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

Improved Depiction of Tissue Boundaries in Vocal Tract Real-time MRI using Automatic Off-resonance Correction

Yongwan Lim, Sajan Goud Lingala, Asterios Toutios, Shrikanth Narayanan, Krishna S. Nayak

Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, California, USA

yongwanl@usc.edu, lingala@usc.edu, toutios@sipi.usc.edu, shri@sipi.usc.edu, knayak@usc.edu

Abstract

Real-time magnetic resonance imaging (RT-MRI) is a powerful tool to study the dynamics of vocal tract shaping during speech production. The dynamic articulators of interest include the surfaces of the lips, tongue, hard palate, soft palate, and pharyngeal airway. All of these are located at air-tissue interfaces and are vulnerable to MRI off-resonance effect due to magnetic susceptibility. In RT-MRI using spiral or radial scanning, this appears as a signal loss or blurring in images and may impair the analysis of dynamic speech data. We apply an automatic off-resonance artifact correction method to speech RT-MRI data in order to enhance the sharpness of air-tissue boundaries. We demonstrate the improvement qualitatively and using an image sharpness metric offering an improved tool for speech science research.

Index Terms: off-resonance correction, deblurring, real-time MRI, speech production, vocal tract shaping

1. Introduction

For speech production research, real-time magnetic resonance imaging (RT-MRI) has several advantages over other modalities such as x-ray fluoroscopy, electromagnetic articulography, and ultrasound [1-3]. RT-MRI provides non-invasive depiction of the dynamics of deep articulatory structures (e.g., pharynx, glottis and epi-glottis) with excellent contrast, and allows for arbitrary imaging planes. In this context, spiral RT-MRI scanning is desirable because it allows for a highly time efficient acquisition, given that spirals provide higher spatio-temporal resolution than alternative schemes. Furthermore, a golden-ratio spiral MRI allows for the flexible selection of temporal resolution retrospectively which is desirable because of variations in speech rate and in the speed of different articulators [4,5].

A major drawback of spiral MRI, however, is the effect of off-resonance referred to as deviation in resonance frequency from the signal receive frequency in MRI, which is due to the magnetic susceptibility difference between air and tissue. This results in signal loss and/or blurring artifacts near the air-tissue boundaries [6]. The artifact is more pronounced with longer spiral readout or on higher strength MRI scanners. To mitigate this artifact, current speech RT-MRI is most often conducted with very short (~2.5 ms) spiral readouts and on lower field (1.5 Tesla) MRI scanners [2,4,5].

From speech analysis point of view, the analysis of the dynamic articulators of interest may be challenging in the presence of off-resonance induced blurring artifacts. For example, average pixel intensities in regions of interest for the vocal tract constriction information [7,8] may be prone to area perturbation due to the blurring artifacts. Also, air-tissue boundary segmentation [9,10,11] that is typically required as pre-processing to acquire vocal tract variables such as area functions [12] may suffer from an ambiguous boundary with poor contrast due to the blurring artifacts. Therefore, it may impair the accurate analysis of the dynamic speech data obtained using spiral RT-MRI.

The goal of this study is to evaluate an off-resonance correction method that is directly applicable to current speech RT-MRI datasets and may be used to further improve the quality of previously corrected RT-MRI data [13-15]. Several potential methods exist in the literature [16-22] and prior work has shown that the deblurred images can be restored by acquiring the off-resonance frequency information called field map [16-18]. However, these methods are not directly applicable to RT-MRI applications because they require additional scan time to acquire a dynamic field map, thereby compromising temporal or spatial resolution [19, 23].

An alternative approach to acquiring the field map is to estimate it directly from the dataset itself, known as "auto-focus" or automatic off-resonance correction [19-22]. The auto-focus methods estimate field map based on a focus metric that can provide information about the degree of blurriness due to the off-resonance effects. For example, the absolute value of the imaginary component of the image at a specific image location can indicate the degree of the off-resonance effect at that location because theoretically the imaginary components of the image should be zero when the image is corrected. Although the methods that estimate the field map have shown comparable results to the methods that acquire the field map, performance still depends on the focus metric used and experimental factors such as MRI sequence parameters, subjects, and off-resonance effects.

