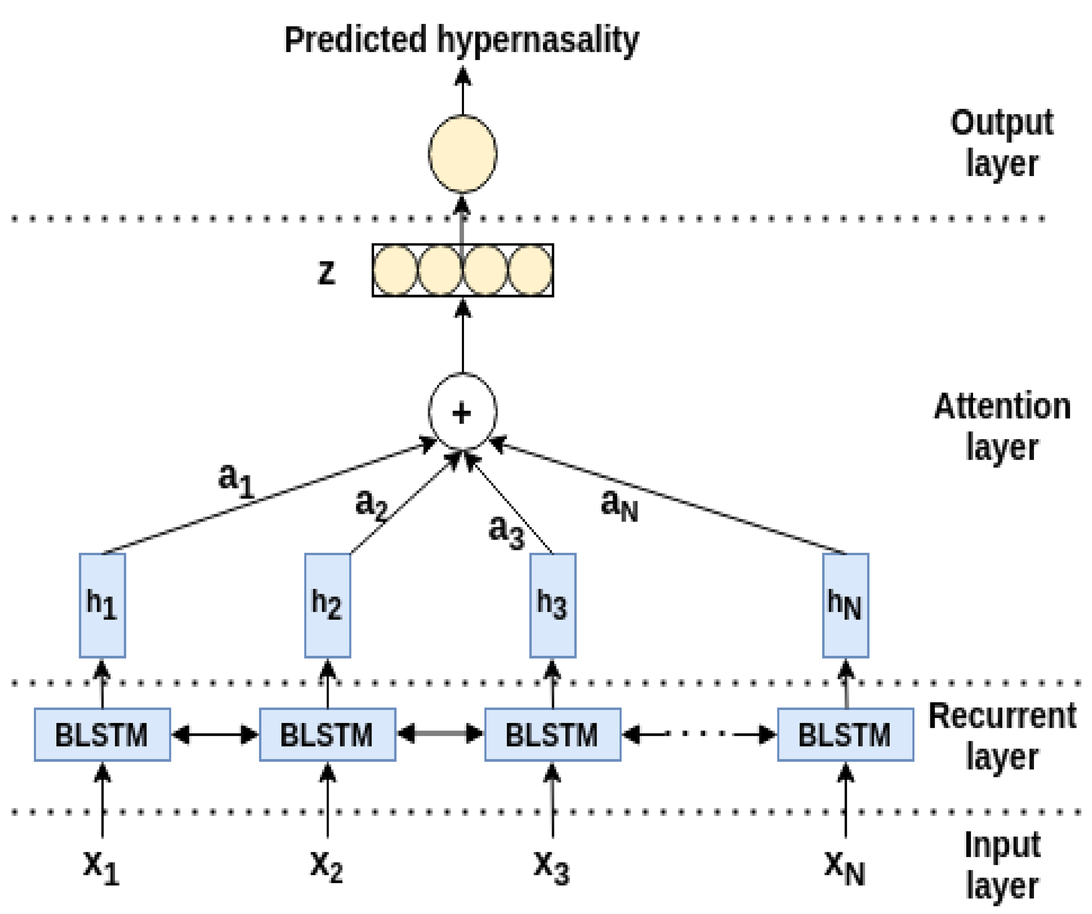

Figure: The architecture of attention BLSTM for hypernasality prediction

Attention BLSTM

- 39-dimensional MFCCs (20 ms/10 ms) computed for each sentence
- 128 memory cells and 50% dropout
- The attention weight vector a is defined as

$$a_i = \frac{e^{w^T h_i}}{\sum_{j=1}^{N} e^{w^T h_j}}. \tag{1}$$

  where h_i: outputs of the BLSTM $i \in [1, N]$, and w: N-dimensional learnable weights
- Weighted average (z) of BLSTM outputs: $z = \sum_{i=1}^{N} a_i h_i$
- z is finally passed through an output layer containing a single node with a linear activation function to predict sentence-level hypernasality
- Mean squared error (MSE) loss function, Adam optimizer, and 0.001 learning rate for 20 epochs

