Improve vocal tract reconstruction and modeling using an image super-resolution technique

Xinhui Zhou [a)]

Speech Communication Laboratory, Institute of Systems Research and Department of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742

Jonghye Woo [b)]

Departments of Neural and Pain Sciences, University of Maryland Dental School, Baltimore, MD 21201
Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD 21218

Maureen Stone [c)]

Department of Neural and Pain Sciences and Department of Orthodontics, University of Maryland Dental School, Baltimore, MD 21201

Jerry L. Prince [d)]

Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore MD 21218

Carol Y. Espy-Wilson [e)]

Speech Communication Laboratory, Institute of Systems Research and Department of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742

(Dated: January 15th, 2013)

Running title: **super-resolution technique based vocal tract modeling**

a) Electronic address: zxinhui@umd.edu
b) Electronic address: jwoo@umaryland.edu
c) Electronic address: mstone@umaryland.edu
d) Electronic address: prince@jhu.edu
e) Electronic address: espy@umd.edu

Abstract:

Magnetic resonance imaging has been widely used in speech production research. Often only one image stack (sagittal, axial or coronal) is used for vocal tract modeling. As a result, complementary information from other available stacks is not utilized. To overcome this, we applied a recently developed super-resolution technique to integrate three orthogonal low-resolution stacks into one isotropic volume. Our results on vowels show that the super-resolution volume produces better vocal tract visualization and reconstruction than any of the low-resolution stacks. Its derived area functions generally produce formant predictions closer to the ground truth, particularly for those formants sensitive to area perturbations at constrictions.

©2013 Acoustical Society of America

PACS numbers: 43.70.Dn, 43.70.Bk

I. Introduction

Magnetic resonance imaging (MRI) has been widely used in speech production research for vocal tract reconstruction and modeling (Baer *et al.*, 1991; Dang *et al.*, 1994; Narayanan *et al.*, 1995; Story *et al.*, 1996; Alwan *et al.*, 1997; Narayanan *et al.*, 1997; Takemoto *et al.*, 2006a; Takemoto *et al.*, 2006b; Zhou *et al.*, 2008). Based on the obtained MR images, the vocal tract shapes were reconstructed, and then area functions or 3-D vocal tract models were built. Although multi-plane scanning may have been performed, only one MR image stack (sagittal, coronal, or axial) is usually used for the vocal tract reconstruction. For example, sagittal stacks were used by Narayanan *et al.* (1997); Takemoto *et al.* (2006b), coronal slices used by Dang *et al.* (1994), and axial slices used by Story *et al.* (1996). However, due to constraints (time, money and subject endurance) in MRI scanning, each image stack typically has an in-plane resolution which is much better than the out-of-plane resolution. So the mid-sagittal tongue shape is better visualized in the sagittal stack; a pharyngeal constriction is better visualized in the axial stack; and an oral cavity constriction may be more accurately estimated from the coronal stack. As a result, using only one image stack to reconstruct the vocal tract shape is not optimal when extra information is available from other image stacks.

It has been shown that combining orthogonal image stacks help the measurement of vocal tract shape. Narayanan *et al.* (1995) extracted the area functions in the front and back regions from the coronal stacks and the axial stacks, respectively. Baer *et al.* (1991) and Zhou *et al.* (2008) overlaid coronal and axial stacks for a better segmentation of the vocal tract. However, these manual combination approaches are empirical and not systematically evaluated. To the best of our knowledge, there is no prior work on automatically integrating orthogonal stacks into one unified volume for vocal tract modeling.

The goal of this study was to apply a recently developed super-resolution image reconstruction technique from our group (Woo *et al.*, 2012) to automatically integrate multiple orthogonal stacks together for improving vocal tract reconstruction and modeling. This technique can combine orthogonal low-resolution stacks into one isotropic super-resolution volume which has already been demonstrated to improve the visualization and the peak signal-to-noise ratio of the tongue volume (Woo *et al.*, 2012).

In order to assess this technique in vocal tact modeling, we obtained a publicly available MR database (ATR. 2008) and, due to its high image resolution, treated its MR stacks as the ground truth of vocal tracts. Then, from each original stack, we simulated three orthogonal stacks (sagittal, coronal and axial) with a commonly used low-resolution value used in cine-MRI and applied the super-resolution technique on them to generate a corresponding super-resolution volume. Assuming the original MR data is the ground truth, we compared the vocal tract area functions derived from high- (original), low- (simulated), and super-resolutions (multiplanar reconstructions) respectively.

In the rest of this paper, we first describe the MRI database and our methodologies in Section II. Then in Section III, we compare the vocal tract visualizations and the area function vocal tract models between the low-resolution stacks and the super-resolution volumes. Finally, a summary is presented in Section V. where is section IV?

