Planning and Execution in Soprano Singing and Speaking Behavior: an Acoustic/Articulatory Study Using Real-Time MRI

Vikram Ramanarayanan, Adam Lammert,
Dani Byrd, Louis Goldstein and Shrikanth Narayanan
University of Southern California, Los Angeles, CA – 90007

ABSTRACT

This paper examines three different behavioral modalities of production in the human vocal tract: read speech, spontaneous speech and singing. We observe and quantify differences in planning and execution of these three behaviors. We use audio-synchronized real-time magnetic resonance imaging technology [1] to record speech and song data from 4 formally trained, professional sopranos. To analyze the data, we propose fully automatic measures of average articulator speed, posture and acoustic spectra to analyze the data. Finally, we provide evidence that these different behavioral modalities involve speech planning mechanisms in different ways.

There have been many attempts at qualitative and quantitative descriptions of speech and song. [2] analyzed pitch patterns in spoken sentences, birdsong and instrumental music themes. Final lengthening and post-skip reversals predominated in all domains, based on which the authors suggest possible shared motor constraints for all three coordinated actions (read speech, spontaneous speech and singing); in addition, arch-like pitch contours were found in music and speech but not birdsong, possibly reflecting an influence of speech patterns on musical structure. [3] also found that music of English and French reflects patterns of durational contrast between successive vowels in spoken sentences, as well as patterns of pitch interval variability in the speech of the respective languages.

In an earlier study [4], we devised a method to measure how the speed of articulators changes over time before, during and after a pause and how the smoothness or abruptness of this progression can give an indication of planning differences. More specifically, in the case of grammatically well-formed pauses, we see a gradual decrease in articulator speed as we move into the pause and a gradual increase as we move out of it. On the other hand, in the case of hesitation or word-search pauses (so-called "ungrammatical" pauses), we see an abrupt spurt in speed as we move out of the pause, which might be due to the re-establishment of global pacing of the speech that occurs after the boundary-adjacent slowing. More recently, we proposed a method to evaluate postural differences ("articulatory setting") between pauses in-between speech as opposed to an absolute rest position in both read and spontaneous speaking styles [5] and showed significant postural differences and differences in degree of active control between these different cases. In this study, we extend these methods to analyze the differences between speech and song and test the hypothesis that each of these coordinated actions have different planning and execution constraints, resulting in different utilizations of the forward map of production (i.e., articulation-to-acoustics).

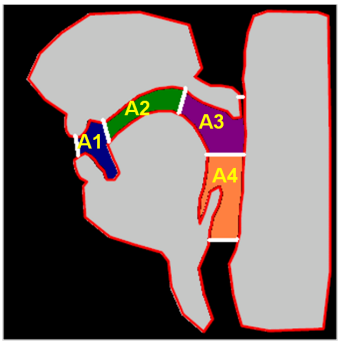

Fig. 1: *A schematic of vocal tract area descriptors (VTADs). VTAD A1 roughly approximates area enclosed from the lips to the tongue tip, A2 – from the tongue tip to dorsum, and A3 – tongue dorsum to pharynx. JawAngle is the obtuse angle measured between the pharyngeal wall and a regression line fitted to the jaw [5].*

The database used consisted of read speech (TIMIT sentences, rainbow passage), spontaneous speech (30 second responses to questions like "tell us about your favorite food") and 7 arias (2 specific and 5 arbitrary, in languages of their choice) elicited from 4 healthy sopranos who are native speakers of American English. 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

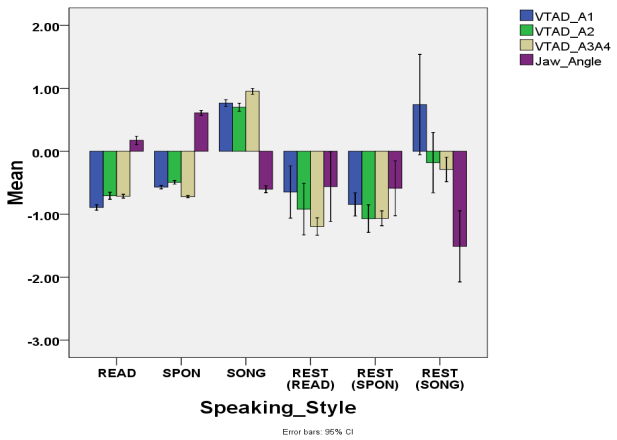

Fig. 2: *Differences in vocal tract postures assumed in read speech, spontaneous speech and singing on average. Please see Figure 1 for details on vocal tract descriptors.*

