Refining rtMRI Landmark-Based Vocal Tract Contour Labels with FCN-Based Smoothing and Point-to-Curve Projection

Mushaffa Rasyid Ridha[1], Sakriani Sakti[1,2]
[1]Japan Advanced Institute of Science and Technology, Japan
[2]Nara Institute of Science and Technology, Japan
{s2210427, ssakti}@jaist.ac.jp

Abstract

Advanced real-time magnetic resonance imaging (rtMRI) allows researchers to study articulatory movements during speech production with high temporal resolution. Nevertheless, accurately outlining articulator contours in high-frame-rate rtMRI presents challenges, given data scalability and image quality issues, which make manual and automatic labeling difficult. The widely used publicly available USC-TIMIT dataset provides rtMRI data with landmark-based contour labels for part of the data derived from unsupervised region segmentation using spatial frequency domain representation and gradient descent optimization. While this method yields high-quality labels, occasional labeling errors exist. Many contour detection methods were trained and tested based on this ground truth, which is not purely a gold standard label, and the resulting contour data remains largely undisclosed to the public. This paper offers a refinement of landmark-based vocal tract contour labels by employing outlier removal, a fully convolutional network (FCN)-based smoothing, and a landmark point-to-edge curve projection approach. In the absence of an established ground truth label, we evaluate the quality of the new labels through subjective assessments of several contour areas, comparing them to the existing data labels.

Keywords: Vocal Track, real-time MRI, landmark-based contour labels, fully convolutional network, point-to-curve projection

1. Introduction

Speech production, a key focus in linguistics, phonetics, and neuroscience, involves creating speech sounds. Researchers investigate how articulators shape acoustic features and brain-controlled speech, vital for linguistic analysis, speech modeling, articulatory control, and the articulatory-sound relationship (Ladefoged and Maddieson, 1996). To study speech production, techniques like X-ray imaging ((Westbury et al., 2005); Vorperian et al., 2009), Electromagnetic Articulography (EMA) (Perkell et al., 1992); (Cai et al., 2018), and real-time Magnetic Resonance Imaging (rtMRI) are utilized. Among these, rtMRI stands out for its non-invasive, radiation-free nature and high-resolution visualization of vocal tract shaping, providing a valuable tool to examine dynamic tongue, lip, and palate movements during speech, enhancing our understanding of speech production mechanisms.

In the pursuit of a more in-depth analysis of dynamic movements like the tongue, lips, and palate in rtMRI, identifying articulator contours is essential. However, this task becomes particularly challenging with high-frame-rate data capturing the intricacies of the dynamic vocal tract. Manual labeling is nearly impossible due to its labor-intensive nature and significant time and budget consumption, while automatic methods face obstacles due to low resolution of rtMRI, image blurring and distortion, limited textural variations, and the rapidly changing shapes of the vocal tract.

The complexity of articulatory data analysis is compounded by the limited availability of publicly accessible rtMRI datasets. USC-TIMIT (Narayanan et al., 2014a), a widely used repository in this domain, offers both rtMRI data and ground truth labels in part of data through spatial frequency domain-based segmentation (Bresch and Narayanan, 2009). While this method provides high-quality labels, it's essential to acknowledge that, like many unsupervised approaches, it can still result in instances of incorrect labeling within the dataset. Consequently, numerous research studies have proposed automatic contour detection methods to enhance labeling accuracy (Asadiabadi and Erzin, 2020a,b; Zhang et al., 2016; Toutios and Narayanan, 2015). However, these methods were trained and tested based on this ground truth, which is not purely a gold label. Furthermore, most of these approaches have not made their results available to the public, and a comprehensive analysis of error-prone contour labels is scarce.

