INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

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

Johannes Töger[1], *Yongwan Lim*[1], *Sajan Goud Lingala*[1], *Shrikanth Narayanan, Krishna Nayak*[1]

[1]Electrical Engineering, University of Southern California, Los Angeles, CA, USA

toger@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 λ, 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 velo-pharyngeal 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 λ 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

http://dx.doi.org/10.21437/Interspeech.2016-168

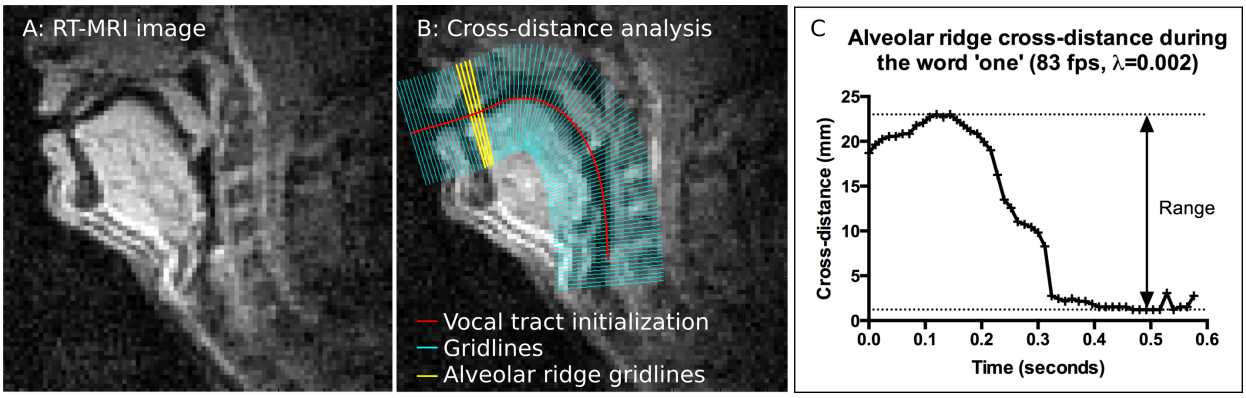

Figure 1: Data analysis. Panel A shows one frame of a RT-MRI image series. Panel B shows the semi-automatic segmentation tool [6]. First, the vocal tract is manually initialized (red). Grid lines are automatically placed over the whole vocal tract, and cross-distances computed. The cross-distance at the alveolar ridge is 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.

4 channels, positioned on each side of the subject's jaw to enable maximal signal from the upper airway.

Images were acquired in the mid-sagittal plane using a bit-reversed real-time spiral sequence based on the RTHawk research platform (HeartVista, Los Altos, CA, USA) [2], [20]. Sequence parameters were as follows: in-plane spatial resolution 2.4x2.4 mm; slice thickness 6 mm; TR 6 ms; TE 0.8 ms; Flip angle 15° and 13 spiral interleaves for full (Nyquist) sampling. The scan plane was manually aligned with the mid-sagittal plane of the subject's vocal tract. The reconstructed temporal resolution varied between 21 and 83 frames per second, as detailed below.

2.3. Constrained reconstruction of MRI data

Images were reconstructed from raw data using constrained reconstruction based on finite differences along the temporal dimension [2]. The reconstructed image f is the solution to the optimization problem

$$\min_f \|A(f) - b\|_2^2 + \lambda \|D_t(f)\|_1. \qquad (1)$$

Here A is the forward model for the MRI acquisition (including Fourier transformation, data sampling strategy and coil sensitivities), f is the reconstructed image, b is the measured raw MRI data, D_t is the temporal finite difference operator, and λ is the regularization parameter that balances data fidelity (first term in Equation 1) and temporal regularization (second term in Equation 1). The reconstructed temporal resolution enters into the reconstruction through the forward model A and the organization of the raw data b, and the degree of temporal regularization through λ.

The reconstruction optimization problem (Equation 1) was solved with the Berkeley Advanced Reconstruction Toolbox (BART) [21]–[23] using the "parallel imaging and compressed sensing" command (`bart pics`). Reconstructions were performed with 2, 3, 4 and 8 MRI spirals per time frame, resulting in corresponding temporal resolutions of 83, 56, 42 and 21 frames per second. The regularization parameter λ was varied between the values 0.00025, 0.0005, 0.001, 0.0015, 0.002, 0.004, 0.008 and 0.0012. In total, 32 reconstructions were performed (4 temporal resolutions and 8 levels for λ).

To visualize the achieved balance between the temporal regularization and the data fidelity for the range of λ 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 λ by locating a distinct corner in the curve [24].

2.4. Data analysis

For each combination of temporal resolution and reconstruction parameter λ, 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 λ. 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.

