A JOINT ACOUSTIC-ARTICULATORY STUDY OF NASAL SPECTRAL REDUCTION IN READ VERSUS SPONTANEOUS SPEAKING STYLES

Vikram Ramanarayanan[*]*, Dani Byrd*[^]*, Louis Goldstein*[^] *and Shrikanth Narayanan*[*] [^]

[*]Signal Analysis and Interpretation Laboratory, Ming Hsieh Department of Electrical Engineering,
[^]Department of Linguistics*;* University of Southern California, Los Angeles, CA-90089-0899

ABSTRACT

Speech styles are one of the primary phenomena of prosodic variation in speech. We present a novel automatic procedure to analyze real-time magnetic resonance images (rt-MRI) of the human vocal tract recorded for read and spontaneously spoken speech. This is applied to rt-MRI data on nasal articulation, jointly used with acoustic analyses of the speech signal, to analyze nasal production differences in read and spontaneous speech, especially focusing on reduction. In this exploratory study, vowel-nasal-vowel (VNV) sequences from one speaker were examined and measures extracted from both acoustic and articulatory signals. Significant differences were observed in the realizations of constriction-forming events for read and spontaneous speaking styles. Such an analysis has implications for understanding speech planning and for informing design of automatic speech analysis algorithms.

Index Terms— speech production, real-time MRI, nasals, vocal tract, image motion analysis, read speech, spontaneous speech, spectral reduction.

1. INTRODUCTION

A joint consideration of direct articulatory and acoustic data of read and spontaneous speech styles can provide an improved understanding of the underlying differences in their production. Such investigations can be of potential use for efforts in both speech science and speech technology. This paper considers such a comparative study, focusing on nasal consonants in read and spontaneous speaking styles, using novel real-time MRI methods to characterize and study vocal tract events. In addition to illuminating certain prosodic features found in read and spoken style variations, this work can help us understand the nature of coupling between the oral tract and nasal tract and can help inform speech representation for production-inspired automatic speech recognition (ASR).

In this paper, we focus on investigating a specific aspect of read and spontaneous speech production differences—acoustic reduction—which is characterized by a reduction in spectral and durational distinctions between sounds as speaking style becomes more informal or the stress on the syllable is reduced. Van Son and Pols [1] report that intervocalic energy differences for nasal consonants are higher in spontaneous speech than in read speech, reversing a trend that is generally seen for all other consonants; while Nakamura et al. [2] give explicit evidence for *acoustic* reduction of the MFCC spectrum for most consonants of English as the speech becomes more and more spontaneous, causing a reduction in ASR performance. Using direct articulatory data in conjunction with the acoustic speech signal, we hope to obtain insights into this behavior from an articulatory perspective.

Varying degrees of coarticulation, velum speed, and timing are some of the important factors that must be taken into consideration in examining the articulatory causes of spectral reduction in nasals [3, 4]. The specific research questions we pose with respect to reduction phenomena are: (1) how do we quantify vocal tract shaping differences (with respect to constriction-forming events) for the different speaking styles, (2) what spectral reduction patterns are observed in vowel-nasal-vowel sequences in read and spontaneous speech, and (3) how is the timing and speed of the velum in read speech different from that in spontaneous speech?

The recent advances in real-time magnetic resonance imaging (MRI) offer an excellent tool to investigate this problem, since real-time MRI allows one to capture the full vocal tract during speech production and quantify the 'choreography' of the articulators [5], making it an ideal technique to compare articulation during read and spontaneous speech production.

The paper is organized as follows: Section 2 details methods of MR data acquisition and reconstruction. Section 3 describes the analysis carried out on the acoustic signal, as well as a method to characterize the midsagittal profile of the vocal tract. Finally, Section 4 discusses our results and summarizes directions for future work.

2. DATA

One native speaker (female) of American English was engaged in a simple dialog on topics of general nature (e.g., "what music do you listen to…", "tell me more about your favorite cuisine …", etc.) while she was lying supine inside the MR scanner. For each speech "turn," audio responses and MRI videos of vocal tract articulation were recorded for 30 seconds and time-synchronized. The same speaker was also recorded/imaged while reading some of the TIMIT sentences and the rainbow passage during a separate scan session. Further details regarding the recording and imaging setup can be found in [5,6]. Midsagittal real-time MR images of the vocal tract were acquired with a repetition time of TR=6.5ms on a GE Signa 1.5T scanner with a 13 interleaf spiral gradient echo pulse sequence. The slice thickness was approximately 3mm. A sliding window reconstruction at a rate of 22.4 frames per second was employed.

