Sensitivity of quantitative RT-MRI metrics of vocal tract dynamics to image reconstruction settings

Johannes Töger¹, Yongwan Lim¹, Sajan Goud Lingala¹, Shrikanth Narayanan, Krishna Nayak¹

¹Electrical Engineering, University of Southern California, Los Angeles, CA, USA
töger@usc.edu, yongwanl@usc.edu, slingala@usc.edu, shri@sipi.usc.edu, knayak@usc.edu

Abstract

Real-time Magnetic Resonance Imaging (RT-MRI) is a powerful method for quantitative analysis of speech. Current state-of-the-art methods use constrained reconstruction to achieve high frame rates and spatial resolution. The reconstruction involves two free parameters that can be retrospectively selected: 1) the temporal resolution and 2) the regularization parameter \( \lambda \), which balances temporal regularization and fidelity to the collected MRI data. In this work, we study the sensitivity of derived quantitative measures of vocal tract function to these two parameters. Specifically, the cross-distance between the tongue tip and the alveolar ridge was investigated for different temporal resolutions (21, 42, 56 and 83 frames per second) and values of the regularization parameter. Data from one subject is included. The phrase 'one two three four five' was repeated 8 times at a normal pace. The results show that 1) a high regularization factor leads to lower cross-distance values 2) using a low value for the regularization parameter gives poor reproducibility and 3) a temporal resolution of at least 42 frames per second is desirable to achieve good reproducibility for all utterances in this speech task. The process employed here can be generalized to quantitative imaging of the vocal tract and other body parts.

Index Terms: MRI, real-time, image reconstruction, vocal tract

1. Introduction

The human upper airway is a complex soft-tissue organ that is involved in several critical functions, including speech, swallowing and breathing. In the context of speech production, the upper airway is often referred to as the vocal tract, and the associated organs that shape it are referred to as articulators. Using 2D real-time magnetic resonance imaging (2D RT-MRI), the anatomy and dynamic function of the upper airway can be visualized and quantified freely in any imaging plane at high frame rates, without known radiation risks to the patient [1], [2].

Using 2D RT-MRI with simultaneous audio recording, the structure and function of the vocal tract can be quantified. The simplest method of analysis is to use average pixel intensities in regions of interest (ROI:s) [3]–[5]. A more sophisticated analysis can be performed by segmenting the air-tissue boundaries along the whole vocal tract [6] or segmenting individual articulators to acquire detailed measures of articulatory function [7].

Previous studies have used 2D RT-MRI to investigate articulatory timing [8], [9], vocal tract shaping in the production of different sounds [10], [11], articulatory setting (the vocal tract configuration in pauses, ready position and rest) [12], vocal tract shaping in professional musicians [13], [14], and paralinguistic mechanisms such as beatboxing [15]. Furthermore, potential clinical applications include velopharyngeal insufficiency [16], characterizing speech post-glossectomy [17], [18], and swallowing disorders [19].

A constant challenge with 2D RT-MRI is the inherent tradeoff between spatial and temporal resolution, the latter being crucial for capturing dynamic speech events [1]. Undersampling of the MRI image data during acquisition combined with compressed sensing can be used to significantly improve temporal resolution [2]. However, image reconstruction involves two free parameters; 1) the reconstructed temporal resolution and 2) a regularization parameter that adjusts the balance between the fidelity to the raw MRI data and the temporal total variation constraint. Previously, these parameters have been chosen heuristically to maximize temporal resolution and subjective image quality [2].

However, no objective measures have previously been used to inform the choice of the free reconstruction parameters. In this work, we propose the use of the reproducibility of the scans as the desired quantity to guide the choice of the free parameters. Reproducibility is an important feature for all research on the vocal tract, since it provides the fundamental limit of the effect size and inter-group differences that can be studied. Furthermore, reproducibility is an important factor for clinical applications, where quantitative measures may be used to inform health care decisions.

Therefore, this work aims to explore the influence of temporal resolution and the reconstruction parameter \( \lambda \) on the mean and reproducibility of quantitative measures of speech derived from 2D RT-MRI scans of the human vocal tract.

2. Methods

2.1. Study population and speech task

The study was approved by the local institutional review board (IRB). Data from one healthy subject was included. The subject was naïve to the purpose of the study. The speech task consisted of counting the numbers "one-two-three-four-five". In total, eight (8) repetitions were performed in the same MRI scanning session.

2.2. Magnetic resonance imaging

All imaging was performed on a GE Signa Excite 1.5T scanner (General Electric, Little Chalfont, UK) with a custom eight-channel upper airway coil [2]. The coil consists of two arrays of
The cross-distance at the alveolar ridge is measured as the average of the cross-distance of four gridlines. The cross-distance range over time was determined, and the cross-distance range for each of the 8 gridlines was measured as the average of the cross-distance of four gridlines. Panel C shows the cross-distance over time during one repetition of the word ‘one’. The cross-distance range is computed as the difference between the maximum and minimum distance over time for each word.

To visualize the achieved balance between the temporal regularization and the data fidelity for the range of $\lambda$ used, the data fidelity (first term of Equation 1) was plotted against the temporal regularization term (second term of Equation 2) in an L-curve analysis [24]. The L-curve has previously been used to identify the optimal value of $\lambda$ by locating a distinct corner in the curve [24].