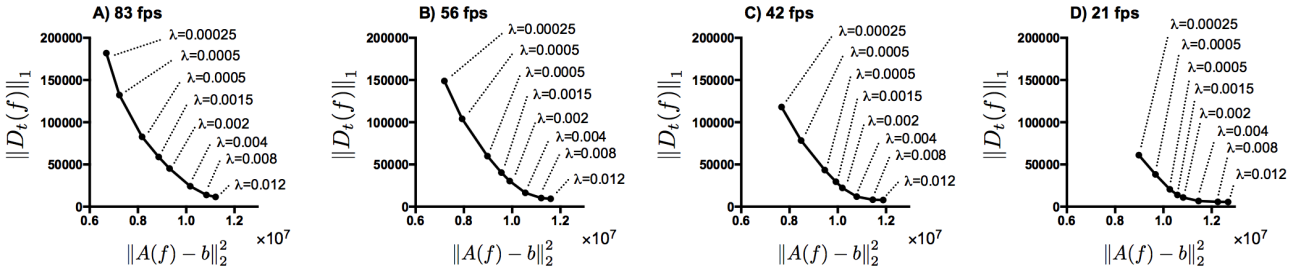

Figure 2: Constrained reconstruction L-curves showing the balance between the data fidelity term (x-axis) and the temporal regularization (y-axis) for different choices of the regularization parameter λ. Panel A shows results for reconstructions at 83 fps, Panel B for 56 fps, Panel C for 42 fps and Panel D for 21 fps. No clear corner can be seen in the curve, giving no guidance to the optimal choice of λ.

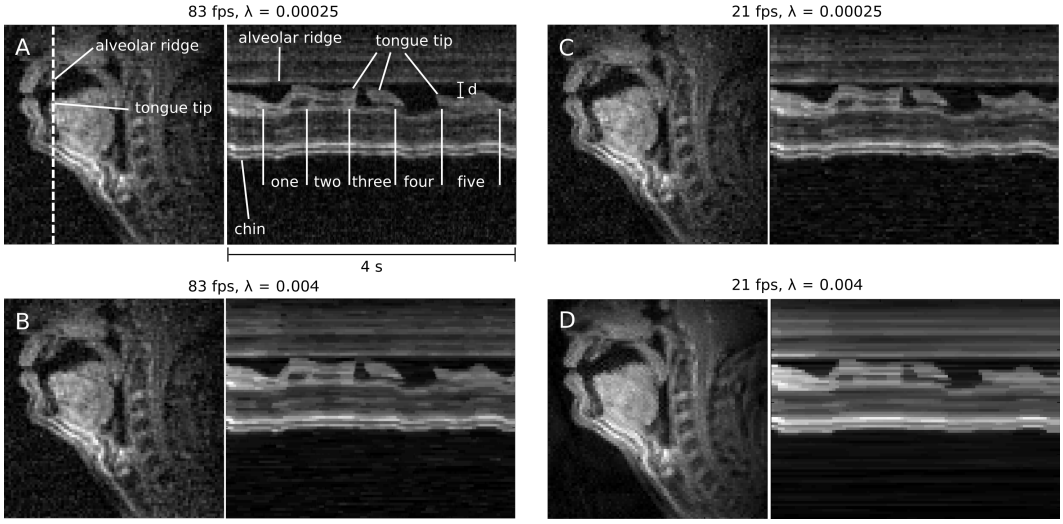

Figure 3: Example image reconstructions at selected temporal resolutions and regularization parameters λ. To the left, one image frame is shown. To the right, one cross-section is shown over time from each image (Panel A, white dashed line). Lower temporal resolution and higher λ gives a less noisy, but temporally smoothed image.
fps = frames per second, d = alveolar ridge cross-distance.

3. Results

Figure 2 shows L-curves for the reconstruction, showing the balance between data fidelity and temporal regularization for different choices of λ. No distinct corner can be found in the curves, giving no clear guidance in the choice of λ. Figure 3 shows RT-MRI reconstructions using selected temporal resolutions and regularization parameters (λ). Visually, a lower temporal resolution or higher λ 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 λ (λ=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 λ, 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 λ, the SD increases for λ 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 λ) for real-time vocal tract MRI using compressed sensing reconstruction.

The results show that the reconstruction parameter λ 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 λ suggests that temporal smoothing leads to underestimation of the cross-distance. Therefore, a sufficiently low λ is needed to capture temporal dynamics.

