Recent Advances in Hypernasal Speech Detection using the Nonlinear Teager Energy Operator

Douglas A. Cairns, John H.L. Hansen, and James F. Kaiser*

Robust Speech Processing Laboratory
Department of Electrical Engineering
Duke University, Box 90291, Durham, NC 27708-0291
http://www.ee.duke.edu/Research/Speech

ABSTRACT

Speakers with a defective velopharyngeal mechanism produce speech with inappropriate nasal resonance. It is of clinical interest to detect hypernasality as it is indicative of an anatomical, neurological, or peripheral nervous system problem. While clinical techniques exist for detecting hypernasality, a preferred approach would be noninvasive to maximize patient comfort and naturalness of speaking. In this study, a noninvasive technique based on the Teager Energy operator is proposed. Employing a proposed model for normal and nasalized speech, a significant difference between the Teager Energy profile for lowpass and bandpass filtered nasalized speech is shown, which is nonexistent for normal speech. An optimum classification algorithm is formulated that detects the presence of hypernasality using a measure of the difference in the Teager Energy profiles. The classification algorithm was evaluated using native English speakers producing front and mid vowels. Results show that the presence of hypernasality in speech can be reliably detected (94.7%) using the proposed classification algorithm.

1. INTRODUCTION

The speech communication process requires a translation of thoughts into spoken language. For a person with an anatomic and/or neurological impairment, the required vocal tract configuration and/or excitation may be compromised. The resulting speech will therefore be of reduced quality. A specific example of a vocal tract dysfunction that causes reduced speech quality, is that of a defective velopharyngeal mechanism. Speakers with this defect produce hypernasal speech (i.e., speech with inappropriate nasal resonance) across voiced elements (vowels). Since a defective velopharyngeal mechanism and the corresponding hypernasal speech can be caused by anatomical defects (cleft palate or other trauma), central nervous system damage (cerebal palsy or traumatic brain injury), or peripheral nervous system damage (Moebius Syndrome), it is important to be able to detect hypernasal speech in a clinical setting.

Since hypernasal speech arises from inappropriate nasal-oral coupling, researchers have attempted to use a nasal/oral ratio to detect hypernasality. Horii proposed a measure of nasal coupling called the *H*orii *O*ral *N*asal *C*oupling index ($HONC$) [1]. A modification of the measurement technique[2], gave rise to another index called the *N*asal *A*ccelerometric *V*ibrational *I*ndex ($NAVI$). Another approach utilizing a nasal-oral ratio is employed by an instrument dubbed $NORAM$ (*N*asal *O*ral *RA*tio *M*eter), where one contact microphone is placed on the alar wing of the nose while another is placed over the lamina of the thyroid cartilage [3]. A third device using a nasal-oral ratio is the Nasometer, based on the work of Fletcher et al. [4]. The measurement apparatus consists of a baffle plate with microphones attached to the top and bottom of the plate. Finally, the measurement of pressure and flow values during the nonvocalic elements (i.e. unvoiced consonants) has also been used to detect hypernasality[5].

While the approaches discussed here are capable of detecting hypernasality, each induces a somewhat artificial speaking situation and is physically intrusive to some extent. In this study, an algorithm is proposed based on the nonlinear Teager Energy operator and signal detection/estimation theory to achieve this goal.

2. ALGORITHM FORMULATION

Research studies have attempted to determine acoustic cues for nasalization [6] which generally include (i) first formant bandwidth increases and intensity decreases, (ii) nasal formants appear, and (iii) antiresonances appear.

A model for normal speech is composed of formants at various frequencies. This can be written as,

$$S_{NORMAL}(\omega) = \sum_{i=1}^{I} F_i(\omega) \quad (1)$$

where $F_i(\omega)$ is the i^{th} frequency domain formant. In contrast, nasalized speech is characterized by formants, antiformants, and nasal formants,

$$S_{NASAL}(\omega) = \sum_{i=1}^{I} F_i(\omega) - \sum_{k=1}^{K} AF_k(\omega) + \sum_{m=1}^{M} NF_m(\omega) \quad (2)$$

Research suggests that intensity reduction of the first formant is a primary cue for nasality [6]. Therefore, (1) and (2) can be rewritten to filter out the higher formants. The lowpass filtered (LPF) equations reduce to,

$$S_{NORMAL-LPF}(\omega) = F_1(\omega) \quad (3)$$

*This work was sponsored by a Young Biomedical Investigative Research Grant from The Whitaker Foundation.

$$S_{NASAL-LPF}(\omega) = F_1(\omega) - \sum_{k=1}^{\acute{K}} AF_k(\omega) + \sum_{m=1}^{\acute{M}} NF_m(\omega) \quad (4)$$

Here, the filtered normal speech has a single component, while the nasalized speech is multicomponent.

