OBJECTIVE MEASURES OF PLOSIVE NASALIZATION IN HYPERNASAL SPEECH

Michael Saxon[1] *Julie Liss*[2] *Visar Berisha*[1,2] *

Arizona State University
[1]School of Electrical, Computer, and Energy Engineering
[2]Department of Speech and Hearing Sciences

ABSTRACT

Hypernasal speech is a common symptom across several neurological disorders; however it has a variable acoustic signature, making it difficult to quantify acoustically or perceptually. In this paper, we propose the nasal cognate distinctiveness features as an objective proxy for hypernasal speech. Our method is motivated by the observation that incomplete velopharyngeal closure changes the acoustics of the resultant speech such that alveolar stops /t/ and /d/ map to the alveolar nasal /n/ and bilabial stops /b/ and /p/ map to bilabial nasal /m/. We propose a new family of features based on likelihood ratios between the plosives and their respective nasal cognates. These features are based on an acoustic model that is trained only on healthy speech, and evaluated on a set of 75 speakers diagnosed with different dysarthria subtypes and exhibiting varying levels of hypernasality. Our results show that the family of features compares favorably with the clinical perception of speech-language pathologists subjectively evaluating hypernasality.

Index Terms— speech, hypernasality, dysarthria, velopharyngeal dysfunction, automatic speech recognition

1. INTRODUCTION

Hypernasality refers to the perception of excessive nasal resonance during speech production. It results from an inability to properly modulate airflow between the nasal and oral cavities due to velopharyngeal dysfunction (VPD) and arises from a cleft lip and palate, or dysarthria secondary to neurological disorders such as Parkinson's disease [1], amyotrophic lateral sclerosis [2], Huntington's disease [3], and ataxia [4].

Detecting and assessing hypernasality are complex tasks that require inferring the ratio of resonances across the pharyngeal, oral, and nasal cavities. A disproportionately high amount of nasal resonance is regarded as atypical and hypernasal. This presents a challenging estimation task, vulnerable to co-modulating variables including word choice, the particular geometry of an individual's resonating cavities, and other covarying dysarthria symptoms (e.g. vocal quality). This results in a highly nonlinear and complex mapping between the percept and the actual acoustic nasal resonance [5], [6].

Current techniques for measuring velopharyngeal function in-clinic employ perception, imaging, and instrumentation. The current state of the art is clinical perception of hypernasality by trained speech-language pathologists [7], however there is a growing body of work suggesting clinical perception is susceptible to the co-modulating variables mentioned above and listener expertise [8]. Reliable perceptual measures of hypernasality require evaluation from multiple clinicians [9] or intensive training according to specific protocols [10]. Direct imaging of the velopharyngeal closing mechanism can provide information about velopharyngeal gap size and shape using X-Ray or multiview videofluoroscopy [11], however these are invasive techniques and not common practice in-clinic. Nasalence is a score on a scale from 0-100 measured by a nasometer worn over the face [12]. While nasalance shows moderate correlation with perceptual judgment of hypernasality [13], it requires a specialized device and a trained clinician to be read and understood.

Existing work in assessing hypernasality directly from speech signals is primarily focused on extracting formant statistics or other spectral features (e.g. the spectral flattening, amplitude reduction, and bandwith increases that accompany nasalization [14], $1/3^{\text{rd}}$ octave band analysis [15], and the voice low tone to high tone ratio [16]). While these methods have all demonstrated some effectiveness in measuring hypernasality, the complex spectral signature of nasalization is difficult to capture with a simple representation. There is also a body of work on measuring nasality using machine learning [17], [18], [19], but these methods are all trained on single-disorder data, so it is difficult to assess if they learn acoustics specific to hypernasality or other co-modulating variables.

In contrast to the existing work in hypernasality assessment, and motivated by existing work on objective phone-level intelligibility measures, we propose a more comprehensive acoustic representation that does not require disease-specific training data. We observe that the unwanted nasalization of phonemes characteristic of hypernasal speech can make certain plosives sound more like nasal sonorants that share the same place of articulation. To that end, we propose

*This work was funded in part by NIH RO1 grant R01DC006859 (MPIs: Liss, Berisha).