3. ANALYSES

Given that most baseline speech analyses have been carried out on the easily available speech signal, it is important to note that the analyses carried out here in the articulatory domain are driven by acoustic-domain preprocessing. In this section, we first describe the acoustic domain pre-processing and analysis, including the segmentation of the speech signal into phones so that context-specific (VCV) analyses can be performed, and extraction of a spectral centroid measure. Then we explain how relevant features were extracted from the MRI videos using automatically-determined air-tissue boundary information.

The SONIC speech recognizer [7] was used to perform a first-pass automatic alignment of the recorded noise-cancelled audio data to the phone sequence. However, occasional misalignment of some phones or groups of phones warranted a second-pass manual correction of these alignments.

Spectral centre-of-gravity (CoG) values were computed for each vowel-nasal-vowel (VNV) sequence extracted automatically from the segmented speech data (band-limited to 5 kHz for both read and spontaneous speech). This helps us approximately obtain the frequency about which maximum spectral energy is distributed without unwanted averaging effects due to noisy sample values at higher frequencies up to 10kHz (since the audio is sampled at 20kHz). This was computed as follows [1]:

$$CoG = \frac{\sum_i f_i E_i}{\sum_i E_i}$$

where f_i is the center frequency of each FFT band of frequencies, and E_i is the spectral power corresponding to each frequency.

3.1. Contour Extraction

The air-tissue boundary of the articulatory structures was automatically extracted using an algorithm that hierarchically optimizes the observed image data fit to an anatomically informed object model using a gradient descent procedure [8]. The object model is chosen such that different regions of interest such as the palate, tongue, velum etc. are each defined by a dedicated region (see Figure 1).

3.2. Velum Speed Measure

The velum contour *alone* was used to create a 'mask' image corresponding to each image of the MRI video sequence, with all pixels inside the contour rendered white and the rest, black. Then by taking the absolute value of the difference between successive masks, we can obtain a measure of how fast the velum is moving. For a detailed description of this process, see [9]. In addition, the velum opening for each MR image was computed as the minimum distance between the velum and pharyngeal wall contours.

3.3. Vocal Tract Area Descriptors (VTADs) Extraction

In the following, the extraction of vocal tract variables, including lip aperture (LA), tongue tip constriction degree (TTCD), tongue dorsum constriction degree (TDCD), tongue root constriction degree (TRCD), and velic aperture (VEL) is described. For each image in the MRI video sequence, LA is computed as the minimum distance between the upper lip and lower lip contour segments. VEL is computed as the minimum distance between the velum and pharyngeal wall contours. In order to extract the tongue-related tract variables (TTCD, TDCD and TRCD), we consider 2 possible cases where the specific articulator (tongue tip, dorsum or root) is critical or non-critical. If the articulator in question is a critical articulator, then the corresponding constriction degree is simply calculated as the minimum distance from the tongue contour to the palate contour. For those frames where the articulator in question is not a critical one, the main problem is defining the point on the palate with respect to which we can measure the constriction degree for that articulator. This problem can be alleviated by using frames in which an articulator is critical in order to define a set of possible 'palate constriction locations;' this, in turn, can then be used to compute the constriction degrees for that articulator for all other frames.

For example, in order to compute TTCD for a vowel /a/, in which the tongue tip is not critical, we use the constriction location (on the palate) of the tongue tip constriction for all /t/, /d/ frames, where the tongue tip is a critical articulator and use the mean of this point cloud as the point on the palate from which to measure minimum distance to the tongue contour. We find that choosing /t/ and /d/ frames, /k/ and /g/ frames, and /a/ and /r/ frames as critical frames for the tongue tip, tongue dorsum, and tongue root respectively works well in practice. Finally, the lowermost boundary of the vocal tract area for our purposes is computed as the minimum distance between the root of the epiglottis and pharyngeal wall contour (see Figure 1). However, due to poor signal-to-noise (SNR) ratio of images in this region, this is not always robust.