In addressing these challenges, this paper offers several contributions:

- The refinement of the landmark-based vocal tract contour data labels[1] is achieved through several steps, including outlier removal, the utilization of a fully convolutional network (FCN)

[1] Note: The refinement of rtMRI landmark-based vocal tract contour labels data proposed in this study available at `https://github.com/ha3ci-lab/USC-TIMIT_rtMRI_Landmarks`. It can be utilized as auxiliary information for the current existing USC-TIMIT dataset.

LREC-COLING 2024, pages 13796–13802
20-25 May, 2024.

to smooth contour shapes, and a point-to-curve projection technique to fit the edge plane of the articulators.

- The investigation of the quality of the new labels through subjective assessments of several contour areas, comparing them to the existing data labels and a comprehensive analysis regarding which contour labels are most prone to issues.

2. Related Works

In rtMRI, vocal tract contour estimation primarily relies on two approaches: landmark-based and segmentation-based. The landmark-based approach involves techniques like the active control model (ACM) (Kass et al., 1988), active shape model (ASM) (Silva and Teixeira, 2015), and articulatory-specific multiple linear regression (MLR) (Labrunie et al., 2018) to identify anatomical landmarks and separate tissue from the airway. Publicly available labeled data is derived from Bresch and Narayanan's work (2009), using unsupervised region segmentation with spatial frequency domain representation and gradient descent optimization.

Conversely, the segmentation-based technique is typically used to assign tissue and background labels through pixel-level segmentation. It then extracts vocal tract contours from the segmented MRI frame. Many methods for vocal tract segmentation (Raeesy et al., 2013; Ruthven et al., 2021) rely on supervised deep learning. Recent work (Silva and Teixeira, 2015) explores multi-task learning for contour detection and labeling.

While many studies have introduced novel techniques, they often rely on existing ground truth, which may not be perfect. In contrast, this paper enhances the ground truth to contribute to the community, by involving outlier detection, contour smoothing with a fully convolutional network, and aligning landmark points with the articulator's edge plane curve.

3. Database Description

This study is based on the USC-TIMIT dataset (Narayanan et al., 2014a,b), a large-scale database of synchronized audio and rtMRI data for speech research. The data were recorded from ten native speakers of American English while uttering the same 460-sentence phonetically balanced dataset used in the MOCHA-TIMIT corpus (Wrench and Richmond, 2000). The visual articulatory data includes midsagittal upper airway MRI data with a 68x68 pixel image resolution over 20x20 cm and a frame rate of 23.18 frames per second.

The USC TIMIT dataset provides ground truth labels for vocal tract contours of just three subjects (F1, F5, and M3). These labels are generated using spatial frequency domain-based segmentation and are available in three formats: 178-points, 180-points, and 181-points, differing in the number of points used. The varying number of landmarks in the original dataset suggests distinct landmark representations, possibly due to different initialization methods. There are 95,223 video frames in total, with frame numbers per format detailed in Table 1. We identified a labeling error involving 3005 frames in the dataset, and this paper offers to refine them. Here, we primarily use the 180-point data, which constitutes over 50% of the available data. Other formats can also be applied if needed.

Table 1: Number of frames for each label format.

Format Points	# Frames (total)	# Frames (each subject)
178	15205	15205 (F1)
180	51252	16882 (F1) 34370 (F5)
181	28766	28766 (M3)

4. Proposed Label Refinement

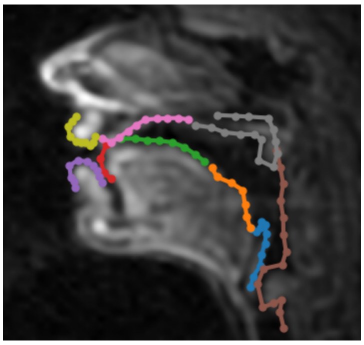


Figure 1: Nine areas of articulator.

There are nine areas to be considered, as illustrated in Figure 1: upper lip (yellow), bottom lip (purple), hard palate (pink), edge tongue (red), middle tongue (green), back tongue (orange), epiglottis (blue), uvular (grey), pharyngeal wall (brown).