II. Database and methodologies

A. MR Database

In this study, we chose to use the ATR MRI database of Japanese vowel production (ATR, 2008). It contains MR sagittal stacks for five sustained vowel productions (/a/, /e/, /i/, /o/,

and /u/). The mid-sagittal MR images of these five sounds are shown in Fig. 1. The reasons for using this database are twofold. First, the MR stacks have a high image resolution (an in-plane resolution of 0.5 mm per pixel and a slice thickness of 2 mm), so they can be used as ground truth for each vowel vocal tracts. Second, the MR image data are supplemented with teeth, thus making teeth compensation easy and extraction of the vocal tract shape precise.

There is an interleaving between 'bright' slices and 'dark' slices in the original MR stacks. This was caused by the gradually weaker slice intensity in the imaging order. Before using this database, we fixed this issue by scaling the intensity of each slice so that the intensity profiles across slices were smoothed.

Note that the ATR MRI database of Japanese vowel production were acquired at and released from ATR Human Information Science Laboratories under the "Research on Human Communication" project funded by the National Institute of Information and Communications Technology. The use of the database and release of the results are under the license agreement with ATR-Promotions Inc.

B. Simulated orthogonal low-resolution MR stacks

By down-sampling the original MR data in the ATR MRI database, we simulated three orthogonal low-resolution stacks (sagittal, coronal, and axial) for each vowel. The simulated stacks have an in-plane resolution of 2 mm per pixel and a slice thickness of 6 mm, which mimics the resolution in our cine-MR data. Prior to down-sampling the original volumes in slice-selection directions, Gaussian filtering with σ=0.5 (in-plane) and σ=2 (slice-selection direction) was applied to avoid an anti-aliasing effect.

C. Super-resolution reconstruction technique

As each simulated low-resolution stack has an in-plane resolution which is much better than the out-of-plane resolution (2 mm vs. 6 mm), the information contained in the three low-resolution stacks is complementary. Intuitively, integrating the three low-resolution stacks into one would be better for modeling than using only one of the low-resolution stacks. The super-resolution technique recently developed (Woo *et al.*, 2012) and applied here integrates three orthogonal low-resolution stacks and generates one isotropic super-resolution volume. It is a maximum a posteriori-Markov random field (MAP-MRF) based reconstruction method. It incorporates edge-preserving regularization to improve signal-to-noise ratio and resolution and yields superior image qualities compared with other reconstruction methods as visually and quantitatively assessed. In addition, image registration is performed in this technique to correct any possible head motions between acquisitions of different image stacks. The resulting voxel resolution in the super-resolution volume here is 2x2x2 mm, three times of that in the low resolution volume.

D. Vocal tract segmentation and area function models

We extracted and compared the area function vocal tract models derived from the low-resolution volumes and the super-resolution volume respectively. There are two steps for getting the vocal tract area functions: airway segmentation and grid line determination. We performed the airway segmentation using thresh-holding at gray values that are approximately halfway from the air to the tissue near the boundary. Manual correction was also performed at our best guess in regions with over-segmentation or under-segmentation. As illustrated in Fig. 2, we used a centerline method based program (Kitamura and Narishige, 2011) to determine the grid lines for the area function extraction. For simplicity, the piriform sinuses and interdental

spaces were excluded in the vocal tract models. For comparison purposes, the same set of grid lines were used for all the image volumes of the same sound. With the obtained area functions, we calculated their corresponding acoustic responses using our MATLAB-based software VTAR (Espy-Wilson *et al.*, 2007).

III. Results

A. Vocal tract visualization in super-resolution volume

Using the vowel /a/ as an example, Fig. 3 shows the 3-D views of the three low-resolution stacks and the super-resolution volume, respectively. Due to the large slice thickness (6 mm), detailed vocal tract structures in the low-resolution stacks might be blurred or distorted along the scanning direction. For example, as shown in Fig. 3, the epiglottis region is blurred in the sagittal stack (Fig. 3a), the lip opening smaller in the axial stack (Fig. 3b), and the pharyngeal region narrower in the coronal stack (Fig. 3c). But these above mentioned issues in the low-resolution stacks are improved in the super-resolution volume (Fig. 3d) because its volume resolution is three times the low resolution. So, as expected, the super-resolution volume provides better vocal tract visualization than any of the three low-resolution stacks. This improvement of vocal tract visualization in the super-resolution volume may help us better understand the roles of detailed structures in speech production.

B. Vocal tract area functions and the corresponding acoustic responses

Figs. 4-6 show the extracted area functions and the corresponding acoustic responses for the five vowels. For each vowel, area functions were extracted from MR data in high resolution, low resolution, and super-resolution respectively. The results from high resolution data are

regarded as the vocal tract ground truth. For the vowel /a/ (Fig. 4a), area functions from three orthogonal low-resolution stacks were extracted respectively. The coronal low-resolution stack produces a relatively large area in the front but a relatively small area in the back compared to the other two low-resolution stacks. The axial stack is the opposite. The sagittal stack and the super-resolution volume produce more uniform areas. For other vowels, we did not extract area functions from low resolution coronal or axial stacks, since constrictions for those vowels might not be visible in those two stacks. For example, the constriction in the front for /i/ is not visible in the coronal and axial stacks.