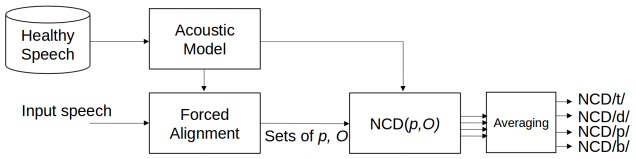

Fig. 1. High-level overview of the NCD feature system.

the nasal cognate distinctiveness (NCD) family of features. We train an acoustic model on healthy English speech. We calculate the feature values by evaluating the likelihood ratio between the plosives and their respective nasal cognates. These features compare favorably with clinical ratings of hypernasality provided by speech-language pathologists for data from 75 speakers diagnosed with different neurological diseases (Parkinson's disease, amyotrophic lateral sclerosis, Huntington's disease, and cerebellar ataxia).

2. NASAL DISTINCTIVENESS FEATURES

2.1. Motivation and Overview

A characteristic of hypernasal speech is the unintentional production of "nasal cognates," nasal sonorants sharing the same place of articulation as certain voiced plosives, when production of the corresponding plosive is intended. This transformation means that the voiced alveolar stop /d/ will sound like the alveolar nasal /n/ and the voiced bilabial stop /b/ will sound like the bilabial nasal /m/. [20] Similarly, the unvoiced counterparts of these stops /t/ and /p/ frequently are present in phonetic environments where they are proceeded or followed by vowels, or proceeded by nasal consonants [21], which means they also can share a propensity to be mapped to the same nasal cognates [20]. Predictable phenomena such as this suggest that perceptually-motivated, phoneme-level objective measures of hypernasality are warranted.

Existing phoneme-level objective measures have been developed targeting more general speech intelligibility assessment, such as the Goodness of Pronunciation algorithm (GoP) [22]. The GoP assesses the pronunciation of a speaker on a phoneme-by-phoneme basis as the log ratio of the probability of the uttered phoneme segment given the correct phoneme from an aligned transcript to the maximum across all phonemes of the uttered segment given a phoneme,

$$GOP(p) = \left| \log \left(\frac{P(O|p)}{\max_{q \in Q} P(O|q)} \right) \right| / |O|$$

where O is the observation with frame count $|O|$, p is the transcript-determined phone, and Q is the full set of phonemes in the language. These probabilities are assessed using an automatic speech recognition (ASR) acoustic model trained on healthy native speech. Similar approaches have been used to measure intelligibility in dysarthric speech [23].

2.2. Feature Computation

In Fig. 1 we provide an overview of the proposed approach. We assume that we have an input speech segment and corresponding transcript for analysis. Furthermore, we assume that the input utterances have several instances of the phonemes of interest (/p/ /b/, /t/, /d/). Similar to the Goodness of Pronunciation feature, the Nasal Cognate Distinctiveness feature computation begins with an ASR acoustic model trained on healthy speech. This acoustic model is used to both force-align the speech to the transcript to sample the plosives and estimate the likelihood ratios between the plosives and their nasal cognates with which the NCD features are computed. Finally, the individual instances of each phoneme are averaged to generate average NCD features.[1]

2.2.1. Acoustic Model

To train our acoustic model, we extract a set of observation feature vectors from each training speech sample. The input speech sampling rate is 16 kHz. We analyze the speech at a frame rate of 10 ms and denote the acoustic features for frame i by O_i. For our implementation we used a triphone model trained with a Gaussian Mixture Model-Hidden Markov Model on 960 hours of healthy native English speech data from the LibriSpeech corpus [24]. We use the Kaldi toolkit training scripts for training the model. The input features to the ASR model are 39-dimensional second order Mel-Frequency Cepstral Coefficient (MFCC) with utterance-level cepstral mean variance normalization and Linear Discriminant Analysis transformation (same approach as in [25]).