In this study, we present an automatic off-resonance correction method that uses a modified focus metric and is applicable to spiral RT-MRI data. We apply the method to the USC-TIMIT dataset [13] and demonstrate the improvement in the air-tissue boundaries qualitatively and using an image sharpness metric. This approach has the potential to improve the analysis of articulator dynamics in spiral RT-MRI.

2. Materials and Methods

2.1. Datasets

Experiments were performed using the USC-TIMIT dataset [13] which is a large corpus of RT-MRI scans of the vocal tract during speech production (RT-MRI image sequences with

http://dx.doi.org/10.21437/Interspeech.2016-664

synchronous noise-cancelled audio) where the stimuli consisted of 460 sentences of the MOCHA-TIMIT corpus [24]. The corpus provides data from ten speakers, each of which consists of 92 video files each with 5 sentences. The degree of the blurring artifacts in their images vary depending on the speech tasks. We selected three representative speakers (M2, F2, and F4) out of the ten speakers, which, on visual assessment, presented the most significant blurring artifacts.

For USC-TIMIT datasets, a 13-interleaf spiral fast gradient echo pulse sequence was used and imaging was performed in the mid-sagittal plane. Imaging parameters used: spatial resolution = 2.4×2.4 mm^2, field of view = 200×200 mm^2, repetition time (TR) = 6.164 ms, echo time (TE) = 0.8 ms, receiver bandwidth = ±125 kHz. Image reconstruction used two anterior coil elements of a 4-channel upper airway receiver coil array with coil sensitivities and a sliding window technique with a frame rate of 23.18 frames/s.

2.2. Automatic Off-resonance Correction

The principle underlying the off-resonance correction is that the deblurred image at an image location can be obtained by compensating the raw MRI signal for the phase accrual due to off-resonance frequency of that image location. This procedure can be done by demodulating the MRI signal with the off-resonance frequency and then using gridding reconstruction with the demodulated MRI signal [16]. In auto-focus [19], the local off-resonance frequency (field map) is chosen as a frequency that minimizes a focus metric calculated at that location among a series of demodulation frequencies.

Auto-focus was performed using the method described in [19] but using a modified focus metric to consider the time dimension and multiple coil elements. The focus metric was:

$$S(x,y,t;\Delta f_i) = \sum_{(x,y,t)\in A(x,y,t)} |Im\{I(x,y,t;\Delta f_i)\}| \qquad (1)$$

where $\{\Delta f_i, i=1,2,\ldots,N_f\}$ is a set of equally spaced frequencies, $A(x, y, t)$ is a $w_x \times w_y \times w_t$ summation window centered at (x,y,t), and $I(x,y,t;\Delta f_i)$ is coil combined image reconstructed at a frequency Δf_i with the removal of image phase unrelated to the off-resonance effect. The metric here is calculated in the coil combined image, $I(x,y,t;\Delta f_i)$. To reduce the likelihood of getting stuck in local minima, the window dimension is extended in the temporal dimension, assuming the off-resonance frequency varies smoothly in time as well as in space

For each location at (x,y,t), the deblurred image $\bar{I}(x,y,t)$ is formed by choosing pixels from the base images $I(x,y,t;\Delta f_i)$ based on the metric as follows:

$$\bar{I}(x,y,t) = I\big(x,y,t;\Delta f(x,y,t)\big) \qquad (2)$$

where $\Delta f(x,y,t) = \underset{\Delta f_i}{\operatorname{argmin}} S(x,y,t;\Delta f_i)$ is the estimated field map that minimizes the metric at each location (x,y,t).