To refine the data, various steps were performed to address many outliers and noise-sensitive contour shapes due to the blurry images in the current dataset, which are described in the following sections.

4.1. Outlier Removal

To remove outliers, we first calculate the average size of each area. Then, for each dataset, we compare the size of each area to the average size. Using a constant threshold value, we remove data when the size significantly differs from the average. For instance, we eliminate data where the uvular area is smaller than 400 square pixels for the F1 dataset and smaller than 300 square pixels for the F5 dataset. In total, we detected and removed 248 outliers from the F1 dataset and 2,757 outliers from the F5 dataset. The uvula in subject F5 appears smaller and more blurry in the MRI images than that of subject F1, resulting in a higher frequency of errors (outliers) for F5 compared to F1.

4.2. FCN-based Smoothing

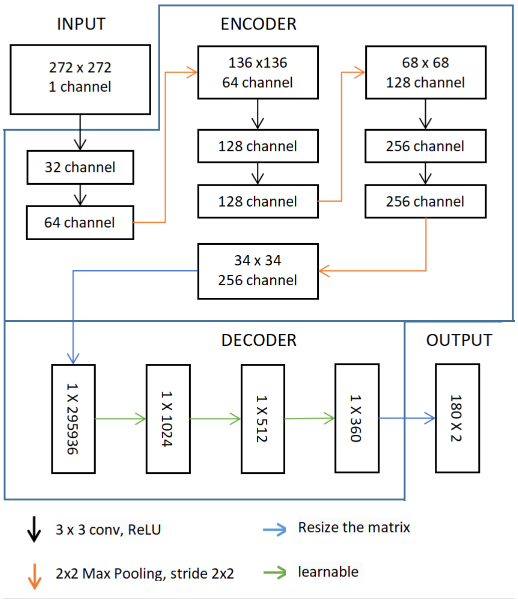


Figure 2: FCN architecture.

Neural networks are effective at handling labeling errors. Generally, simpler models exhibit greater resilience to input noise, as complex models are more prone to overfitting and sensitivity to noisy inputs. Our objective, given the rtMRI image data, is to produce smoother versions of the coordinate points for 180 landmarks. Inspired by the U-Net approach (Ronneberger et al., 2015), we have implemented a more straightforward version of a U-Net-based FCN. The network comprises a four-layer encoding path, followed by a three-layer decoding path, as depicted in Figure 2.

4.3. Point-to-Curve Projection

Last, our goal is to project landmark points onto the edge plane of the articulators. To achieve this, we generate the edge of the MRI image using the adaptive threshold Gaussian method (Gaur et al., 2014) as it is more robust than other classical techniques, Prewitt (1970), Sobel (1968), and Canny (1986), for thresholding images with varying illumination. This method produces all detailed edge information, as illustrated in Figure 3(a). Next, we eliminate unnecessary edges that are too far from the smoothed points generated by FCN, resulting in only the edges corresponding to specific areas of the articulator, as shown in Figure 3(b). This process is repeated for all nine areas of the articulator. Finally, we project the FCN points onto the nearest neighbor edge pixels to fit them onto the edge curve of the articulator, as depicted in Figure 3(c).

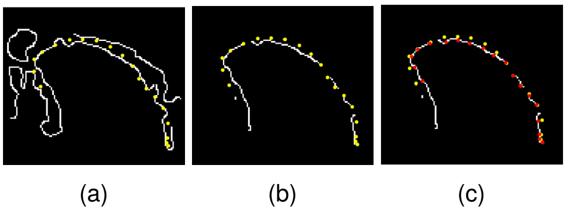

(a) (b) (c)

Figure 3: Yellow dots are the original landmark points; (a) Edges generated by the adaptive threshold Gaussian; (b) Removal of unnecessary edges; and (c) Projection of landmark points (yellow dots) onto the edge curve of the articulator (red dots).