2.2.2. NCD Feature Computation

The NCD is formulated for phoneme $p \in S = [/\text{t}/,/\text{d}/,/\text{p}/,/\text{b}/]$, frame $O_i \in O$, the observation corresponding with p based on forced alignment to the transcript,

$$NCD(p, O) = \Sigma_i \log \left(\frac{P(O_i|p)}{P(O_i|\text{cog}(p))} \right) / |O|$$

where $\text{cog}(p)$ is a "cognate mapping function" that maps the stops in the set S to their corresponding nasal cognate, and $|O|$ is the total number of frames in observation O.

The probabilities in the numerator and denominator of the formula are assessed using the Viterbi alignments in the ASR model. To assess the denominator probability the $\text{cog}(p)$ function is called first, swapping the given plosive with its cognate in the triphone context.

Given a set of recordings of a speaker reading from a set of transcripts, the four NCD features are evaluated as follows. First, the transcripts are force-aligned at the phoneme level using the ASR model. With this alignment the $NCD(p)$ feature can be computed for each all phonemes $p \in S$ in the input

[1] Code is available at `https://github.com/michaelsaxon/ncd`

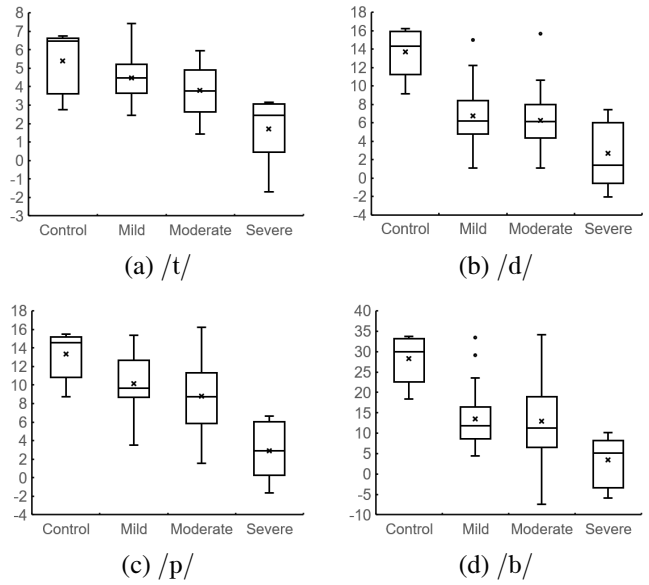

Fig. 2. Box plots for the NCD feature distribution separated by nasality severity. The y-axis in each plot represents the NCD feature value for the phoneme under consideration.

utterances. This produces a set of many output NCD values, with each corresponding to an occurrence of one of the four phonemes in consideration in the transcripts. The NCD values are then averaged within the four phonemes to return four output features, NCD for /t/, /d/, /p/, and /b/.

3. RESULTS

The method was evaluated using a dataset of 75 speakers with a variety of dysarthria subtypes exhibiting a range of hypernasality severities. The dysarthria is secondary to Parkinson's disease (N=38), Huntington's disease (N=6), amyotrophic lateral sclerosis (N=15), and cerebellar ataxia (N=16). Each speaker spoke five phonetically rich test sentences. The sentences were: "the supermarket chain shut down because of poor management", "much more money must be donated to make this department succeed", "in this famous coffee shop they serve the best doughnuts in town", "the chairman decided to pave over the shopping center garden", and "the standards committee met this afternoon in an open meeting."

We evaluate the NCD features against clinical perception of hypernasality, as measured by speech language pathologists (SLPs). There is considerable evidence that the inter-rater reliability between clinicians evaluating hypernasality is susceptible to other co-modulating variables [9]. As a result, we recruit a group of 15 SLPs to asses each speaker's degree of nasality on a 1-7 scale and average their scores - we use the average of the 15 clinical ratings as our ground truth. The inter-rater reliability of the SLPs was moderate, with a Pearson Correlation Coefficient of 0.66 and an average inter-clinician mean absolute error of 1.44 on the 7-point scale. A group of 5 healthy control speakers, speaking the same sentences was also included; all were assigned the minimum nasality score of 1.