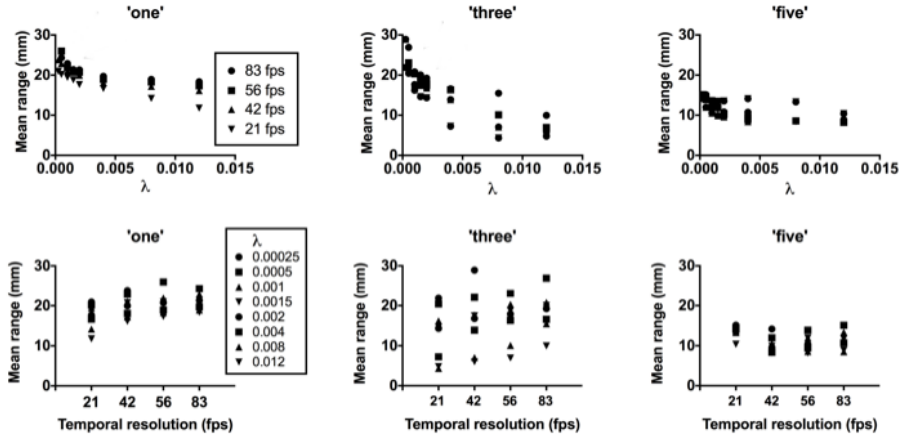

Figure 4: Mean of the alveolar ridge cross-distance range over 8 repetitions. In the top row, a clear trend towards lower cross-distance ranges can be seen for the words 'one' and 'three'. In the bottom row, there is a weaker trend towards higher mean cross-distance for higher temporal resolutions for the words 'one' and 'three'.

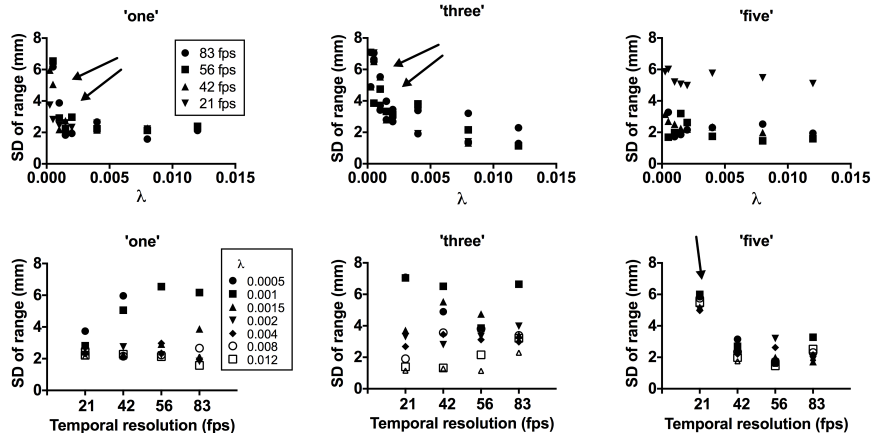

Figure 5: Quantitative reproducibility results: standard deviation (SD) of the alveolar ridge cross-distance over 8 repetitions. In the top row, it can be seen that for λ smaller than 0.002 (arrows), the SD increases for the words 'one' and 'three', which indicates poor reproducibility. In the bottom row, no clear trends with respect to temporal resolution can be seen for the words 'one' and 'three'. However, for the word 'five', the SD is higher for the lowest temporal resolution (21 fps, arrow).

However, setting λ 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 λ, 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 λ=0.002 is a good tradeoff for this combination of speech task and automated image analysis. Other choices of λ 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 λ. Lingala et al. [2] used a visual analysis of reconstructed image quality to determine an appropriate λ. However, their λ can not be directly compared to the one used here due to different normalizations of image data intensity.

In contrast to the parameter λ, 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 reproduciblity. 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 λ 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 λ 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

[1] S. G. Lingala, B. P. Sutton, M. E. Miquel, and K. S. Nayak, "Recommendations for real-time speech MRI," *J. Magn. Reson. Imaging*, vol. 43, no. 1, pp. 28–44, Jan. 2016.

[2] S. G. Lingala, Y. Zhu, Y. Kim, A. Toutios, S. Narayanan, and K. S. Nayak, "A fast and flexible MRI system for the study of dynamic vocal tract shaping," *Magn. Reson. Med.*, Jan. 2016.

[3] M. Proctor, A. Lammert, A. Katsamanis, L. Goldstein, C. Hagedorn, and S. Narayanan, "Direct estimation of articulatory kinematics from real-time magnetic resonance image sequences," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, pp. 281–284, 2011.

[4] E. Bresch, N. Katsamanis, L. Goldstein, and S. Narayanan, "Statistical multi-stream modeling of real-time MRI articulatory speech data," *Proc. 11th Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH 2010*, pp. 1584–1587, 2010.

[5] A. C. Lammert, M. I. Proctor, and S. S. Narayanan, "Data-Driven Analysis of Realtime Vocal Tract MRI using Correlated Image Regions," *INTERSPEECH Annu. Conf. Int. Speech Commun. Assoc.*, no. September, pp. 1572–1575, 2010.

[6] J. Kim, N. Kumar, S. Lee, and S. Narayanan, "Enhanced airway-tissue boundary segmentation for real-time magnetic resonance imaging data," *Proc. 10th Int. Semin. Speech Prod.*, pp. 222–225, 2014.