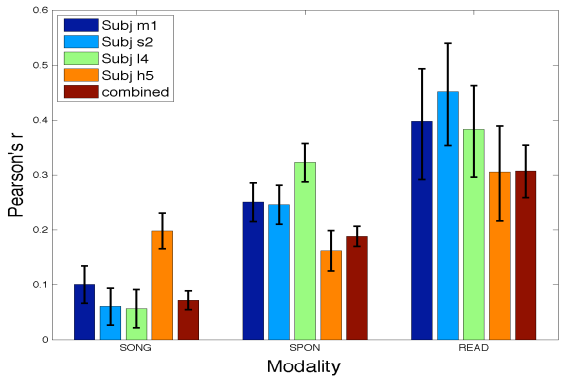

Fig. 3: *Correlations between articulatory change and acoustic change. Substantial differences in the strength of this correlation can be seen across behavioral modalities.*

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. Field-of-view (FOV), which can be thought of as a zoom factor, was set depending on the subjects head size.

Our results are as follows: Firstly, we find no significant differences in articulator speed on average (as measured using a techniques described in [4]) before, during and after pauses in musical arias, as opposed to the case of grammatical speech where we see a gradual decrease in speed. This might be in conformation with the hypothesis that articulatory/acoustic targets during pauses in singing are a critical part of the score, and thus, in effect, have a fixed timing specified in a similar manner to the musical phrases. Secondly, we find significant differences between the postures assumed (as measured using techniques described in [5]) during read speech, spontaneous speech and singing, which suggest that each of these have different dynamic planning and execution constraints (please see Figure 2). More specifically, we see a significantly more open vocal tract posture in the case of singing as opposed to read and spontaneous speech, which makes sense given the vocal tract shaping constraints required in the case of singing. In addition, these rest postures vary much more than those during pauses at phrasal breaks in both speech and song, suggesting that linguistic pauses are under greater active control than rest positions.

These differences are apparent in the interaction of articulation and acoustics, even at the coarsest level. We correlated articulator speed with the overall speed of spectral change (defined as the sum of absolute changes in mel-frequency cepstral coefficients (MFCCs) 2–13 for successive frames). Interestingly, singing shows a very small correlation when compared with read speech (see Figure 3), which potentially results from several phenomena. For example, acoustic changes in speech may arise primarily from oral and pharyngeal articulations in the midsagittal plane, while singing modulates the acoustic signal in other ways (e.g., in the larynx or off the midsagittal plane). Alternatively, singing may also involve manipulation of the forward map with the goal of minimizing the acoustic correlates of articulation. Either way, it is clear that the forward map is being utilized in different ways for each modality. Extensions of this work will include more detailed analysis and examination of additional modalities, such as beatboxing for which we have data. [Supported by NIH]

REFERENCES

[1] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," *The Journal of the Acoustical Society of America*, vol. 115, p. 1771, 2004.

[2] A. Tierney, F. Russo, and A. Patel, "Empirical comparisons of pitch patterns in music, speech, and birdsong," *Journal of the Acoustical Society of America*, vol. 123, no. 5, p. 3721, 2008.

[3] A. Patel, J. Iversen, and J. Rosenberg, "Comparing the rhythm and melody of speech and music: The case of British English and French," *The Journal of the Acoustical Society of America*, vol. 119, p. 3034, 2006.

[4] V. Ramanarayanan, E. Bresch, D. Byrd, L. Goldstein, and S. Narayanan, "Analysis of pausing behavior in spontaneous speech using real-time magnetic resonance imaging of articulation," *The Journal of the Acoustical Society of America*, vol. 126, no. 5, 2009.

[5] V. Ramanarayanan, D. Byrd, L. Goldstein, and S. Narayanan, "Investigating articulatory setting – pauses, ready position, and rest – using real-time MRI," *Interspeech, Makuhari, Japan*, 2010.