For the datasets analyzed in this study, auto focus was achieved using estimation in two stages (coarse and fine estimation) with the following parameters: the size of the summation window is w_x (= w_y) = 21~25 pixels and w_t = 7~11 time frames for coarse estimation and w_x (= w_y) = 7~9 pixels and w_t = 7~11 time frames for fine estimation, the range of the frequency is −120 to 120 Hz, and the step of the frequency is 20 Hz (N_f = 13).

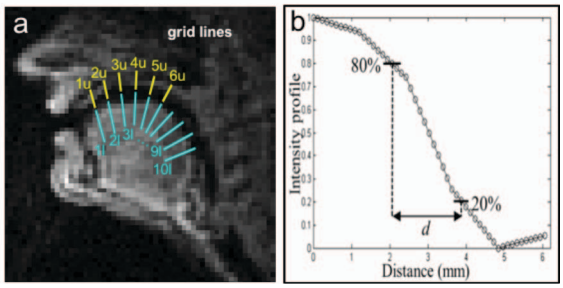

Figure 1: *Gridlines (cut-view lines) of lower tissue boundaries (marked as cyan solid lines; 1l~10l) and upper tissue boundaries (marked as yellow solid lines; 1u~6u) are extracted to obtain intensity profiles (panel a) and normalized intensity profile of the grid line (panel b) is plotted where a sharpness is measured using a reciprocal of the distance (d).*

2.3. Sharpness Evaluation

We define a quantitative metric for edge sharpness at air-tissue boundaries. Ten intensity profiles (1l-10l) perpendicular to tongue boundary and six (1u-6u) perpendicular to the hard palate boundary were selected by using the grid extraction method [10] (Figure 1(a)) and linearly interpolated to generate ten times higher spatial resolution. The distance (d) (mm) between the points of 80% and 20% of the maximum intensity value was measured (Figure 1(b)) and averaged over all time frames. Finally, sharpness is defined as the reciprocal of the averaged distance ($1/d$).

3. Results

Figure 2 compares images reconstructed using different off-resonance correction methods. Image in Figure 2(a) is the original blurry image without the correction method, images in Figure 2(b) and (c) are images after the correction with 2D spatial (i.e., w_t=1) and 3D spatio-temporal summation windows, respectively, and image in Figure 2(d) is the image corrected with a static field map. As shown with arrows in Figure 2, images in Figure 2(c) and (d) represent sharper and more clear boundary near the velum while off-resonance induced blurring artifacts are apparent near the velum in Figure 2(a) and (b). Therefore, the window extended in the temporal dimension is more desirable in order to avoid estimation errors. Note that although not shown in this work, using the static field map often yields estimation errors in the region where the articulators move rapidly.

Figure 3 contains representative images for three speakers (M2, F2, and F4) without and with the proposed auto-focus procedure. For every image reconstructed with the correction method, posterior to the alveolar ridge, the hard palate becomes more intense and sharper up to the velum compared to the original images. For speaker F2, the intensity around the alveolar ridge and the air-tongue boundary appears sharper in the deblurred images.

Figure 4 contains time-intensity profiles for the original and the deblurred image sequence where the profiles are extracted at the solid lines in the sample image frames. For speaker M2, the intensity in the hard palate in the deblurred image sequence is more constant along time than the intensity value in the