2.4. Data analysis
For each combination of temporal resolution and reconstruction parameter $\lambda$, the resulting image data was analyzed using a semi-automatic segmentation tool [6]. The method requires a manual initialization of the vocal tract mid-line, and then automatically segments the airway cross-distance along the length of the vocal tract. The same initialization was used for all combinations of temporal resolution and $\lambda$. Figure 1 shows a schematic view of the analysis. First, 90 gridlines are placed along the initialization of the vocal tract. The tool then automatically segments the airway cross-distance for each line.

The cross-distance at the alveolar ridge was chosen as the target metric in this work. The rationale for this was that the potentially rapid motion and complex geometry of the tongue tip poses a challenge both to the image reconstruction and the segmentation method. Therefore, the cross-distance between the tongue and the alveolar ridge was measured as the average of the cross-distances of four gridlines (shown in yellow in Figure 1B). The average of four gridlines was taken to provide stability of the measures with respect to noise and segmentation errors.

Each occurrence of the spoken words were automatically aligned using Gentle, a freely available forced alignment tool [https://lowerquality.com/gentle/]. For each occurrence of a word, the cross-distance range over time was determined, defined as the difference between the maximum and the minimum cross-distance during the word (Figure 1C).

After determining the cross-distance range for each of the 8 repetitions of each word, the variability was computed as standard deviation (SD) over the ranges. Low SD was interpreted as good reproducibility, and high SD as poor reproducibility.
3. Results

Figure 2 shows L-curves for the reconstruction, showing the balance between data fidelity and temporal regularization for different choices of the regularization parameter $\lambda$. No distinct corner can be found in the curves, giving no guidance to the optimal choice of $\lambda$.

Figure 3 shows RT-MRI reconstructions using selected temporal resolutions and regularization parameters ($\lambda$). Visually, a lower temporal resolution or higher $\lambda$ gives a temporally smoothed result.

The semi-automatic segmentation tool gave interpretable results in 30/32 reconstructed image datasets (94%). The two reconstructions where the tool failed were for the combination of the lowest value of $\lambda$ ($\lambda=0.00025$) and the two highest temporal resolutions (83.3 and 55.5 fps).

Figure 4 shows the mean alveolar ridge cross-distance range over the 8 repetitions for each word. The words ‘two’ and ‘four’ are not shown, for reasons of keeping the presentation clear. There was a clear trend towards lower mean values for higher $\lambda$, and a weaker trend towards higher mean values for higher temporal resolutions.

Figure 5 shows quantitative reproducibility results, as standard deviation (SD) of the cross-distance range over the 8 repetitions for each word. When comparing different values of $\lambda$, the SD increases for $\lambda$ below 0.002 for the words ‘one’ and ‘three’. When comparing different temporal resolutions, there are no clear trends for the words ‘one’ and ‘three’. However, for the word ‘five’, the SD is higher for the lowest temporal resolution (21 fps).

4. Discussion

This work investigates the sensitivity of quantitative metrics of dynamic vocal tract function to choice of reconstruction parameters (temporal resolution and regularization parameter $\lambda$) for real-time vocal tract MRI using compressed sensing reconstruction.

The results show that the reconstruction parameter $\lambda$ has a significant influence on the mean of the resulting quantitative measures of vocal tract function, at least for the speech task and analysis method employed in this work. The trend towards lower values for higher $\lambda$ suggests that temporal smoothing leads to underestimation of the cross-distance. Therefore, a sufficiently low $\lambda$ is needed to capture temporal dynamics.
However, setting $\lambda$ too low (e.g. lower than 0.002) gives increased variability in derived quantitative articulatory measures, signaling poor reproducibility. This may be due to increased noise and aliasing artifacts that appear for low $\lambda$, as observed in a previous study [2]. The increased noise may in turn influence the automated image analysis negatively, leading to poor reproducibility. To achieve high reproducibility and avoid potential temporal smoothing of the images, choosing $\lambda=0.002$ is a good tradeoff for this combination of speech task and automated image analysis. Other choices of $\lambda$ may be beneficial for other speech tasks and analysis methods, and need further investigation.

We found no distinct corner in the L-curve to guide the choice of $\lambda$. Lingala et al. [2] used a visual analysis of reconstructed image quality to determine an appropriate $\lambda$. However, their $\lambda$ can not be directly compared to the one used here due to different normalizations of image data intensity.

In contrast to the parameter $\lambda$, there was a weaker dependence on temporal resolution, both for mean and standard deviation results. In one of the utterances (the word ‘five’), the lowest temporal resolution (21 fps) gives high standard deviation, signaling poor reproducibility. This may be due to fast movement of the tongue tip in this task that may be blurred by the low temporal resolution. However, it is not clear why this effect is not evident in the other utterances. Further investigation of other speech tasks is needed to determine the benefit of high temporal resolution, e.g. in fast and/or natural speech.

5. Conclusions

The results of this study show that 1) the reconstruction parameter $\lambda$ in constrained reconstruction of 2D RT-MRI can significantly influence quantitative measures of vocal tract function, 2) choosing a too small of a value for $\lambda$ gives poor reproducibility, and 3) a reconstructed temporal resolution of at least 42 fps is needed to achieve good reproducibility for tongue tip motion for a simple scripted speech task at a normal pace. Even higher temporal resolution may be beneficial for fast and/or natural speech, or in specialized applications such as singing or beatboxing.

6. Acknowledgements

This work was supported the NSF and by NIH grant DC007124. Tanner Sorensen at the Department of Linguistics and the Signal Interpretation and Analysis Laboratory, University of Southern California, Los Angeles, CA, is acknowledged for fruitful discussions that greatly improved this work.
7. References