Although the area functions from low-resolution stacks and super-resolution volumes have small differences, the super-resolution data in the vowels /a/ and /e/ (Fig. 4) produce formant patterns closer to the ground truth than the low-resolution data, specifically for the higher formants F3 and F4. These F3 and F4 differences are due to the area differences in the laryngeal cavity. When the area functions in the pharyngeal region are replaced with the ground truth (indicated by the yellow line in Fig. 4a or 4b), the acoustic responses are corrected to be almost identical to the ground truth. So the super-resolution data produces better area estimation in the laryngeal cavity, which leads to a better prediction of F3 and F4.

For the vowel /o/ (Fig. 5), there is not much difference in formants between the outputs from the super-resolution data and the low-resolution data. Replacing the areas of the pharynx with the ground truth (indicated by the yellow line) corrects F4 and replacing the areas in the lip region with the ground truth corrects F2. For the vowels /i/ and /u/ (Fig. 6), the difference in formants between the outputs from the super-resolution data and the low-resolution data is also small. While replacing the laryngeal constriction with the ground truth corrects the formant errors, the errors are much smaller in these two high vowels because the pharynx is relatively

larger and the area error is a smaller percentage. As indicated in Fig. 6, replacing the tongue constrictions with ground truth will correct F1 for /i/ and F2 for /u/.

IV. Discussion

Our results indicate that, in order to have formant predictions close to the ground truth, it is crucial for the vocal tract model to get accurate areas at the constrictions. This is because even a small area change in constriction might lead to a large percentage change in area and, based on the formant sensitivity function, it will affect those formants sensitive to that area perturbation. The super-resolution volume does help measure the areas in the laryngeal cavity for low vowels (/a/ and /e/,. but there might be some other alternatives to help improve the estimation of small areas. First, instead of using the thresholding segmentation method, a more advanced algorithm such as those based on deformable models might help. Second, we can apply an MRI scan with a relatively high resolution exclusively to those constricted regions.

V. Summary

Through integrating information from multiple orthogonal low-resolution MR stacks, a super-resolution volume can provide better vocal tract visualization than any one of the low-resolution MR stacks. Overall, area function vocal tract models derived from the super-resolution volumes produce better prediction of formants, particularly when those formants are sensitive to area perturbations at constrictions (such as the laryngeal cavity). However, phonetic qualities (F1 and F2) of vowels in area functions are not affected by the low image resolution. Our results also suggest that applying more advanced segmentation methods to get accurate areas in constrictions may improve vocal tract modeling accuracy.

Acknowledgements

This work was supported by a grant from the National Institutes of Health (grant No. R01CA133015). The authors would like to thank Dr. Emi Murano at Johns Hopkins University and Dr. Tatsuya Kitamura at Konan University for their help in this work.

References

Alwan, A., Narayanan, S., and Haker, K. (**1997**). "Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part II. The rhotics," Journal of the Acoustical Society of America **101**, 1078-1089.

ATR (**2008**). "The ATR MRI database of Japanese vowel production " (ATR Human Information Science Laboratories).

Baer, T., Gore, J. C., Gracco, L. C., and Nye, P. W. (**1991**). "Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels," Journal of the Acoustical Society of America **90**, 799-828.

Dang, J. W., Honda, K., and Suzuki, H. (**1994**). "Morphological and Acoustical Analysis of the Nasal and the Paranasal Cavities," Journal of the Acoustical Society of America **96**, 2088-2100.

Espy-Wilson, C. Y., Zhou, X., Tiede, M., and Boyce, S. (**2007**). "New features in VTAR: A Matlab-based computer program for vocal tract acoustic modeling," Journal of the Acoustical Society of America **121**, 3136.

Kitamura, T., and Narishige, T. (**2011**). "Development of plug-in for ImageJ to extract vocal tract area function," in *Proceedings of International Seminar of Speech Production*.

Narayanan, S. S., Alwan, A. A., and Haker, K. (**1995**). "An Articulatory Study of Fricative Consonants Using Magnetic-Resonance-Imaging," Journal of the Acoustical Society of America **98**, 1325-1347.

Narayanan, S. S., Alwan, A. A., and Haker, K. (**1997**). "Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part I. The laterals," Journal of the Acoustical Society of America **101**, 1064-1077.

Story, B. H., Titze, I. R., and Hoffman, E. A. (**1996**). "Vocal tract area functions from magnetic resonance imaging," Journal of the Acoustical Society of America **100**, 537-554.