5. Experimental Setup

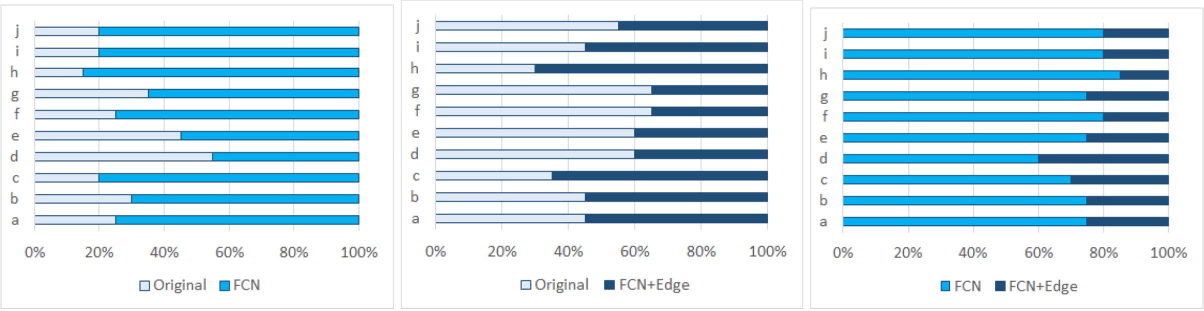
5.1. Model Parameter

The FCN was implemented using PyTorch 2.0.1 (Paszke et al., 2019), and training was performed on a dual NVIDIA GeForce RTX 3090 graphics card. The input is MRI image data, with the original 68x68 pixels. We then upscale it by a factor of 4 into 272x272 pixels using Single-scale SR Network (EDSR) (Lim et al., 2017), while the outputs are coordinate points of 180 landmarks with x and y values. The Adam optimizer (Kingma and Ba, 2017) was applied with hyperparameters $\beta 1 = 0.9$, $\beta 2 = 0.999$, and $\epsilon = 1e{-}4$ to adjust network weights. In each experiment, the network was trained for 100 epochs.

5.2. Subjective Evaluation

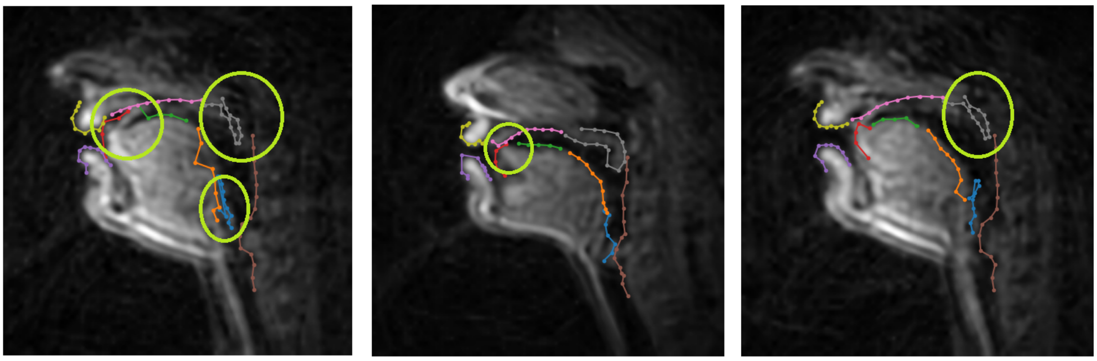
As there is no available gold-level ground truth data for this dataset, we conducted a subjective evaluation involving 20 participants (see Section 9 for ethical considerations). This evaluation entailed an A/B preference test to judge which set of landmark points best fit the MRI image. We compared three different label sets generated from different methods: the original ground truth labels from the USC-TIMIT dataset (denoted as "Original"), the labels produced by FCN smoothing (denoted as "FCN"), and the labels produced by FCN smoothing and point-to-edge projection (denoted as "FCN+Edge").