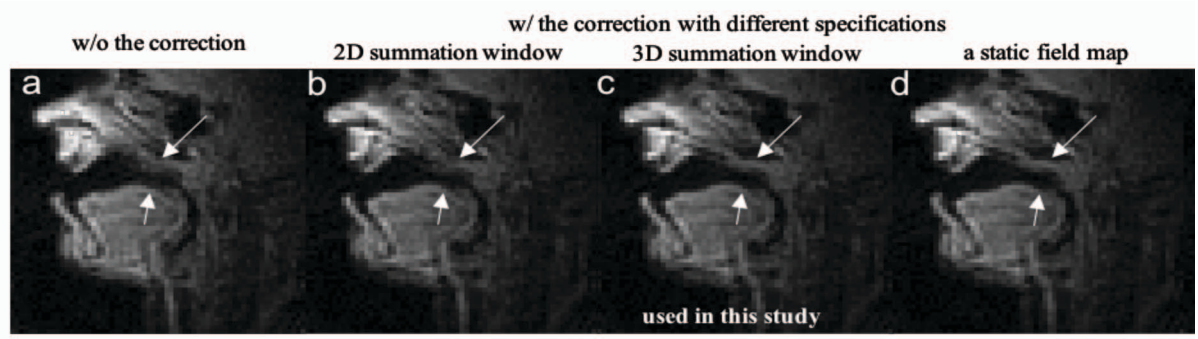

Figure 2: *Comparison of the results with different specifications in the off-resonance correction method. Original blurred image without off-resonance correction (a), result images after correction with a 2D spatial summation window (window size: 21×21×1) (b), with a 3D spatio-temporal summation window (window size: 21×21×7) which is used in this work (c), and with a static field map which was once estimated from a temporal averaged image and applied to whole time frame images with the estimated static field map (d). While off-resonance artifacts are apparent near velum in (a) and (b), (c) and (d) represent sharper and more clear boundary near velum.*

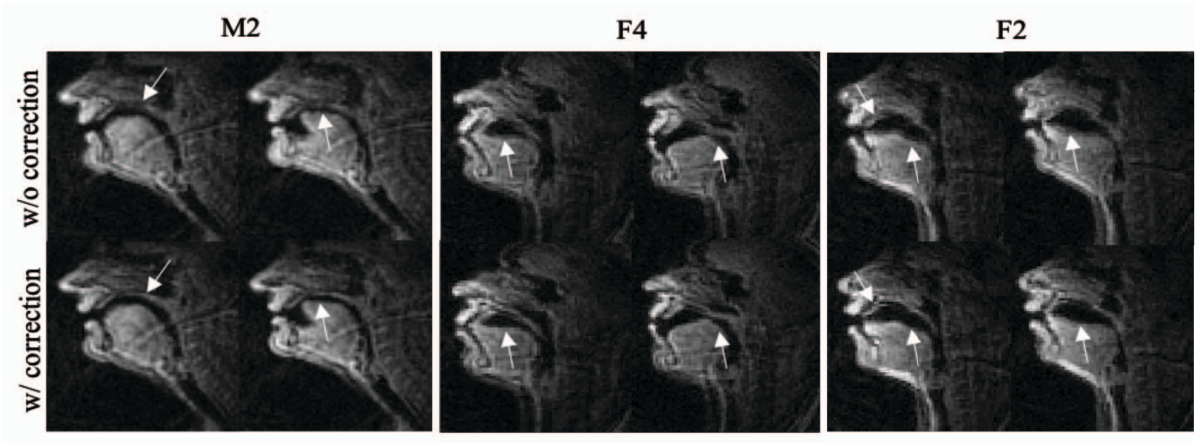

Figure 3: *Representative mid-sagittal image frames of vocal tracts for three speakers; M2, F4, and F2 in the USC-TIMIT dataset. Images in top row are the images without the off-resonance correction method and images in bottom row are the result images deblurred by the correction method. Images for each of columns are reconstructed at same time frame, respectively. Arrows point out the regions that are affected by the off-resonance effects.*