Takemoto, H., Adachi, S., Kitamura, T., Mokhtari, P., and Honda, K. (**2006a**). "Acoustic roles of the laryngeal cavity in vocal tract resonance," Journal of the Acoustical Society of America **120**, 2228-2238.

Takemoto, H., Honda, K., Masaki, S., Shimada, Y., and Fujimoto, I. (**2006b**). "Measurement of temporal changes in vocal tract area function from 3D cine-MRI data," Journal of the Acoustical Society of America **119**, 1037-1049.

Woo, J., Murano, E. Z., Stone, M., and Prince , J. L. (**2012**). "Reconstruction of high resolution tongue volumes from MRI," IEEE Trans on Biomedical Engineering **59**, 3511-3524.

Zhou, X., Espy-Wilson, C. Y., Tiede, M., Boyce, S., Holland, C., and Choe, A. (**2008**). "An MRI-based articulatory and acoustic study of 'retroflex' and 'bunched' American English /r/ sounds," Journal of the Acoustical Society of America **123**, 4466-4481.

Fig. 1. Original midsagittal images of the five vowels in the ATR MRI database.

Fig. 2. (Color online) From vocal tract segmentation to area function. Left: grid lines for extracting area function; Right: the extracted area function.

Fig. 3. (Color online) 3-D views of image stacks in low resolution (2x2x6 mm) or super resolution (2x2x2 mm) (from left to right: axial, sagittal, and coronal views). (a) simulated sagittal stack in low resolution, (b) simulated axial stack in low resolution, (c) simulated coronal stack in low resolution, (d) super-resolution volume.

Fig. 4. (Color online) Area functions of **/a/** and **/e/** and their corresponding acoustic responses. (a) **/a/** and (b) **/e/.** Right side: area functions are extracted from MR data in original high resolution, low resolution, and super-resolution, and left side: area functions are the same to the left side, except that the pharynx regions are made the same as the ground truth, indicated by the thick yellow line.

Fig. 5. (Color online) Area functions of **/o/** and their corresponding acoustic responses. (a) area functions are extracted from MR data in original high resolution, low resolution, and super-resolution, (b) area functions are the same as in (a), except that the pharynx and lips regions are made the same as the ground truth, indicated by the thick yellow line (the dashed lines connect the region in the area function with the affected formant in the acoustic response).

Fig. 6. (Color online)) Area functions of **/i/**, **/u/** and their corresponding acoustic responses. (a) /i/, (b) /u/ (the dashed lines connect the region in the area function with the affected formant in the acoustic response.)