3.1. Individual Feature Analysis

Figure 2 contains box plots for the four phoneme NCD features. The speakers were divided into four groups based on nasality severity for this analysis: control, mild, moderate, and severe. To perform the separation the real range of non-control assessed nasality was divided roughly in three, with the mild nasality $N \in [1.3, 2.7)$, moderate $N \in [2.7, 4.1)$ and severe $N \in [4.1, 5.6]$.

The feature trends very convincingly move for the voiceless phonemes /t/ and /p/, with the control and mild nasality speakers exhibiting the highest values of Nasal Cognate Distinctness. The moderate nasality speakers then exhibit lower feature values and the severe nasality speakers exhibit the lowest. The means, medians, and quartiles for all of the values decrease as nasality increases across groups. These expected trends are not all exhibited in the voiced phonemes /d/ and /b/, however. For both phonemes the means, medians, and quartiles hardly move at all or do not move together between the mild and moderate nasality groups. For /b/, the moderate nasality NCD feature range even spans the entire range of values exhibited by all other groups. Despite these inconsistencies, for all phonemes the NCD score completely separates the control range from the severe nasality range.

3.2. Predicting the Nasality Score

Multiple regression analysis was used to test if the NCD measure for the four nasal cognates predicted the average clinician nasality ratings. The results of the regression analysis indicated that the four predictors explained 47% of the variance ($R = 0.687, F(4, 79) = 16.798, p < 0.05$). It was found that the NCD for /t/ significantly predicted the hypernasality rating ($\beta = -0.316, p < 0.05$), as did the NCD for /p/ ($\beta = -0.278, p < 0.05$).

Factor	B	SE B	β	p
/t/	-0.225	.093	-0.316*	0.018
/d/	-0.061	.043	-0.201	0.163
/p/	-0.077	.030	-0.278*	0.013
/b/	0.000	.016	0.002	0.987
R^2		0.473		
F		16.798**		

Table 1. Summary of Linear Regression Analyses for Variables Predicting Clinical Hypernasality Scores from the NCD Measures (N=80). *$p < 0.05$, **$p < 0.001$

Figure 3 shows model-predicted nasality in Table 1 against the SLP-assessed clinical hypernasality measure.

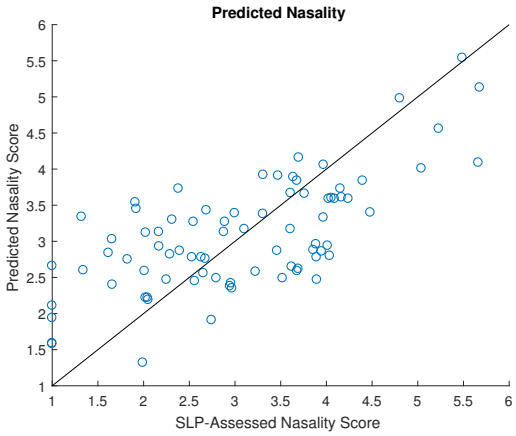

Fig. 3. Output of the linear regression model predicting nasality using /t/, /d/, /p/, and /b/ as shown in Table 1.

3.3. Discussion

The NCD features are formulated as a log probability ratio between the expected class of a given transcript plosive and its nasal cognate. Increasingly positive values correspond with a higher confidence that the speaker has correctly articulated the plosive rather than its nasal cognate, and values closer to zero or negative represent plosives that sound more like their nasal cognate than the intended stop. This directionality is exhibited as expected in the phoneme-by-phoneme analysis of the feature, where across speaker classes the NCD of a given phoneme decreases as nasality increases.

Figure 2 shows that the NCD features very clearly separate the control and severe nasality groups with all phonemes. However, for the voiced phonemes /d/ and /b/ the moderate and mild means and medians are close and the quartile ranges overlap considerably. Performance is worst for /b/, which exhibits both high cross-group overlap of the quartile ranges and insignificance as a predictor of the subjective hypernasality scores in the multiple regression analysis.