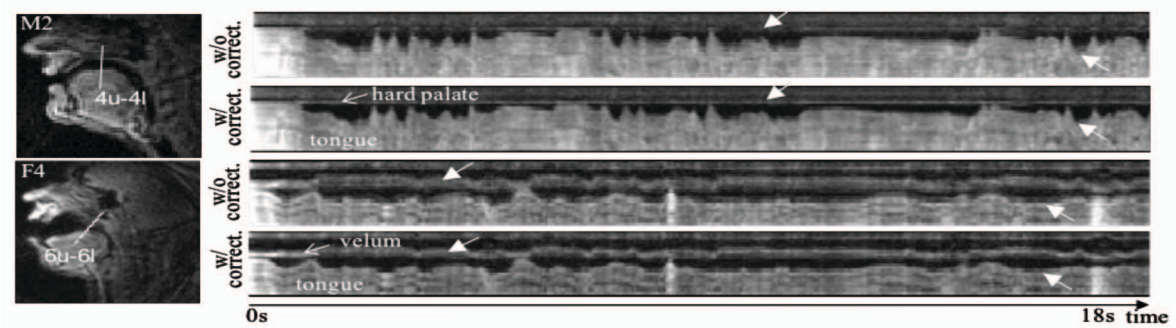

Figure 4: *Demonstration of improved image sharpness in tissue boundaries in intensity profile along time. An image of an example frame is shown (left) and the intensity time profiles are extracted at the solid lines in the left images (right) where the solid lines correspond to grid lines 4u-4l and 6u-6l shown in Figure 1, respectively.*

original image sequence. This result agrees with the fact that the hard palate, which is a bony structure covered by a thin layer of tissue, does not change its shape during speech production [11]. Furthermore, the intensity profile from the deblurred image exhibits sharper boundary between tongue and air. For speaker F4, the intensity profile from the deblurred image provides a clear delineation of the velum movements as well as a sharp boundary between tongue and air.

To quantitatively validate the improvement in the depiction of the vocal tract boundary, the sharpness scores were

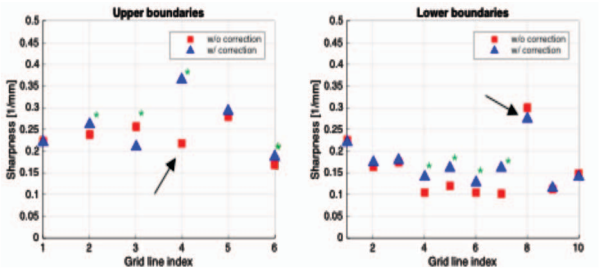

Figure 5: *Improvement of sharpness at different tissue boundary locations. For speaker M2, the sharpness values are measured at the gridlines of six upper boundaries (left) and ten lower boundaries (right) and averaged along time. The sharpness values marked as asterisk (*) represent significant differences between w/o correction and w/ correction ($p<0.01$).*

calculated at the different locations of tissue boundaries described in Figure 1. As shown in Figure 5, for the upper tissue boundaries, the deblurred image at the boundaries (2u, 4u, and 6u) shows higher sharpness scores than those of the original blurry image while the deblurred image at the boundaries (1u, 3u, and 5u) shows less improvement or lower sharpness score than those of the original blurry image. For the lower tissue boundaries, tongue boundaries at the middle of airway (4l-7l) shows the significant improvements of the sharpness scores in the deblurred image.

Figure 6 shows the sharpness score along time measured at the boundary (4u) where the sharpness is highly improved in Figure 5. The sharpness score at the boundary (4u) shows improvement over most time frames. This may be due to the fact that the hard palate is greatly restored over most of time in the deblurred image for speaker M2 as shown in Figure 4.

4. Discussion

We have shown that the automatic off-resonance correction method improves the depiction of the dynamic tissue boundaries that has the potential to improve the performance of the segmentation of vocal tract that is essential to analyze the articulatory movement patterns.

There are two possible advantages of using the auto-focus method. First, it would be applicable to any previously acquired spiral RT-MRI datasets because it does not require acquisition of the field map. Second, it would allow for a longer spiral readout sequence with a fewer number of spiral interleaves that is a highly time efficient acquisition [4]. This is because the auto-focus method can compensate for the more pronounced blurring artifacts in images with a longer spiral acquisition.