Database

- 38 children (28 CP and 10 controls) balanced across different severity levels
- 5-point Americleft Speech Protocol scale (0-normal, 1-borderline, 2-mild, 3-moderate, 4-severe) and 5 raters and inter-rater correlation 0.797
- For each sentence-level speech sample, 26 copies (6 noisy + 8 utterance-level speed perturbation + 8 word-level speed perturbation+ 4 vocal tract length perturbations (VTLP))

Analysis of attention weights

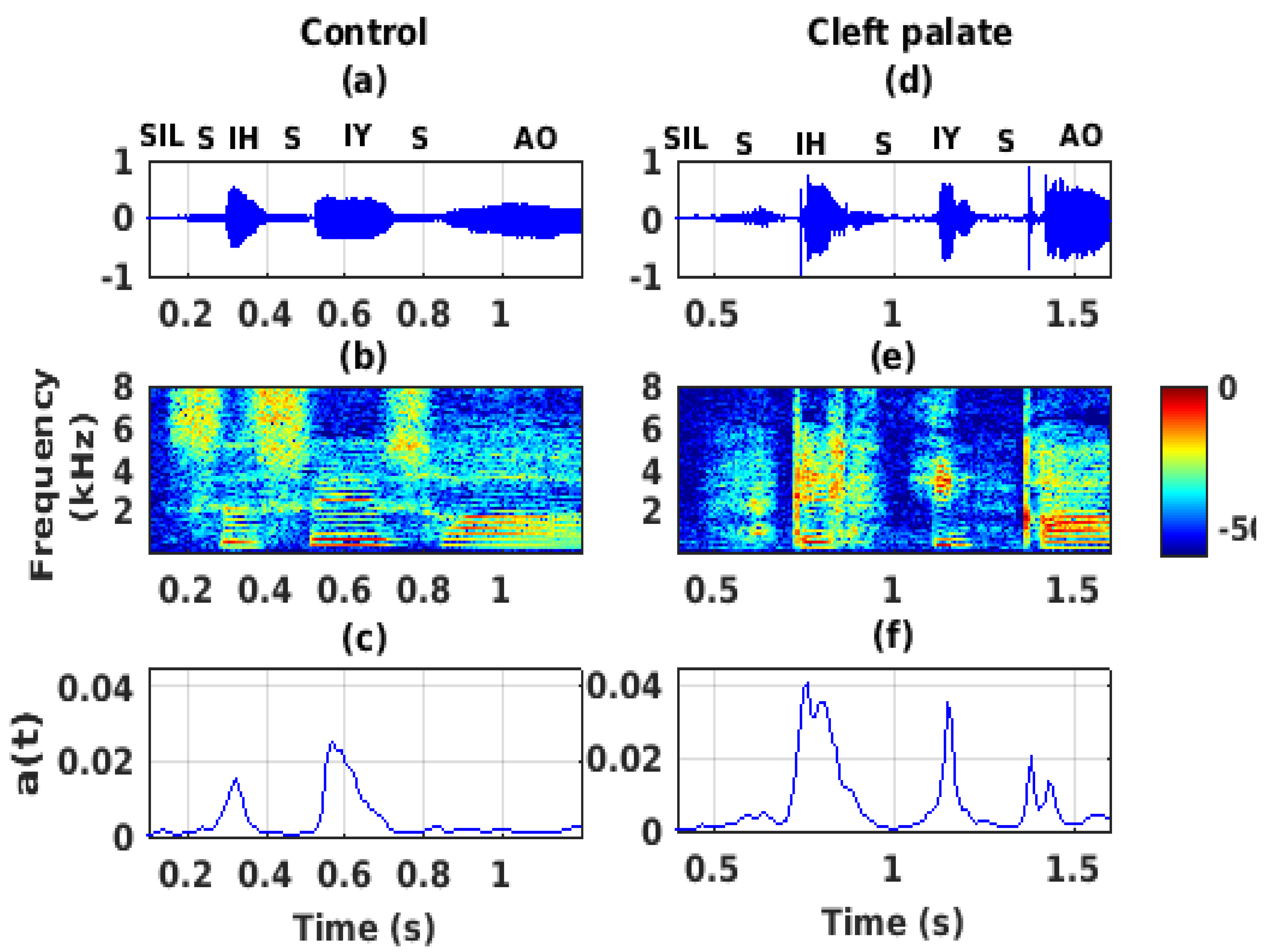

Figure: Plot of attention weights for the target speech "sissy saw". (a)-(c) and (d)-(f) speech waveform, spectrogram, and attention weight contours for control and CP speech, respectively.