In total, there are 60 questions that we randomly selected, comprising three sets of 20 questions. Each set compares only two models out of the three available. To avoid bias and complexity, the participant was unaware that there are three systems, and the questions are designed to compare the quality of two randomly selected systems with two images (A and B). Participants are then asked to compare



(a) Original vs. FCN (b) Original vs. FCN+Edge (c) FCN vs. FCN+Edge

Figure 4: AB preference tests on two landmark contour labels among three options (original, FCN, and FCN+Edge), considering nine local areas: (a) upper lip, (b) bottom lip, (c) hard palate, (d) edge of the tongue, (e) middle tongue, (f) back of the tongue, (g) epiglottis, (h) uvular, (i) pharyngeal wall and (j) the overall landmark-based vocal tract.



(a) Original (b) FCN (c) FCN+Edge

Figure 5: Errors and inaccuracies of the landmark contour labels identified in each method.

the accuracy of nine local areas from those two images, as visualized in Figure 1: upper lip, bottom lip, hard palate, edge tongue, middle tongue, back tongue, epiglottis, uvular, pharyngeal wall, as well as the overall landmark-based vocal tract.

6. Experiment Results

Figure 4 shows the results of A/B preference tests on two landmark contour labels among three options (original, FCN, and FCN+Edge), considering nine local areas: (a) upper lip, (b) bottom lip, (c) hard palate, (d) edge of the tongue, (e) middle tongue, (f) back of the tongue, (g) epiglottis, (h) uvular, (i) pharyngeal wall and (j) the overall landmark-based vocal tract.

In a comparison between the labels from the original landmarks of the USC-TIMIT dataset and the smoothed versions from the FCN output (denoted as "Original vs. FCN"), the subjects judged that the FCN labels are better than the original labels in almost all areas. Specifically, for the uvular area, 85% of the subjects chose FCN labels, as well as for the palate, pharyngeal wall, and overall areas, where 80% of the subjects preferred FCN labels. The typical errors and inaccuracies of the landmark contour labels in the uvular area in the original dataset can be seen in Figure 5 (a). The only one area where subjects preferred the original label is the edge of the tongue area, with 55% versus 45%. An error example of an FCN label in this area can be seen in Figure 5 (b).

In a comparison between the labels from the original landmarks of the USC-TIMIT dataset and the smoothed versions from the FCN with the point-to-edge projection approach (denoted as "Original vs. FCN+Edge"), unexpectedly, the FCN+Edge labels only judged to be better in some areas, resulting in an almost 50%-50% distribution overall. This might be because both the original landmarks and the FCN+Edge labels rely on edge prediction, which might still not be robust in handling blurry areas in the images, often resulting in unnecessary landmark contour labels. The main areas where FCN+Edge is unable to provide better labels are in the uvular area. An example of an error from FCN+Edge can be seen in Figure 5 (c).

Lastly, in the comparison between the labels from the FCN alone and the FCN with the point-to-edge projection approach (denoted as "FCN vs. FCN+Edge"), the subjects once again judged that the FCN-only labels are better than the FCN+Edge labels in all areas. As before, the most significant difference is observed in the uvular area, where 85% of the subjects preferred the FCN-only labels. Overall, the results reveal that FCN-only labels significantly outperform the original and FCN+Edge label data. The FCN successfully demonstrates greater resilience to input noise and generates smoother labels. Consequently, the data we release is based on these FCN labels.

7. Conclusion

This paper contributes to the field by offering a refinement of landmark-based vocal-tract contour labels. This refinement includes outlier removal, FCN-based smoothing, and a landmark point-to-edge curve projection approach. Despite the absence of an established ground truth label, we evaluated the quality of the new labels through subjective assessments of various contour areas, comparing them to the existing data labels. The results reveal that FCN-only labels significantly outperform the original and FCN+Edge label data. The most critical area prone to errors in the original data is the uvular area. Nevertheless, the results demonstrate that the new labels, with the outlier removal and FCN-based smoothing, significantly enhance accuracy and reliability, providing improved vocal-tract label data for the research community.