An important issue in the automatic correction method is the parameter selection for the focus metric. In the presence of substantial noise in the images, a summation window with an increased spatial area coverage is desirable, because the increased area can be considered as a noise-reducing low pass filter. However, at the same time it would reduce the ability to track fast spatial variations of the off-resonance frequency. The increased temporal coverage of the summation window can prevent from getting stuck in local minima in the area corresponding to static or slow articulators such as the hard palate and pharyngeal wall, but it is hard to capture the fast variation of the off-resonance in the vocal tract for fast-moving

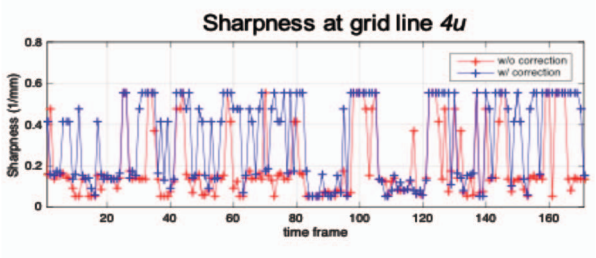

Figure 6: *Sharpness score along time. The sharpness score is measured along time at the boundary in the gridline 4u in which the most improvement of the sharpness score is observed among the different boundaries in Figure 5.*

articulators. As a result, there is a trade-off and the size of summation window $A(x,y,t)$ has to be chosen carefully.

We have measured the sharpness score in particular locations at tissue boundaries to quantitatively validate the effectiveness of the off-resonance correction. The more desirable way to evaluate the effectiveness would be to conduct segmentation of the vocal tract and then evaluate the improvement in the segmentation result. This can be done by comparing the segmentation results with manual segmentation results. However, because of the very large number of frames in the RT-MRI datasets, performing a manual segmentation is not practical. Hence, we are investigating a methodology to evaluate the segmentation results without manual reference.

In this work, we reconstructed the images based on gridding reconstruction. Future research is to combine the off-resonance correction method with constrained reconstruction such as parallel imaging compressed sensing reconstruction to achieve higher temporal resolution and better delineations of articulatory motion [5].

5. Conclusions

In this work, we present an automatic off-resonance correction method for vocal tract RT-MRI. We have applied the method to the USC-TIMIT dataset and shown that the method improves image sharpness of the vocal tract area such as the tongue boundaries, alveolar ridge, and velum which has the potential to improve the analysis of the dynamics of articulators.

6. Acknowledgements

This work was supported by NIH grant R01DC007124 and NSF grant 1514544. We acknowledge the support and collaboration of the Speech Production and Articulation kNowledge (SPAN) group at the University of Southern California, Los Angeles, CA, USA.

7. References

[1] E. Bresch, Y.-C. Kim, K. Nayak, D. Byrd, and S. Narayanan, "Seeing speech: Capturing vocal tract shaping using real-time MRI," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 123–132, 2008.

[2] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time Magnetic Resonance Imaging for speech production," *JASA*, vol. 115, no. 4, pp. 1771–1776, 2004.

[3] A. D. Scott, M. Wylezinskaa, M. J. Bircha, and M. E. Miquela, "Speech MRI: Morphology and function," *Phys. Med.* vol. 30, no. 6, pp. 604–618, 2014.

[4] Y-C. Kim, S. Narayanan, and K. Nayak, “Flexible retrospective selection of temporal resolution in real-time speech MRI using a golden-ratio spiral view order,” *Magn. Reson. Med.*, vol. 65, no. 5, pp. 1365–1371, 2011.

[5] S. Lingala, Y. Zhu, Y-C. 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.*, doi: 10.1002/mrm.26090, 2016.

[6] C. Meyer, B. S. Hu, D. G. Nishimura, and A. Macovski, “Fast spiral coronary artery imaging,” *Magn. Reson. Med.*, vol. 28, no. 2, pp. 202–213, 1992.

[7] 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”, in *Proc. of Interspeech*, pp. 281–284, Florence, Italy, Aug. 2011.

[8] A. Lammert, V. Ramanarayanan, M. Proctor, and S. Narayanan, “Vocal tract cross-distance estimation from real-time MRI using region-of-interest analysis,” in *Proc. of Interspeech*, pp. 959–962, Lyon, France, Aug. 2013

[9] M. Proctor, D. Bone, A. Katsamanis, and S. Narayanan, “Rapid semi-automatic segmentation of real-time magnetic resonance images for parametric vocal tract analysis,” in *Proc. of Interspeech*, pp. 1576–1579, Makuhari, Japan, Sep. 2010.