[7] E. Bresch and S. Narayanan, "Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images," *IEEE Trans. Med. Imaging*, vol. 28, no. 3, pp. 323–338, 2009.

[8] M. Proctor, L. Goldstein, A. Lammert, D. Byrd, A. Toutios, and S. Narayanan, "Velic coordination in French nasals: A real-time magnetic resonance imaging study," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, no. August, pp. 577–581, 2013.

[9] D. Byrd, S. Tobin, E. Bresch, and S. Narayanan, "Timing effects of syllable structure and stress on nasals: A real-time MRI examination," *J. Phon.*, vol. 37, no. 1, pp. 97–110, 2009.

[10] E. Bresch, D. Riggs, L. Goldstein, D. Byrd, S. Lee, and S. Narayanan, "An analysis of vocal tract shaping in English sibilant fricatives using real-time magnetic resonance imaging," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, pp. 2823–2826, 2008.

[11] C. Smith, "Complex Tongue Shaping in Lateral Liquid Production Without Constriction-Based Goals," *Proc. 10th Int. Semin. Speech Prod.*, pp. 413–416, 2014.

[12] V. Ramanarayanan, A. Lammert, L. Goldstein, and S. Narayanan, "Are articulatory settings mechanically advantageous for speech motor control?," *PLoS One*, vol. 9, no. 8, p. e104168, 2014.

[13] E. Bresch and S. Narayanan, "Real-time magnetic resonance imaging investigation of resonance tuning in soprano singing," *J. Acoust. Soc. Am.*, vol. 128, no. 5, p. EL335, 2010.

[14] P. W. Iltis, E. Schoonderwaldt, S. Zhang, J. Frahm, and E. Altenmüller, "Real-time MRI comparisons of brass players: A methodological pilot study.," *Hum. Mov. Sci.*, vol. 42, pp. 132–145, 2015.

[15] M. Proctor, E. Bresch, D. Byrd, K. Nayak, and S. Narayanan, "Paralinguistic mechanisms of production in human 'beatboxing': a real-time magnetic resonance imaging study.," *J. Acoust. Soc. Am.*, vol. 133, no. 2, pp. 1043–54, 2013.

[16] A. D. Scott, R. Boubertakh, M. J. Birch, and M. E. Miquel, "Towards clinical assessment of velopharyngeal closure using MRI: Evaluation of real-time MRI sequences at 1.5 and 3T," *Br. J. Radiol.*, vol. 85, no. 1019, pp. 1083–1092, 2012.

[17] M. Stone, J. M. Langguth, J. Woo, H. Chen, and J. L. Prince, "Tongue motion patterns in post-glossectomy and typical speakers: a principal components analysis.," *J. Speech. Lang. Hear. Res.*, vol. 57, no. 3, pp. 707–17, Jun. 2014.

[18] C. Hagedorn, A. Lammert, M. Bassily, Y. Zu, U. Sinha, L. Goldstein, and S. S. Narayanan, "Characterizing Post-Glossectomy Speech Using Real-time MRI," in *International Seminar on Speech Production, Cologne, Germany*, 2014.

[19] Y. Zu, S. S. Narayanan, Y.-C. Kim, K. Nayak, C. Bronson-Lowe, B. Villegas, M. Ouyoung, and U. K. Sinha, "Evaluation of swallow function after tongue cancer treatment using real-time magnetic resonance imaging: a pilot study.," *JAMA Otolaryngol. Head Neck Surg.*, vol. 139, no. 12, pp. 1312–9, 2013.

[20] S. Narayanan, K. Nayak, S. B. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," *J. Acoust. Soc. Am.*, vol. 115, no. 4, pp. 1771–1776, 2004.

[21] J. I. Tamir, F. Ong, J. Y. Cheng, M. Uecker, and M. Lustig, "Generalized Magnetic Resonance Image Reconstruction using The Berkeley Advanced Reconstruction Toolbox," in *ISMRM Workshop on Data Sampling and Image Reconstruction, Sedona 2016*, 2016.

[22] M. Uecker, F. Ong, J. I. Tamir, D. Bahri, P. Virtue, J. Y. Cheng, T. Zhang, and M. Lustig, "Berkeley Advanced Reconstruction Toolbox," in *In Proc. Intl. Soc. Mag. Reson. Med. 23:2486*, 2015.

[23] M. Uecker, P. Virtue, F. Ong, and M. Murphy, "Software toolbox and programming library for compressed sensing and parallel imaging," *ISMRM Work. Data Sampl. Image Reconstr. Sedona 2013*, 2013.

[24] P. C. Hansen, "Analysis of Discrete Ill-Posed Problems by Means of the L-Curve," *SIAM Rev.*, vol. 34, no. 4, pp. 561–580, 1992.