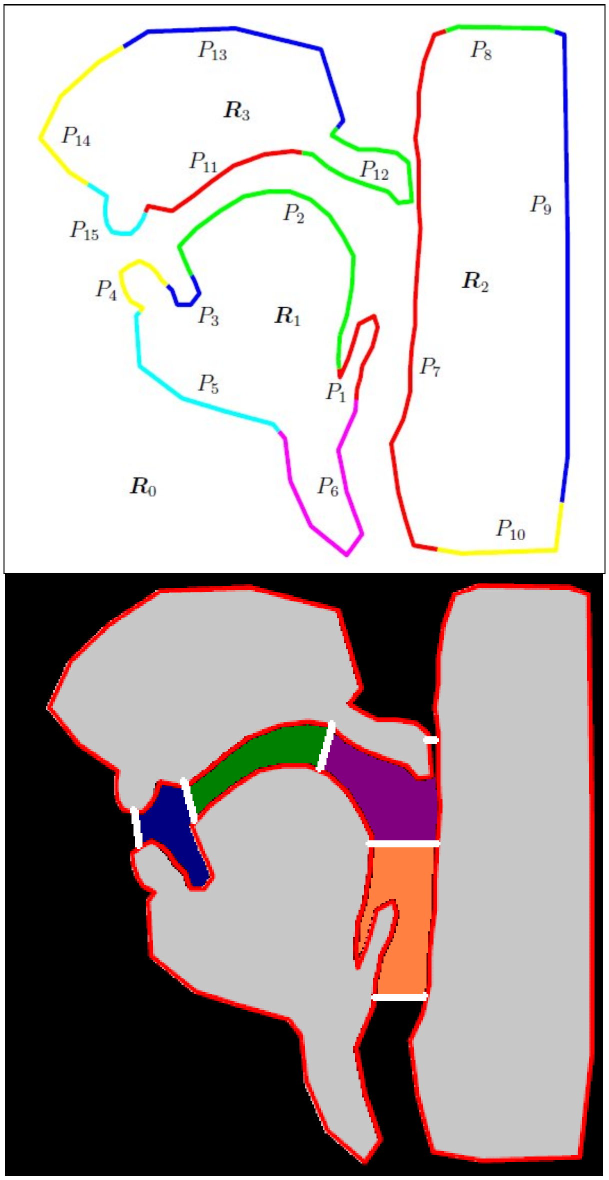

Figure 1 *(Top) Contour outlines extracted for each image of the vocal tract. Note the template definition such that each articulator is described by a separate contour. (Bottom) A schematic depicting the concept of vocal tract area descriptors (adapted from [8]).*

Once these tract variables are computed, we can then use them to partition the vocal tract midsagittal cross-sectional area, into the area between the LA and TTCD (which we call A1 [blue in Fig]), the area between the TTCD and TDCD (or A2 [green]), the area between the TDCD and TRCD (or A3 [purple]), and the area below the TRCD as A4 [orange]. (We are not using A4 for conclusive analyses, due to the reason described above). Once these areas are obtained, we can formalize the differences in vocal tract shaping more concretely.

4. RESULTS AND DISCUSSION

4.1. Results for all nasals in general

A one-way parametric[1] analysis of variance (ANOVA) with post-hoc Tukey test was used to test the hypothesis that the means of the z-scores of spectral CoG values calculated for each VNV instance in the read and spontaneous speech utterances were the same. All nasals (/m/,/n/,/ng/) in spontaneous speech were found to have a significantly lower ($p<=0.05$) spectral centroid as compared to those in read speech, which agrees with the van Son and Pols study [1]. Van Son and Pols assert that the steepness of spectral slope and therefore CoG frequency is determined by speech effort, generally viewed as greater in read speech.

One-way parametric ANOVAs with post-hoc Tukey tests were also used to test the hypothesis that the mean values of the areas (defined in Section 3.3) computed individually over the course of each VNV instance in the read and spontaneous speech utterances belonged to the same distribution. The average area of the A2 vocal tract region (defined between TTCD and TDCD) for each nasal instance was found to be significantly *higher* ($p=0$) for read nasals as opposed to spontaneously produced nasals. In addition, the A1 region (defined between the LA and TTCD) was found to be significantly *lower* ($p<=0.05$) for read nasals as opposed to spontaneously produced nasals. These observations could be due to a more anterior tongue-tip constriction location in the read speech cases; however, no conclusive evidence was found for the same. Vowel coarticulation effects are also a potential factor, e.g., if the /i/…/i/ vowel contexts force the consonant to be shaped more /i/-like. Repeating these experiments separated for different vowel contexts showed that the reported phenomenon was still significant irrespective of the underlying vowel context. Furthermore, the variances of the areas computed over the course of each VNV instance in spontaneous speech were found to be generally higher than in read speech, although statistical significance was obtained only for A1. This behavior points to a greater degree of variability in time and space for the realization of gestures in spontaneous produced VNVs, which in turn is an indicator of articulatory reduction.

