Support vector machines can decode speech patterns from high speed dynamic spiral FLASH images of the mouth

S. LaConte[1], J. Lisinski[1], and B. Sutton[2]
[1]School of Biomedical Engineering and Sciences, Virginia Tech, Blacksburg, VA, United States, [2]Bioengineering, University of Illinois, Urbana-Champaign, Urbana, IL, United States

INTRODUCTION

Speech is a vital but fragile human characteristic. It is susceptible to developmental disorders such as Down's syndrome and cleft palate as well as to insult from cerebrovascular accidents, traumatic brain injury, and neurodegenerative diseases such as multiple sclerosis and Parkinson disease. Unfortunately, though, existing monitoring methods have drawbacks that include being indirect and insensitive, being incapable of measuring all of the relevant anatomy, and/or requiring insertion into the mouth and thus interfering with the speech production itself. Because of MRI's excellent soft-tissue contrast and non-invasive nature, it holds great potential as a method to visualize the dynamics of structures in the oropharyngeal region.

Using a recently developed multi-shot, field-corrected, dynamic spiral FLASH sequence [1], we have recorded the oropharyngeal cavity (including the tongue, lips, and soft palate) at 15.8 frames per second. Here we report initial findings exploring the extent to which speech-related information is captured by this MR pulse sequence. During image acquisition, we asked a subject to perform a visually guided speech task, consisting of alternating 20 sec. blocks of slow and fast counting. We performed a support vector machine (SVM) analysis of these data and obtained 88% prediction accuracy when classifying individual frames as either "fast" or "slow" speech. This achievement could ultimately provide the basis for MRI-based lip-reading and has potential applications in speech therapy and diagnosis.

METHODS

Imaging: We used a custom 6-shot FLASH spiral sequence (TR/TE=6.7/(0.9 or 1.4) ms, 64 x 64 matrix with 120 mm FOV and 6.5 mm slice thickness, collected at 15.8 fps. The alternating echo time was used to perform magnetic field inhomogeneity corrected image reconstruction [2]. Images were acquired using a single-slice midsagittal acquisition. The subject performed a visually guided counting task, cued to count slowly for 20 s and then rapidly for 20 s for a total of 3 blocks of each condition (1896 images/120 s)
Analysis: SVM classification analysis was performed with *3dsvm* [3] in AFNI [4]. Each frame corresponded to either fast or slow counting. Cross-validation was used to estimate classification accuracy. Specifically every combination of training an SVM model with 4 blocks (2 slow and 2 fast) and testing with the remaining 2 (1 slow and 1 fast) was applied.

RESULTS

The cross-validation analysis led to individual accuracies of 91.9%, 80.0%, 95.0%, 84.6%, 90.2%, and 88.2% - the mean prediction accuracy was 88.3%. Figure 1B shows a thresholded SVM weight vector map that indicates the anatomical regions that were most important to the classifier. Inspection of individual voxel time series demonstrates the rich signal structure present in these data.

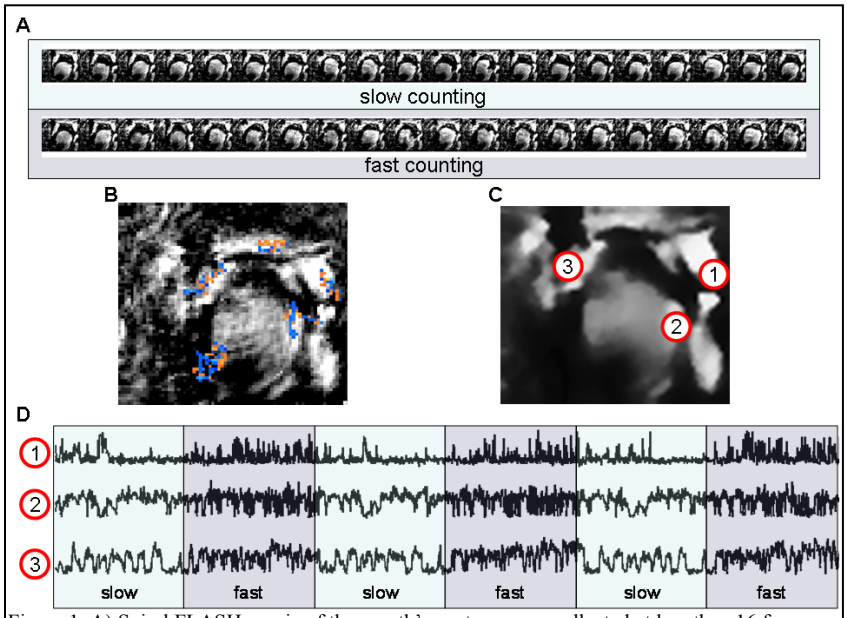

Figure 1. A) Spiral FLASH movie of the mouth's anatomy was collected at less than 16 fps during slow and fast counting. B) SVM model (88% accurate) shows importance of lips, tip and base of tongue, as well as soft palate. C) 3 regions identified using the SVM model. D) Single voxel time series extracted from the regions shown in C (lips, tongue, soft palate).

DISCUSSION AND CONCLUSION

Fast imaging of the articulating anatomy in the oral and pharyngeal cavities is challenging because of air-tissue susceptibility mismatches that give rise to large field inhomogeneities. This study demonstrates the capabilities of our recently developed MR sequence. One novel means of extracting quantitative information from these data is through supervised learning-based analyses. Here we used the support vector machine to show that these data can capture important information about speech production. Moreover, the SVM model can be used to identify important anatomical structures related to different oral motor activity. Applications and extensions of this work include decoding the precise syllables or words being spoken (MRI-based mouth reading), extracting more refined behavioral descriptions for combined structural and functional studies, and monitoring speech as a tool for speech therapy and diagnosis.

REFERENCE [1] Sutton, et al. J Magn Reson Imaging. 2010 Nov;32(5):1228-37. [2]. Sutton, et al. IEEE Trans Med Imaging. 2003 Feb;22(2):178-88. [3] LaConte, S.M. et al. 2005. NeuroImage 26, 317-329. [4] Cox, R. W. 1996. Comp. and Biomed. Res 29, 162-173.
ACKNOWLEDGEMENT Support: The Bob and Janice McNair Foundation, NIH 5R03DC009676-02, R03EB012464-01, R33DA026086; and USAMRMC W81XWH-08-2-0144.