[10] J. Kim, N. Kumar, S. Lee, and S. Narayanan, "Enhanced airway-tissue boundary segmentation for real-time magnetic resonance imaging data," in *10-th Int. Seminar on Speech Production (ISSP)*, pp. 222–225, Cologne, Germany, 2014.

[11] 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, Mar. 2009.

[12] C. P. Browman and L. M. Goldstein, “Towards an articulatory phonology”, *Phonology Yearbook*, vol. 3, no. 1, pp. 219–252, 1986.

[13] S. Narayanan, A. Toutios, V. Ramanarayanan, A. Lammert, J. Kim, S. Lee, K. Nayak, Y.-C. Kim, Y. Zhu, L. Goldstein, D. Byrd, E. Bresch, P. Ghosh, A. Katsamanis, and M. Proctor, “Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC),” *The Journ. of the Acoust. Soc. of Am.*, vol. 136, no. 3, pp. 1307–1311, 2014.

[14] J. Kim, A. Toutios, Y.-C. Kim, Y. Zhu, S. Lee, and S. S. Narayanan, “USC-EMO-MRI corpus: An emotional speech production database recorded by real-time magnetic resonance imag- ing,” in *International Seminar on Speech Production (ISSP)*, Cologne, Germany, May 2014.

[15] A. Toutios and S. Narayanan, “Advances in real-time magnetic resonance imaging of the vocal tract for speech science and technology research,” *APSIPA Transactions on Signal and Information Processing*, 2016.

[16] D. C. Noll, C. H. Meyer, and J. M. Pauly, D. G. Nishimura, “A homogeneity correction method for magnetic resonance imaging with time-varying gradients,” *IEEE Trans. Med. Imaging*, vol. 10, no. 4, pp. 629–637, 1991.

[17] L. C. Man, J. M. Pauly, and A. Macovski, “Multifrequency interpolation for fast off-resonance correction,” *Magn. Reson. Med.* vol. 37, no. 5, pp. 785–792, 1997.

[18] K. S. Nayak, C. M. Tsai, C. H. Meyer, and D. G. Nishimura, “Efficient off-resonance correction for spiral imaging,” *Magn. Reson. Med.* vol. 45, no. 3, pp. 521–524, 2001.

[19] L. C. Man, J. M. Pauly, and A. Macovski, “Improved automatic off-resonance correction without a field map in spiral imaging,” *Magn. Reson. Med.* vol. 37, no. 6, pp. 906–913, June 1997.

[20] D. C. Noll, J. M. Pauly, C. H. Meyer, D. G. Nishimura, and A. Macovski, “Deblurring for non-2D Fourier transform magnetic resonance imaging,” *Magn. Reson. Med.* vol. 25, no. 2, pp. 319–333, June 1992.

[21] W. Chen and C.H. Meyer, “Fast automatic linear off-resonance correction method for spiral imaging,” *Magn. Reson. Med.* vol. 56, no. 2, pp. 457–462, Aug. 2006.

[22] T. B. Smith and K. S. Nayak, “Automatic off-resonance correction in spiral imaging with piecewise linear autofocus,” *Magn. Reson. Med.* vol. 69, no. 1, pp. 82–90, Jan. 2013.

[23] B. P. Sutton, C. A. Conway, Y. Bae, R. Seethamraju, and D. P. Kuehn, “Faster dynamic imaging of speech with field inhomogeneity corrected spiral fast low angle shot (FLASH) at 3 T,” *J Magn. Reson. Imaging*, vol. 32, no. 5, pp. 1228–1237, Nov. 2010.

[24] A. A Wrench and H. K. William, “A multichannel articulatory database and its application for automatic speech recognition,” in *Proc. of 5th Seminar of Speech Production: Models and Data*, pp. 305–308, Kloster Seeon, Germany, 2000.