In order to check if the above-mentioned articulatory behavior had any influence on the spectral centroid trends, the Pearson's correlation between each area and the spectral centroid was computed. Results indicate minimal correlation between the spectral centroid measure and each area ($\rho=0.34$ for A1, $\rho=-0.39$ for A2, $\rho=-0.01$ for A3), which suggests the spectral reduction phenomena might be characterized by at least two separate non-linearly related or unrelated phenomena. In order to investigate whether the degree of velum opening contributed in any significant way to our CoG observations, z-scores of

[1] P-P plots and Kolmogorov-Smirnov tests ($p=0.05$) were used to ascertain parametricity wherever required.

the maximum width of velum opening for each VNV instance were computed for both read and spontaneous speech. There were no significant differences (p>=0.05) observed between the means of these two samples, suggesting that the extent to which the velum lowers is similar in both read and spontaneous speech for our speaker. A one-way ANOVA was also used to compare the z-scores of the average as well as the maximum speed of velum movement extracted for each VNV sequence in the read and spontaneous speech samples. Again, no significant differences (p>=0.05) were found between read and spontaneous samples in either case, which suggests that the extent and speed of velum opening have a minimal effect on spectral reduction, at least for our speaker. However, the variance of the speed measure was much higher in the case of spontaneous nasals, as would be expected. The length of nasal duration for each VNV sequence was also computed for read and spontaneous speech utterances and compared. The z-scores of the two samples showed no significant differences when tested using a one-way ANOVA (p>=0.05). Absolute values of rate of change of areas (A1, A2 and A3) were not found to be significantly different for read and spontaneous nasals.

4.2. Alveolar nasals

It is interesting to note that the normalized short-term acoustic energy of the nasal /n/ in read speech was much higher than in spontaneous speech (p<=0.05). However, no significant differences were found between the normalized short-term energies of the other nasals (/m/,/ng/) for the two speaking styles.

In addition, the palate constriction location coordinates were found for each alveolar nasal when the constriction degree was minimum. The variance of these values was found to be much higher for spontaneous alveolar nasals as compared to the read nasals. This observation agrees with studies suggesting that spectral reduction is characterized by a blurring in constriction location position (cf:[10]).

4.3. Conclusions and future work

In this paper, we presented a new approach towards examining MRI data of speech articulation, and applied it to analyzing spectral reduction phenomena in nasal consonants. Future work includes validating and extending these results to more speakers. Since all the articulatory measures automatically extracted from the MRI data are dependent on the specific attributes of the data such as the SNR, field-of-view scale (which can be thought of as a zoom factor), scan protocol, etc., it is important to come up with robust methods to account for these factors such that comparisons across speakers can be made.

5. ACKNOWLEDGEMENTS

Work described in this paper was supported by NIH Grant DC007124, the USC Imaging Sciences Center, and the USC Center for High Performance Computing and Communications (HPCC).

6. REFERENCES

[1] Van Son, R. J. J. H. and Pols, L. C. W. (1999): An acoustic description of consonant reduction. *Speech Communication.* 28: 125-140.

[2] Nakamura, M., Iwano, K., and Furui, S. (2008). Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. *Computer Speech and Language*. 22: 171– 184.

[3] Byrd, D., Tobin, S., Bresch, E. and Narayanan, S. (2009). Timing effects of syllable structure and stress on nasals: A real-time MRI examination. *Journal of Phonetics*. 37(1): 97-110.

[4] Moll, K. & Daniloff, R. (1971). Investigation of the timing of velar movement during speech. *Journal of the Acoustical Society of America.* 50: 678-684.

[5] Narayanan, S., Nayak, K., Lee, S., Sethy, A., and Byrd, D. (2004). An approach to real-time magnetic resonance imaging for speech production. *Journal of the Acoustical Society of America*. 115(4): 1771-1776.

[6] Bresch, E., Nielsen, J., Nayak, K., and Narayanan, S.. (2006). Synchronized and noise-robust audio recordings during realtime MRI scans. *Journal of the Acoustical Society of Americ*a.120(4): 1791-1794.

[7] Pellom, B. (2001), "SONIC: The University of Colorado Continuous Speech Recognizer", University of Colorado, #TRCSLR- 2001-01, Boulder, Colorado.

[8] Bresch, E., and Narayanan, S. (2009). Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images. *IEEE Transactions on Medical Imaging*. 28(3): 323-338.

[9] Ramanarayanan, V., Bresch, E., Byrd, D., Goldstein, L., Narayanan, S. (2009). Analysis of pausing behavior in spontaneous speech using real-time magnetic resonance imaging of articulation. *Journal of the Acoustical Society of America*. 126 (5): EL160-EL165.

[10] Van Son, R. J. J. H. and Pols, L. C. W. (1995): Acoustic consonant reduction: a comparison. *Proceedings of the Institute of Phonetic Sciences, University of Amsterdam.* 19: 83-91.