8. Limitations

Despite providing new ground truth labels with a better accuracy and reliability, this study has some limitations. First, the proposed approach utilizes several well-known techniques, which may limit its technical novelty. However, as mentioned earlier, the primary aim of this study is to contribute to improved ground truth labels. The results have indeed shown that even simple yet effective techniques, such as outlier removal and FCN-smoothing, can provide better labels. While the edge-to-curve projection approach was not particularly effective in improving the labels in this case due to its reliance on edge prediction, which might still not be robust in handling blurry areas in the images, we will address this issue by exploring other methods that are more robust in such situations. Furthermore, in the future, we will investigate more sophisticated and novel approaches to further refine the landmark contour labels.

Second, this study lacks objective evaluation primarily due to the absence of a true gold standard for ground truth labels. During the planning of the subjective evaluation, we intended to involve a minimum of 30 participants. However, after the final data collection and compilation, only 20 participants were available for the study. This may be due to participants' unfamiliarity with articulatory data. In future studies, we plan to collect data from a larger and more diverse group.

Lastly, this study focused on the specific USC-TIMIT rtMRI dataset. However, since the proposed method used the FCN framework to generate landmarks and neural networks function like a black box when mapping a vector to another vector, this method can be applied to other datasets with minimal modification.

9. Ethical Considerations

This study is based on the publicly available USC-TIMIT rtMRI vocal tract dataset. The participation of human subjects in evaluating the new ground truth labels, especially during the subjective assessment, has undergone review and approval by our institution's Institutional Review Board and Research Ethics Committee. Consequently, the experiments were conducted in accordance with institutional ethical guidelines.

For crowdsourced adult participants, we implemented a 'first in, first out' (FIFO) selection method, ensuring that they met specific criteria. These criteria included a requirement for English proficiency above TOEIC 700 to ensure that participants could comprehend the task guidelines. Additionally, we needed to screen for color blindness, as figures and landmarks were identified using various colors. Other than that, there was no discrimination in the selection of participants.

We have taken steps to ensure that all participants in this study are well-informed about the research's objectives, data usage, and data protection related to personally identifiable information. Before participating in the experiments, individuals provided written informed consent, and, after the experiments, they received compensation as per our institution's hourly work policy.

10. Acknowledgements

Part of this work is supported by JSPS KAKENHI Grant Numbers JP21H05054 and JP23K21681.

11. Bibliographical References

Sasan Asadiabadi and Engin Erzin. 2020a. Automatic vocal tractlandmark tracking in rtmri using fully convolutional networks and kalman filter. In *Proc. the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7339–7343.

Sasan Asadiabadi and Engin Erzin. 2020b. Vocal tract contour tracking in rtmri using deep temporal regression network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:3053–3064.

E. Bresch and S. Narayanan. 2009. Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images. *IEEE Transactions on Medical Imaging*, 28(3):323–338.

John Canny. 1986. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698.

Sanjay Gaur, Jayashri Vajpai, and Sandip Mehta. 2014. Adaptive local thresholding for edge detection. *International Journal of Computer Applications*, 2:8887.

Michael Kass, Andrew Witkin, and Demetri Terzopoulos. 1988. Snakes: Active contour models. *Proc. the International Journal of Computer Vision*, 1(4):321–331.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Mathieu Labrunie, Pierre Badin, Dirk Voit, Arun A Joseph, Jens Frahm, Laurent Lamalle, Coriandre Vilain, and Louis-Jean Boë. 2018. Automatic segmentation of speech articulators from real-time midsagittal mri based on supervised learning. *Speech Communication*, 99:27–46.

Peter Ladefoged and Ian Maddieson. 1996. *The Sounds of the World's Languages*. Wiley.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. 2017. Enhanced deep residual networks for single image super-resolution. *arXiv preprint arXiv:1707.02921*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*.