When considering these inter-phone performance inconsistencies, differences in the phonetic environments in which the test phonemes appear are noteworthy. In the five test sentences, /t/ appears 18 times, /d/ 9 times, /p/ 6 times and /b/ only 3 times. Of these appearances, /t/ has 8 word-internal appearances, 7 word-final appearances, and 3 word-initial appearances. The phoneme /d/ has 5 word-initial appearances, 3 word-internal appearances, and 1 word-final appearance. The phoneme /p/ has 4 word-internal appearances and 2 word-initial appearances, and /b/ exclusively has 3 word-initial appearances. It is likely that these disparities in overall occurrence and word-internal occurrence play an important role in explaining the performance disparity.

Additionally, it is important to note that the NCD features are intended to assess a physical phenomenon, the realized allophones, not the underlying phonemes themselves. The phonetic transcriptions provided for HMM-based ASR systems fall somewhere between broad phonetic transcriptions and allophonic narrow transcriptions, allowing for possible confusion scenarios. For example, a /t/ may be realized in English as [t], [ɾ], or [ʔ] depending on phonetic environment. All three could be compared to [n] in the NCD model even though the glottal stop [ʔ] is unaffected by VPD and shares no place of articulation with [n].

The NCD features tend to be high-variance because they require reliable phoneme-level alignment to compute; higher frequency phonemes exhibit reduced variability through averaging. Accordingly, in this study the more frequent phonemes are more useful predictors of hypernasality. This suggests that future datasets to evaluate methods like NCD should include a higher frequency of plosive consonants balanced across the categories, in consistent environments in which the correct allophones are reliably produced.

4. CONCLUSION

In this paper, we proposed the nasal cognate distinctiveness features as an objective and noninvasive proxy for hypernasal speech. The features are motivated by the simple observation that alveolar stops /t/ and /d/ map to the alveolar nasal /n/ and the bilabial stops /p/ and /b/ map to bilabial nasal /m/ when the energized column of air is shunted into the nasal passage during speech production. The feature is measured by first training an acoustic model on healthy speech and, for a test speaker, evaluating the likelihood ratio between the plosives and their respective nasal cognates. For healthy speakers that exhibit no signs of hypernasality, this ratio is large and decreases with increasing levels of hypernasality. This is confirmed on speech samples from 75 speakers diagnosed with different dysarthria subtypes and exhibiting varying levels of hypernasality. The results show that the features are strongly correlated with clinical perception.

As we saw in Fig. 2, some of the features are variable. Future work will focus on characterizing and mitigating this variability. This mitigation will require collecting samples of disordered speech that contain a more balanced representation of the phonemes under evaluation, enabling more adequate representation of /b/, /d/, and /p/, and allowing similar analysis of /k/, /g/, and their cognate, /ŋ/. This study has shown that standard phonetically diverse sentences are suboptimal for nasality assessment tasks. Future participants should be asked to produce sets of utterances that contain multiple attempts at the precise allophones that have nasal cognates.

In addition, alignment of the speech at the phoneme level can be challenging for speech from speakers with impaired articulation. To that end, in the future we will investigate integrating these ASR-based, alignment-reliant methods alongside new, strictly acoustic methods for hypernasality estimation that do not rely on alignment. Future work will more directly target predication of the SLP-assessed nasality ratings.

5. REFERENCES

[1] D.G. Theodoros, B.E. Murdoch, and E.C. Thompson, "Hypernasality in Parkinsons disease: A perceptual and physiological analysis," *J Med Speech-Lang Pathol*, vol. 3, no. 2, pp. 73–84, 1995.

[2] J. Duffy, *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management*, Mosby, 1995.

[3] M. Novotny, J. Rusz, R. Cmejla, H. Ruzickova, J. Klempir, and E. Ruzicka, "Hypernasality associated with basal ganglia dysfunction: evidence from Parkinson's disease and Huntington's disease," *PeerJ*, vol. 4, pp. e2530, 2016.

[4] M.L. Poole, J.S. Wee, J.E. Folker, L.A. Corben, M.B. Delatycki, and A.P. Vogel, "Nasality in Friedreich ataxia," *Clin Linguist Phon*, vol. 29, no. 1, pp. 46–58, Jan 2015.