List of Figures

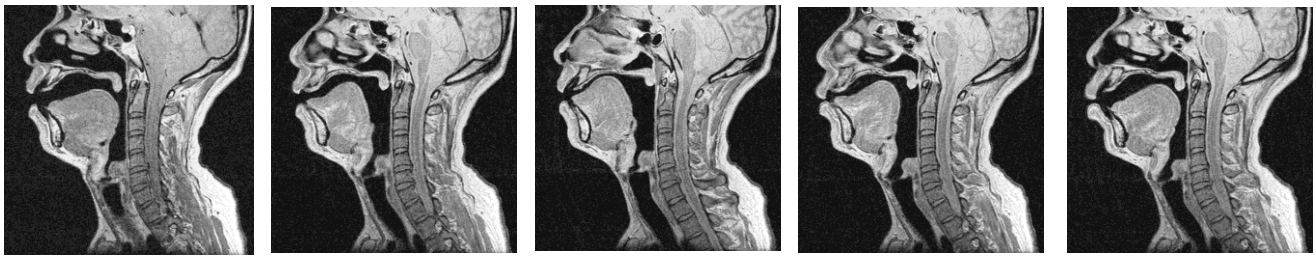


Figure 1

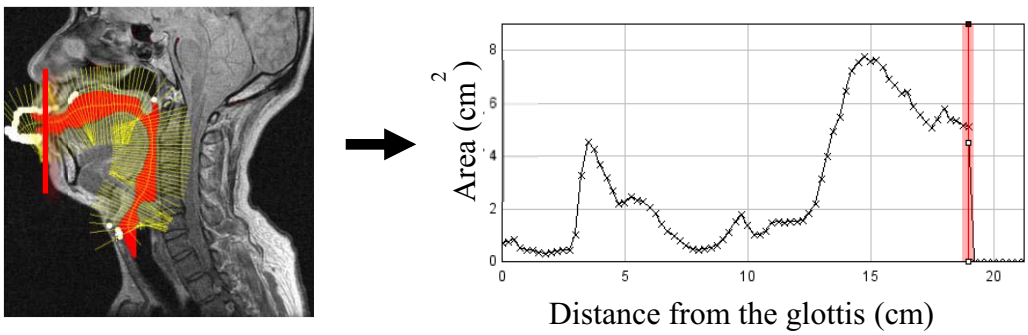


Figure 2