The multicomponent nature of the lowpass filtered nasalized speech can be exploited through the use of the Teager Energy operator (TEO). The TEO was first used by Teager in his work on speech production [7], and documented by Kaiser as [8]

$$\Psi_d[x(n)] = x^2(n) - x(n+1)x(n-1). \quad (5)$$

The TEO can be shown to be sensitive to multicomponent signals. For multicomponent signals, the output of the TEO is the TEO of each component plus cross terms. While past speech studies using the TEO have employed bandpass filtering to avoid this cross-term property, we chose to exploit it for nasalized speech. To illustrate this property, the inverse Fourier transform of (3) and (4) is taken and the TEO applied. The result for lowpass normal and nasalized speech will be,

$$\Psi_d[s_{NORMAL-LPF}(n)] = \Psi_d[f_1(n)] \quad (6)$$

$$\Psi_d[s_{NASAL-LPF}(n)] = \Psi_d[f_1(n)] - \sum_{k=1}^{\acute{K}} \Psi_d[af_k(n)] + \sum_{m=1}^{\acute{M}} \Psi_d[nf_m(n)] + \sum_{j=1}^{\acute{K}+\acute{M}+1} \Psi_{cross}[f_1(n), af_1(n), nf_1(n), ..)] \quad (7)$$

At this point, consider a comparison between the output of the TEO of speech that has been bandpass filtered (BPF) around the first formant,

$$\Psi_d[s_{NORMAL-BPF}(n)] = \Psi_d[f_1(n)] \quad (8)$$

$$\Psi_d[s_{NASAL-BPF}(n)] = \Psi_d[f_1(n)] \quad (9)$$

and the TEO of lowpass filtered speech as in (6) and (7). As can be seen, the output of the TEO for lowpass filtered and bandpass filtered nasalized speech is appreciably different, while the output of the TEO is the same for normal speech in each case. This comparison forms the basis of the hypernasal detection system.

The proposed TEO based detection system was inspired by prior studies by Cairns and Hansen in classifying speech as stressed/normal [9] or normal/hypernasal [10]. The results of these studies showed that the shape of the Teager Energy profile is a useful criteria for analysis and classification of non-normal speech. For normal speech, the comparison should show little or no difference, while for hypernasal speech, the comparison should show a measurable difference. To illustrate this, Fig.'s 1(a) and (b) show Teager Energy profiles for LPF and BPF versions of one frame of normal and hypernasal speech. The difference resulting from the additional terms in (7) is clear. To quantify the difference, the correlation coefficient $r = \frac{C}{\sigma_{LPF}\sigma_{BPF}}$ between the two TEO profiles is determined, and a likelihood ratio detector formulated to make a normal/hypernasal decision.

3. HYPERNASAL DETECTION SYSTEM

In order to perform analysis, a pitch detector is used to mark pitch epochs across a consonant-vowel-consonant (CVC) utterance (i.e., pitch synchronous analysis). A speech window is passed to the formant tracker which locates the first formant and computes the TEO profile of the bandpass filtered first formant. The same speech window is also lowpass filtered to exclude the second and higher formants, and the TEO profile computed. The cross-correlation coefficient between the two TEO profiles is then determined, and a likelihood ratio detector computes a normal/hypernasal decision for the current speech window. This procedure terminates when the final pitch epoch is encountered ($n = N$). An overall normal/hypernasal decision is then made based on the percentage of speech windows classified as normal.

3.1 Pitch Detector: The pitch detector employed estimates the glottal closure instant (GCI). The GCI is determined using a modified dyadic wavelet transform (D_yWT) pitch detector. Kadambe and Boudreaux-Bartels first formulated a D_yWT based pitch detector [11]. See [10] for further pitch detector details.