[5] K. Bettens, L. Bruneel, Y. Maryn, M. De Bodt, A. Luyten, and K. M. Van Lierde, "Perceptual evaluation of hypernasality, audible nasal airflow and speech understandability using ordinal and visual analogue scaling and their relation with nasalance scores," *J Commun Disord*, vol. 76, pp. 11–20, Jul 2018.

[6] M. de Stadler and C. Hersh, "Nasometry, videofluoroscopy, and the speech pathologist's evaluation and treatment," *Adv. Otorhinolaryngol.*, vol. 76, pp. 7–17, 2015.

[7] A.W. Kummer and L. Lee, "Evaluation and Treatment of Resonance Disorders," *Language, Speech, and Hearing in Schools*, vol. 27, pp. 271–281, Jul 1996.

[8] K. Strobel-Schwarthoff E. Nkenke S. Paal, U. Reulbach and M. Schuster, "Evaluation of speech disorders in children with cleft lip and palate," *J Orofac Orthop*, vol. 66, no. 4, pp. 270–278, Jul 2005.

[9] R. H. Scarmagnani, A.C. Oliveira, A.P. Fukushiro, M.H. Salgado, I.E. Trindade, and R.P. Yamashita, "Impact of inter-judge agreement on perceptual judgment of nasality," *Codas*, vol. 26, no. 5, pp. 357–359, 2014.

[10] K. Brunnegard, A. Lohmander, and J. van Doorn, "Comparison between perceptual assessments of nasality and nasalance scores," *Int J Lang Commun Disord*, vol. 47, no. 5, pp. 556–566, 2012.

[11] A.S. Woo, "Velopharyngeal dysfunction," *Semin Plast Surg*, vol. 26, no. 4, pp. 170–177, Nov 2012.

[12] Pentax, "Nasometer ii: Model 6450," 2016.

[13] 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.

[14] T. Pruthi and C. Espy-Wilson, "Acoustic parameters for the automatic detection of vowel nasalization," in *Proc. Interspeech 2007*, 2007, pp. 1925–1928.

[15] R. Kataoka, D.W. Warren, D.J. Zajac, R. Mayo, and R.W. Lutz, "The relationship between spectral characteristics and perceived hypernasality in children," *The Journal of the Acoustical Society of America*, vol. 109, no. 5, pp. 2181–2189, 2001.

[16] Y.J. Tsai, C.P. Wang, and G.S. Lee, "Voice low tone to high tone ratio, nasalance, and nasality ratings in connected speech of native mandarin speakers: a pilot study.," *The Cleft Palate-Craniofacial Journal*, vol. 49, no. 4, pp. 437–46, 2012.

[17] S. Murillo, D. H. Peluffo, and G. Castellanos, "Support vector machine-based approach for multi-labelers problems," in *European Symposium on Artificial Neural Networks, Computational Inteligence and Machine Learning*, 2013.

[18] M. Golabbakhsh, F. Abnavi, M. Kadkhodaei Elyaderani, F. Derakhshandeh, F. Khanlar, P. Rong, and D.P. Kuehn, "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.

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

[20] L.D. Shriberg and R.D. Kent, *Clinical Phonetics*, Wiley Communications Series. Macmillan, 1982.

[21] H. J. Giegerich, *English Phonology: An Introduction*, Cambridge Textbooks in Linguistics. Cambridge University Press, 1992.

[22] S.M. Witt and S.J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Commun.*, vol. 30, no. 2-3, pp. 95–108, Feb. 2000.

[23] M.J. Kim, Y. Kim, and H. Kim, "Automatic intelligibility assessment of dysarthric speech using phonologically-structured sparse linear model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 694–704, April 2015.

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

[25] M. Tu, A. Grabek, J. Liss, and V. Berisha, "Investigating the role of L1 in automatic pronunciation evaluation of L2 speech," in *Proc. Interspeech 2018*, 2018, pp. 1636–1640.