Joseph S. Perkell, Marc H. Cohen, Mario A. Svirsky, Melanie L. Matthies, Iñaki Garabieta, and Michel T. T. Jackson. 1992. Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *The Journal of the Acoustical Society of America*, 92(6):3078–3096.

Judith MS Prewitt et al. 1970. Object enhancement and extraction. *Picture processing and Psychopictorics*, 10(1):15–19.

Zeynab Raeesy, Sylvia Rueda, Jayaram K. Udupa, and John Coleman. 2013. Automatic segmentation of vocal tract mr images. In *Proc. the 10th IEEE International Symposium on Biomedical Imaging*, pages 1328–1331.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 9351:234–241.

Matthieu Ruthven, Marc E. Miquel, and Andrew P. King. 2021. Deep-learning-based segmentation of the vocal tract and articulators in real-time magnetic resonance images of speech. *Computer Methods and Programs in Biomedicine*, 198:105814.

Rafael De Assuncao Sampaio and Marcel Parolin Jackowski. 2017. Vocal tract morphology using real-time magnetic resonance imaging. In *Proc. the 30th Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 359–366.

Samuel Silva and António Teixeira. 2015. Unsupervised segmentation of the vocal tract from real-time mri sequences. *Computer Speech & Language*, 33(1):25–46.

Irwin Sobel, Gary Feldman, et al. 1968. A 3x3 isotropic gradient operator for image processing. *a talk at the Stanford Artificial Project in*, 1968:271–272.

Asterios Toutios and Shrikanth S. Narayanan. 2015. Factor analysis of vocal-tract outlines derived from real-time magnetic resonance imaging data. In *International Congress of Phonetic Sciences*.

Houri K Vorperian, Shubing Wang, Moo K Chung, E Michael Schimek, Reid B Durtschi, Ray D Kent, Andrew J Ziegert, and Lindell R Gentry. 2009. Anatomic development of the oral and pharyngeal portions of the vocal tract: An imaging study. *The Journal of the Acoustical Society of America*, 125(3):1666–1678.

Dawei Zhang, Minghao Yang, Jianhua Tao, Yang Wang, Bin Liu, and Danish Bukhari. 2016. Extraction of tongue contour in real-time magnetic resonance imaging sequences. In *Proc. The IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 937–941.

12. Language Resource References

Cai, Zexin and Qin, Xiaoyi and Cai, Danwei and Li, Ming and Liu, Xinzhong and Zhong, Haibin. 2018. *The DKU-JNU-EMA Electromagnetic Articulography Database on Mandarin and Chinese Dialects with Tandem Feature based Acoustic-to-Articulatory Inversion*.

Narayanan, Shrikanth and Toutios, Asterios and Ramanarayanan, Vikram and Lammert, Adam and Kim, Jangwon and Lee, Sungbok and Nayak, Krishna and Kim, Yoon-Chul and Zhu, Yinghua and Goldstein, Louis and Byrd, Dani and Bresch, Erik and Ghosh, Prasanta and Katsamanis, Athanasios and Proctor, Michael. 2014a. *Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)*.

Narayanan, Shrikanth and Toutios, Asterios and Ramanarayanan, Vikram and Lammert, Adam and Kim, Jangwon and Lee, Sungbok and Nayak, Krishna and Kim, Yoon-Chul and Zhu, Yinghua and Goldstein, Louis and Byrd, Dani and Bresch, Erik and Ghosh, Prasanta and Katsamanis, Athanasios and Proctor, Michael. 2014b. *USC-TIMIT: A database of multimodal speech production data*.

John Westbury, Paul Milenkovic, Gary Weismer, and Raymond Kent. 2005. X-ray microbeam speech production database. *The Journal of the Acoustical Society of America*, 88(S1):S56–S56.

Wrench, Alan A. and Richmond, Korin. 2000. *Continuous speech recognition using articulatory data*. ISCA.