3.2 Formant Tracker: Once pitch synchronous information is available, the first formant track is estimated for the vowel in the CVC. It has been suggested that formants can be modeled as amplitude modulated, frequency modulated (AM-FM) signals. Given an AM-FM model for speech, Maragos, Kaiser, and Quatieri [12] developed the energy separation algorithm (ESA) to isolate the AM and FM components. The separation equations are represented as [12]

$$f(n) \approx \frac{1}{2\pi T} \arccos(1 - \frac{\Psi[y(n)] + \Psi[y(n+1)]}{4\Psi[x(n)]}) \quad (10)$$

$$|a(n)| \approx \sqrt{\frac{\Psi[x(n)]}{1 - (\frac{\Psi[y(n)]+\Psi[y(n+1)]}{4\Psi[x(n)]})^2}}, \quad (11)$$

where $y(n) = x(n) - x(n-1)$, $\Psi[\bullet]$ is the discrete TEO, $f(n)$ is the FM contribution at sample n, and $a(n)$ is the AM contribution at sample n. It was later shown that the FM signal could be used to iteratively refine the formant center frequency[14],

$$f_c^{i+1} = \frac{1}{N} \sum_{n=1}^{N} f(n). \quad (12)$$

where f_c^{i+1} is the formant center frequency on iteration $i+1$. Given an initial formant estimate, the speech section is BPF and (12) applied to determine the center frequency on the next iteration. The filter and refine procedure, called the iterative energy separation algorithm (IESA), continues until the formant center frequency converges. The advantage here is that not only are formant locations found, the individual formants TEO profiles are produced as a byproduct of the computations.

3.3 LPF/Teager Energy Operator: Estimation of the TEO profile from lowpass filtered speech is the second

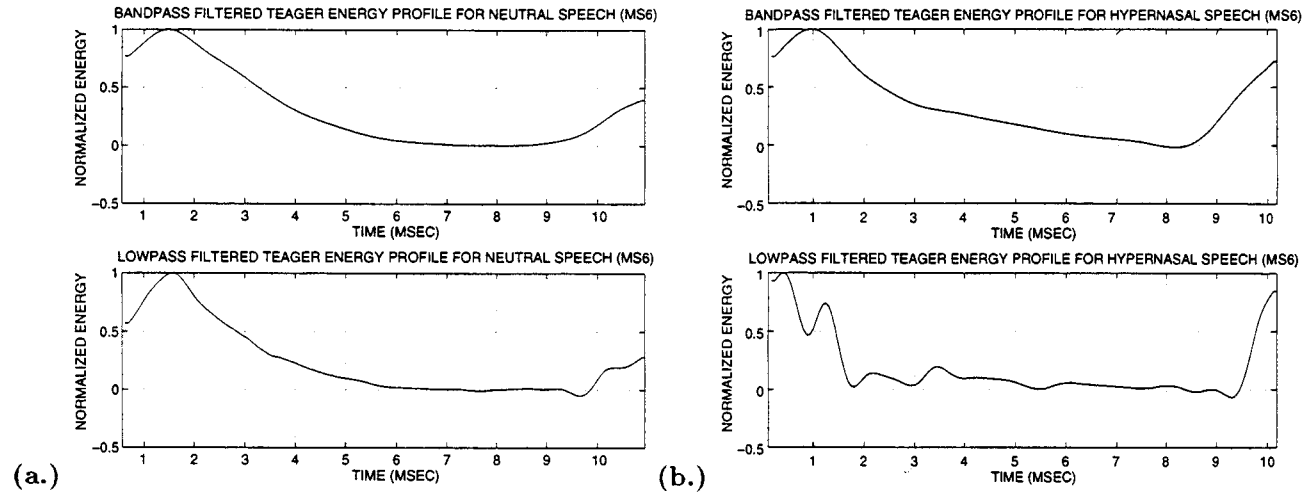

Figure 1: Teager Energy profiles for lowpass and bandpass filtered (a) normal and (b) hypernasal speech.

part of the comparison required for a normal/hypernasal decision. To preserve the TEO shape, a linear phase finite impulse response (FIR) filter is designed with a cutoff frequency that is dependent on the vowel of the CVC (1.3kHz for front vowels, 1.0kHz for mid vowels). The TEO profiles for LPF and BPF speech are then passed to the likelihood ratio detector.

3.4 Likelihood Ratio Detector: To cast the hypernasal detection problem, consider the antiformants and nasal formants as the signal to be detected, and the true first formant as the noise. The binary hypothesis problem becomes,

$$H_1: \quad \mathbf{x} = [\mathbf{af} + \mathbf{nf}] + \mathbf{f_1} \tag{13}$$

$$H_0: \quad \mathbf{x} = \mathbf{f_1}. \tag{14}$$

To formulate the likelihood ratio, the densities for H_1 and H_0 must be known or estimated from data. Since there are unknown parameters under each hypothesis in this case, this is referred to as the doubly composite signal hypothesis problem. This problem is solved by integrating the density functions over the unknown parameters (assuming the range of parameters is known). This gives

$$\lambda(\mathbf{r}) = \frac{\int_\alpha \int_\beta p(\mathbf{r}|H_1, \alpha, \beta)}{\int_\gamma \int_\delta p(\mathbf{r}|H_0, \gamma, \delta)} \tag{15}$$

Note that numerator and denominator are integrated separately to obtain the marginal densities $p(\mathbf{r}|H_1)$ and $p(\mathbf{r}|H_0)$. The likelihood ratio can then be computed from the marginal densities.