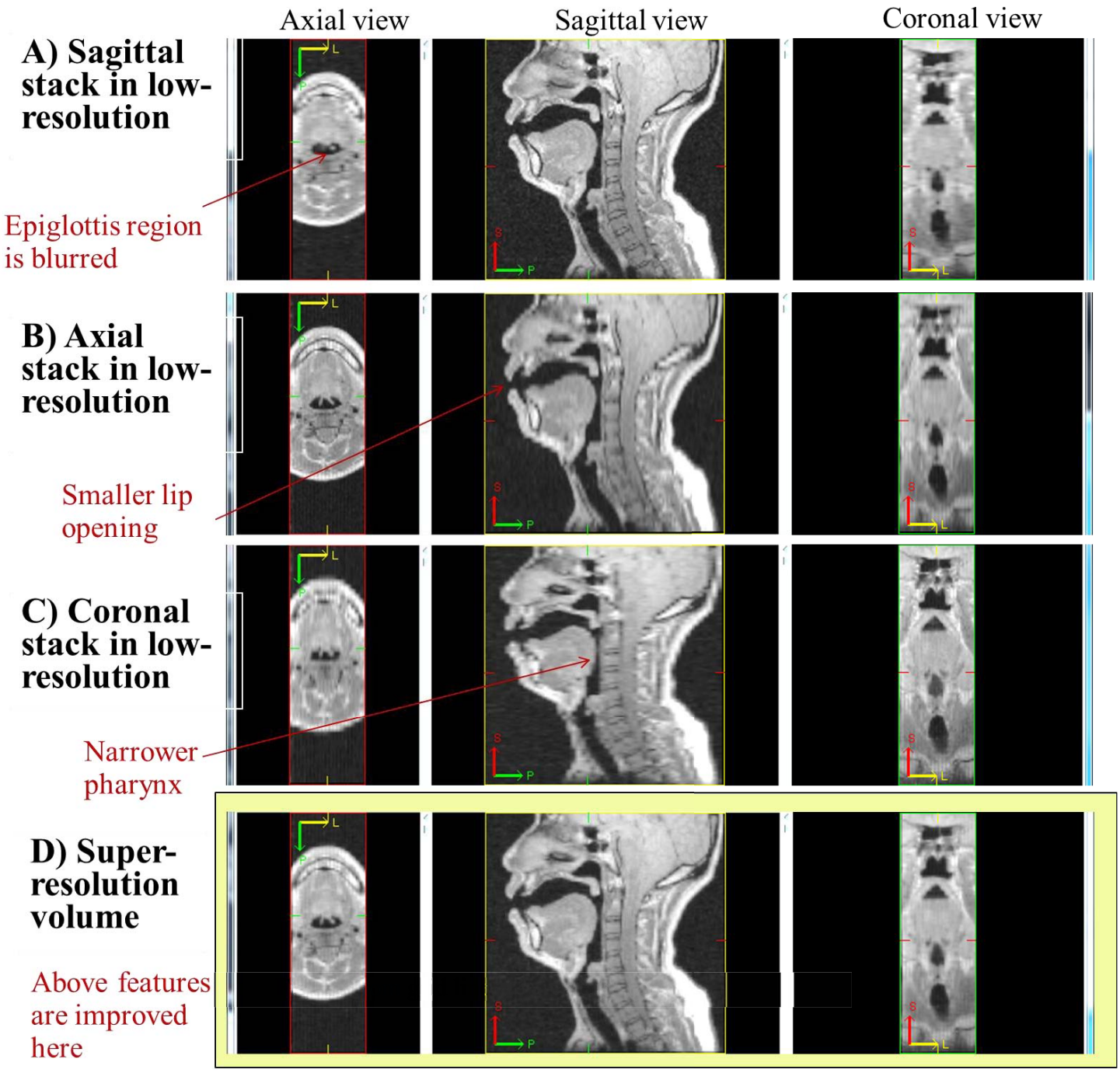
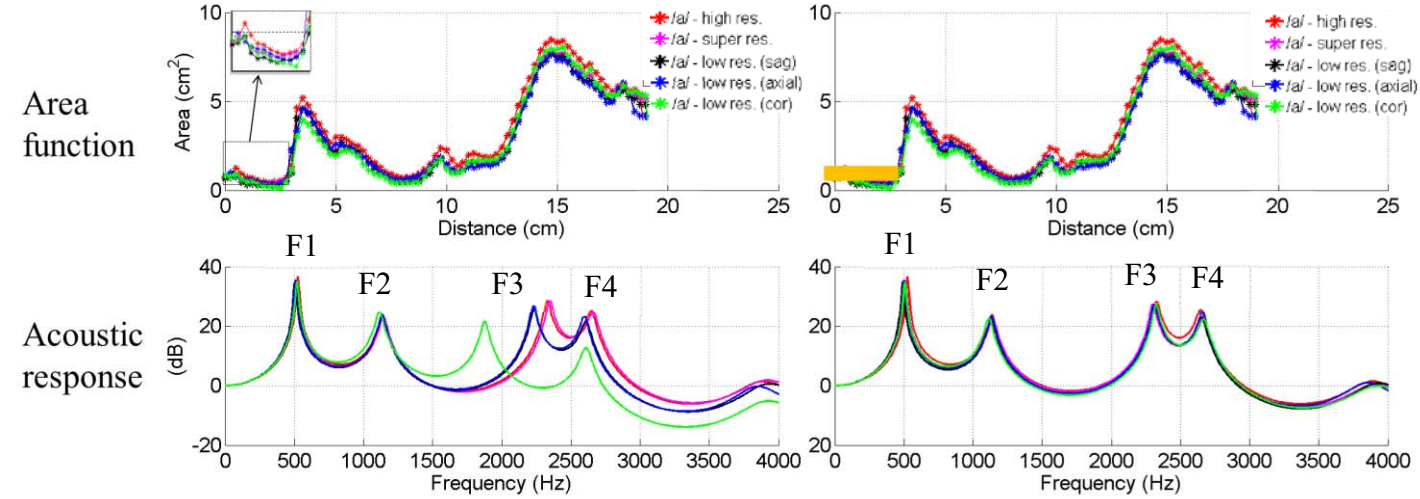
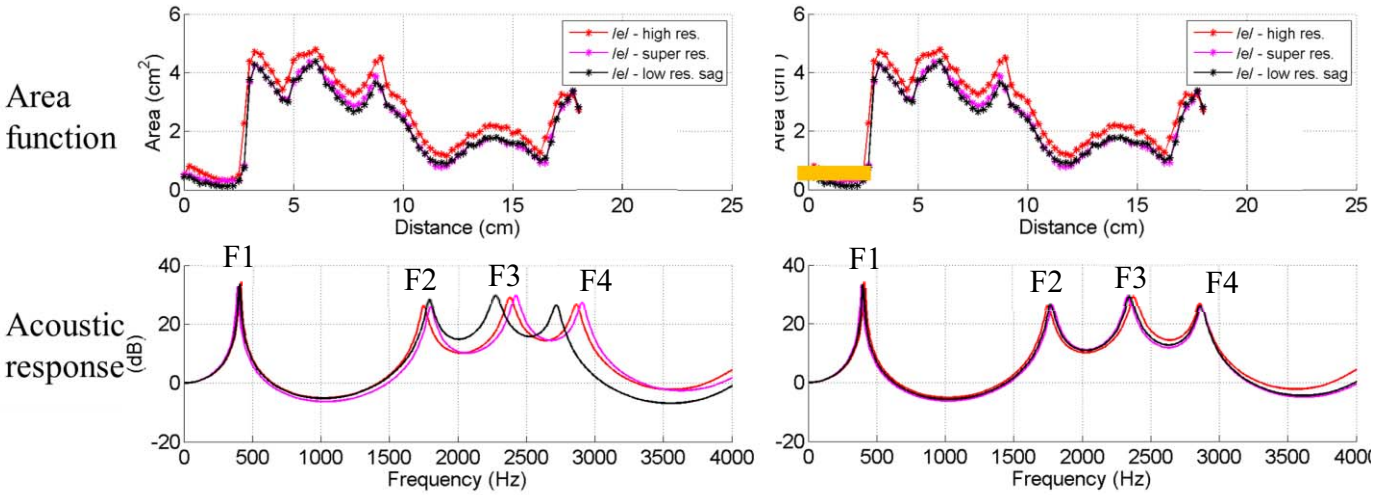