Distribution of attention weights across broad phoneme categories

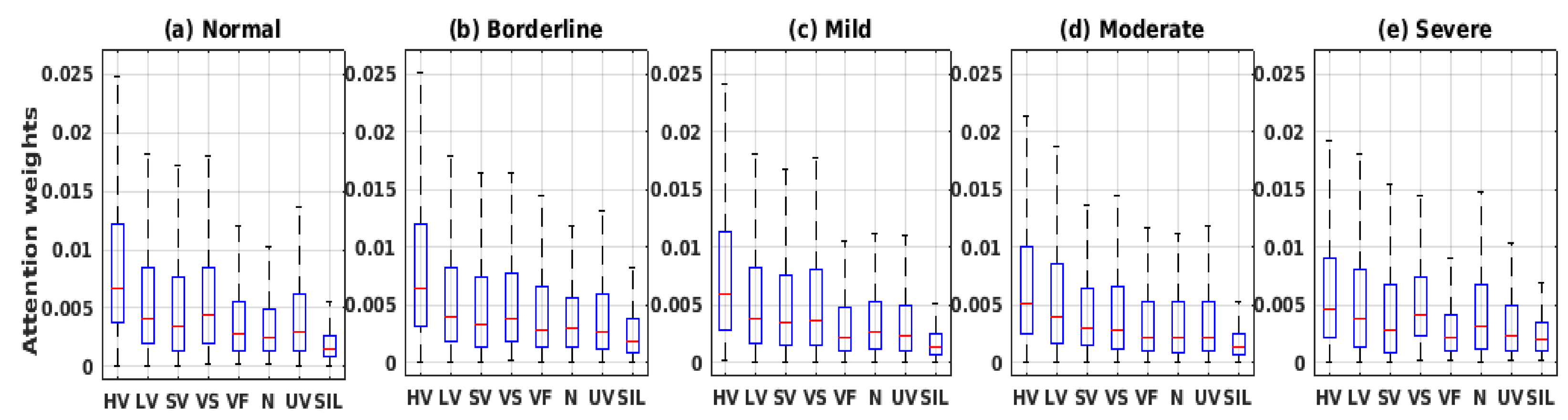

Figure: Boxplots representing the distribution of attention weights for high vowel (HV), low vowel (LV), semivowel (SV), voiced stops (VS), voiced fricatives (VF), nasals (N), unvoiced consonant (UV), and silence (SIL) sound categories across (a) normal, (b) borderline, (c) mild, (d) moderate, and (e) severe hypernasality levels.

Comparative evaluation

Table: Comparison among different neural network architecture ($*p < 0.001$). (Pearson's correlation coefficient (PCC) and mean absolute error (MAE))

Model	PCC	MAE
BLSTM-attention	**0.684***	**0.618**
FFDNN	0.559*	0.760
BLSTM-frame-wise	0.600*	0.724
BLSTM-final frame	0.636*	0.713
BLSTM-mean pooling	0.659*	0.650

Summary and Conclusion

- Attention weights primary focus on voiced regions
- Phoneme specific analysis:
  - High vowels in mild hypernasality; low and high vowels in moderate hypernasality; vowels and voiced stops in severe hypernasality
- Attention-BLSTM outperforms conventional frame-wise training
- Sentence-level averaging to get sentence-level scores
- Frame and sentence-level attention as a future scope

Acknowledgement

- This work was funded in part by NIH-NIDCR grant DE026252 and NIH-NIDCD grant R01DC006859.

References

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

[2] K. Bettens, F. L. Wuyts, and K. M. Van Lierde, "Instrumental assessment of velopharyngeal function and resonance: A review," Journal of communication disorders, vol. 52, pp. 170–183, 2014.

[3 C. M. Vikram, A. Tripathi, S. Kalita, and S. R. M. Prasanna, "Estimation of hypernasality scores from cleft lip and palate speech." in Interspeech, 2018, Hyderabad, India, pp. 1701–1705..

e-mail: vchikkaw@asu.edu