4. EVALUATION METHOD

The data analyzed in this study was collected from a group of eleven (6 male, 5 female) native speakers of English. Each speaker was judged to have normal nasal resonance by a speech pathologist. A speaker was asked to repeat the carrier phrase "*Please say* **blank** *again*," in order to capture a natural speaking style. The **blank** was filled by a series of CVC utterances containing the vowels /i/, /a/, or /A/[1]. The normal resonance data is composed of words with a plosive as the initial consonant (/p/, /k/, /t/, /g/, /d/, /b/) and /t/ as the final consonant. The hypernasal data is composed of words with /m/ or /n/ as the initial and final consonants. Example carrier phrases with CVCs are normal: "*Please say pAt again*" and nasalized: "*Please say mAm again*".

To ensure that a speaker was producing normal or hypernasal data, a Nasometer (Kay Elemetrics model 6200) was used to monitor each subject while speaking. Speech was recorded directly to a digital audio tape (DAT) machine (final sample rate was 16kHz).

5. RESULTS

For evaluation, receiver operating characteristic (ROC) curves were constructed. The second performance measure for the detection system was the percentage of words correctly identified as hypernasal and normal. The probability density functions used in this work are based on experimental evidence, and used to form the integration range for the doubly composite signal hypothesis detector (see Fig. 2).

Fig. 3 illustrates a representative ROC curve for the vowel /i/ for male speakers who were able to produce distinct normal and hypernasal speech (9 of the 10 speakers). It is clear that the detection algorithm achieves outstanding performance.

The second performance measure, the percentage of speech type correctly identified, can show if the approach is a reasonable solution across speakers. As evidenced by overall correct identification rates of 94.7% (normal) and 94.7% (hypernasal) for /i/ (see Fig. 4), and 93.0% (normal) and 93.3% (hypernasal) for /A/, the detection system consistently identifies normal and hypernasal speech.

To facilitate a comparison, a criteria must be established to label a speaker as either normal or hypernasal. Let the criteria be that a speaker is judged normal/hypernasal if greater than 50% of the experimental

[1]Single-symbol ARPAbet symbols are used to indicate phonemes. [13]

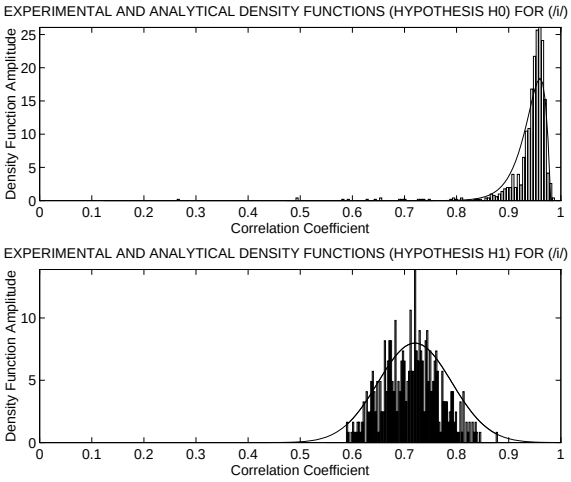

Figure 2: Analytical density function fitted to experimental data for vowel /i/ (male speaker set).

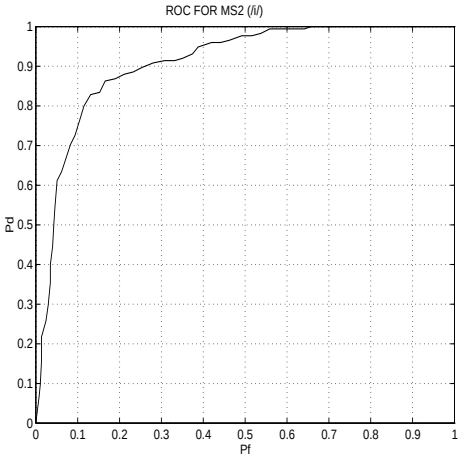

Figure 3: Example ROC for /i/ for Male speakers.

data is identified as that type of speech. Given this criteria, the results show that 16 of the 17 evaluations (94.1%) correctly label a speaker as normal, and 16 of 16 evaluations correctly label a speaker as hypernasal (speaker FS1 was unable to nasalize). Since the standard used in this study for determining whether a speaker is normal or hypernasal is the Nasometer, these results indirectly show that the proposed algorithm correlates with listener judgements of hypernasality as well as the Nasometer.