Axial view
Sagittal view
Coronal view
A) Sagittal stack in low-resolution
Epiglottis region is blurred
B) Axial stack in low-resolution
Smaller lip opening
C) Coronal stack in low-resolution
Narrower pharynx
D) Super-resolution volume
Above features are improved here

Figure 3

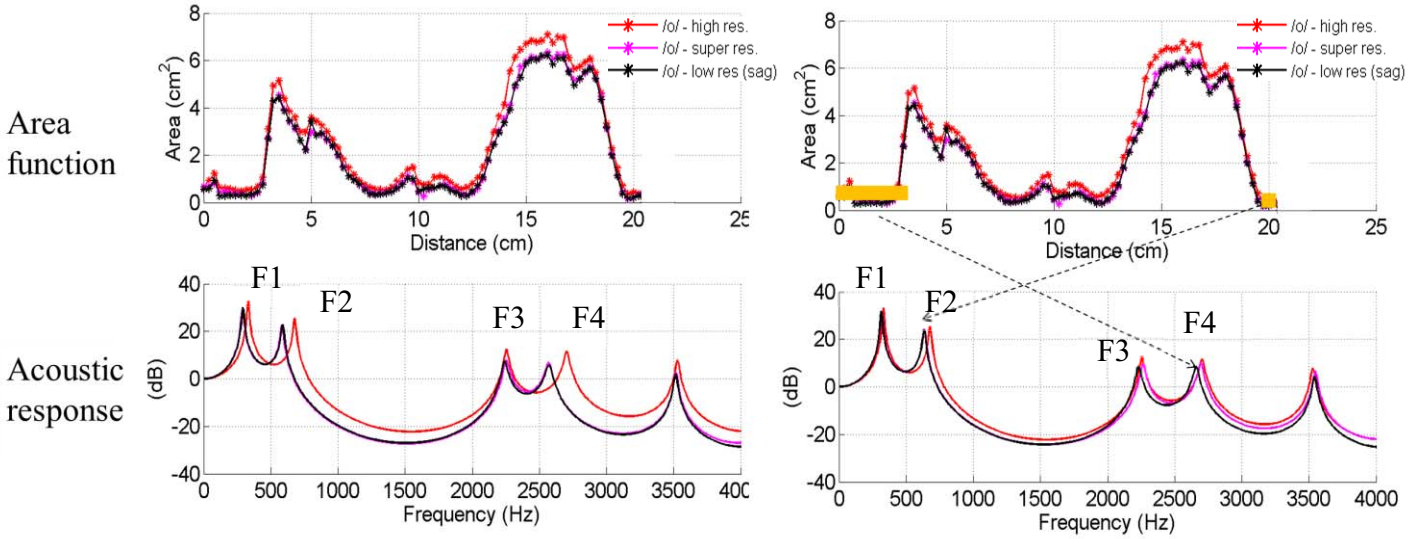

Area function
Acoustic response
/a/ - high res.
/a/ - super res.
/a/ - low res. (sag)
/a/ - low res. (axial)
/a/ - low res. (cor)
Area (cm2)
Distance (cm)
F1
F2
F3
F4
(dB)
Frequency (Hz)

(a)


Area function
Acoustic response
/e/ - high res.
/e/ - super res.
/e/ - low res. sag
Area (cm2)
Distance (cm)
F1
F2
F3
F4
(dB)
Frequency (Hz)

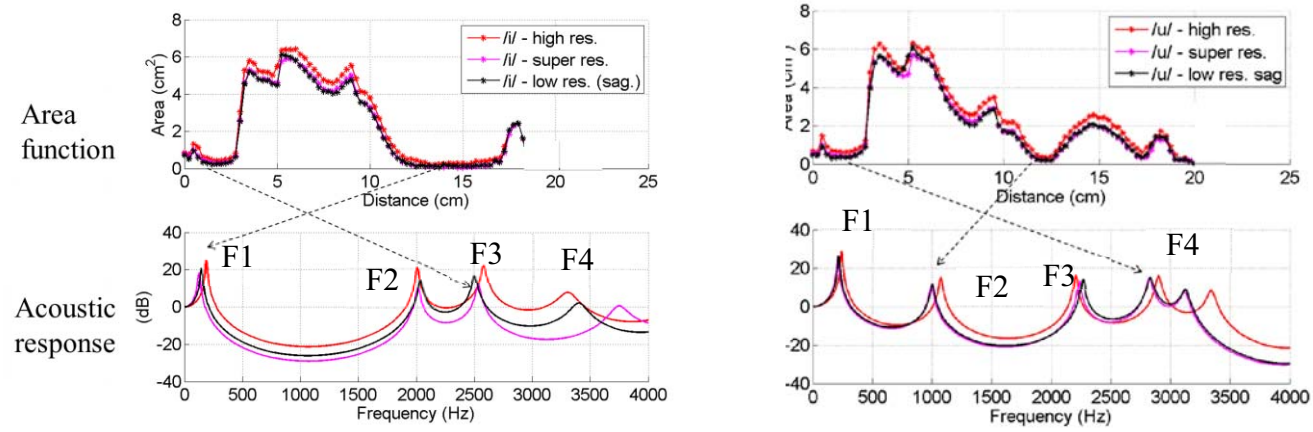
(b)

Figure 4



(a) (b)