6. CONCLUSIONS

In this study, a hypernasal detection algorithm based on the nonlinear Teager Energy operator is proposed. Using a property of the Teager Energy operator, a useful measure of hypernasality is derived. This measure was tested on native English speakers for front (/i/) and mid (/A/) vowels, achieving average identification rates for normal/hypernasal speech of 94.7/94.7% and 93.0/93.3% respectively. These results indicate that this algorithm is a promising approach for noninvasively detecting hypernasal resonance. Also, the lack of specialized hardware

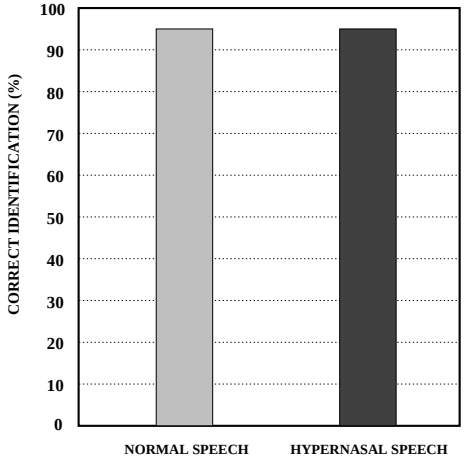

Figure 4: Correct identification rates for /i/.

required allows this algorithm to be easily utilized in clinical and nonclinical settings.

References

[1] Y. Horii, J. Lang, "Distributional Analysis of an Index of Nasal Coupling (HONC) in Simulated Hypernasal Speech", *Cleft Palate J.* , 18(4):279-285, 1981.

[2] M.A. Redenbaugh, A.R. Reich, "Correspondence Between NAVI and Listeners' Direct Magnitude Estimations of Hypernasality", *J. Speech Hear. Res.,* (28):273-281, 1985.

[3] J. Karling, O. Larsen, R. Leanderson, K. Galyas, A. Serpa-Leitao, "NORAM-An Instrument Used in the Assessment of Hypernasality", *Cleft Palate J.* , 30(2):135-140, 1993.

[4] S. G. Fletcher, L. E. Adams, M. J. McCutcheon, "Cleft Palate Speech Assessment Through Oral-Nasal Acoustic Measures", *Communicative Disorders Related to Cleft Lip and Palate*, (Boston: Little and Brown), pp. 246-257, 1989.

[5] D. W. Warren, A. B. Dubois, "A Pressure-Flow Technique for Measuring Velopharyngeal Orifice Area During Continuous Speech", *Cleft Palate J.* , (1):52-71, 1964.

[6] G. Fant, "Nasal Sounds and Nasalization", *Acoustic Theory of Speech Production*, (The Hague: Mouton), 1960.

[7] H.M. Teager and S. M. Teager, "Evidence for Nonlinear Production Mechanisms in the Vocal Tract", in *Speech Production and Speech Modeling*, pp. 241-261, 1990.

[8] J.F. Kaiser, "On a Simple Algorithm to Calculate the 'Energy' of a Signal", *Proc. ICASSP-90*, pp. 381-384, April 1990.

[9] D.A. Cairns, J.H.L. Hansen, "Nonlinear analysis and classification of speech under stressed conditions", *J. Acoust. Soc. Am.* , 92(5):3392-3400, Dec. 1994.

[10] D.A. Cairns, J.H.L. Hansen, J. Riski, "A Noninvasive Technique for Detecting Hypernasal Speech Using a Nonlinear Operator", *IEEE Trans. Biomed. Eng.*, 43(1):35-45, Jan. 1996

[11] S. Kadambe, G.F. Boudreaux-Bartels, "Application of the Wavelet Transform for Pitch Detection of Speech Signals", *IEEE Trans. on Infor. Theory*, (38):917-924, 1992.

[12] P. Maragos, J. Kaiser, T. Quatieri, "Energy separation in signal modulations with application to speech analysis", *IEEE Trans. Sig. Proc.,* 41(10):3024-3051, Oct. 1993.

[13] J.R. Deller, J.G. Proakis, J.H.L. Hansen, *Discrete Time Processing of Speech Signals,* Macmillan Series for Prentice-Hall Publishers, New York, New York, 1993.

[14] H. Hanson, P. Maragos, A. Potamianos, "A System for Finding Speech Formants and Modulations via Energy Separation," *IEEE Trans. Speech, Audio Proc.* (2):436-442, 1994.