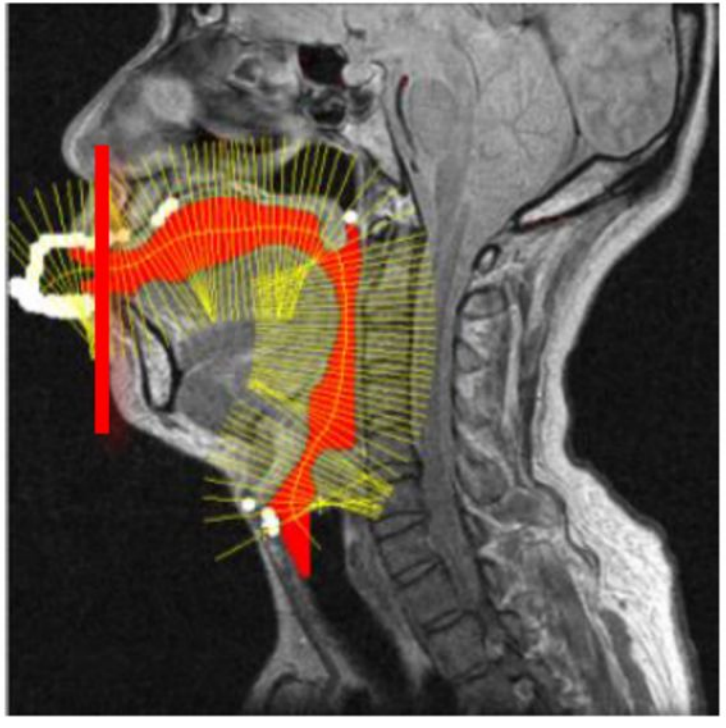
Figure 5

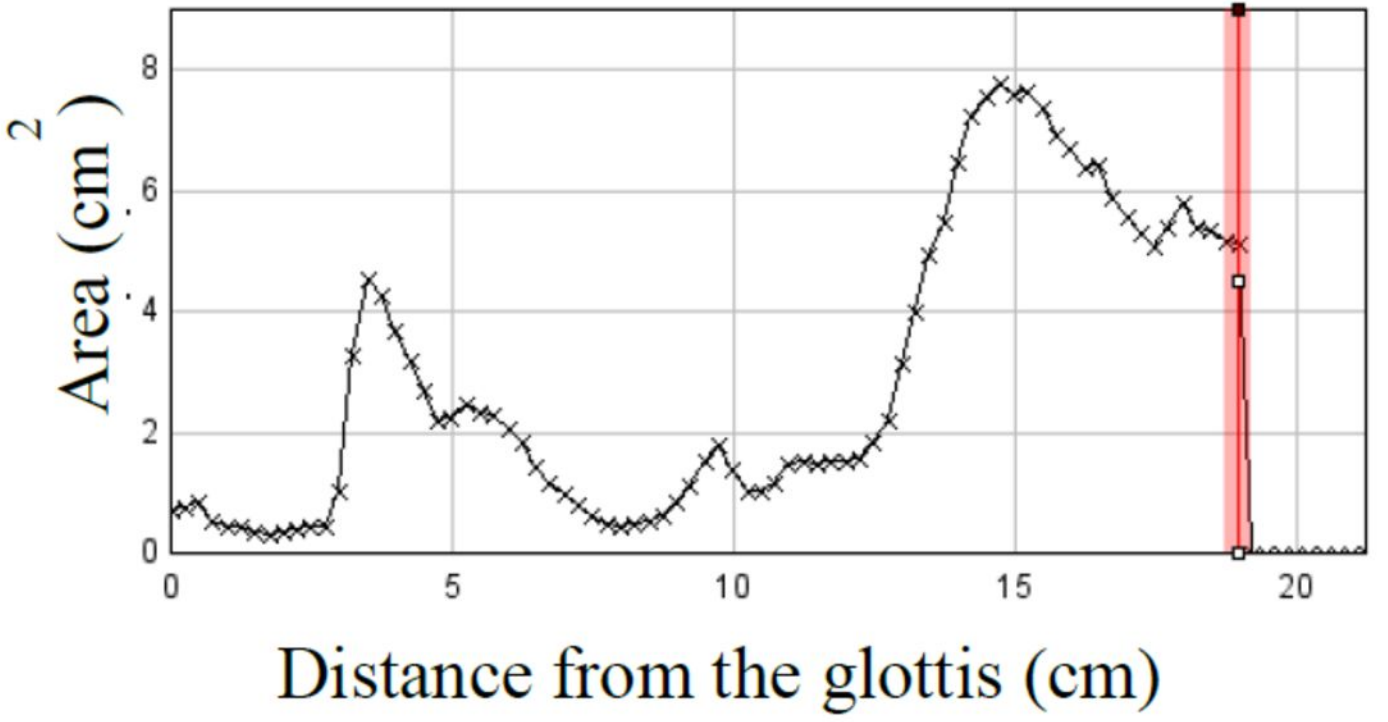


(a) (b)

Figure 6


/a/
/e/
/i/
/o/
/u/




Area (cm 2)
0
2
4
6
8
0
5
10
15
20
Distance from the glottis (cm)

Axial view

Sagittal view

Coronal view

A) Sagittal stack in low-resolution

Epiglottis region is blurred

B) Axial stack in low-resolution

Smaller lip opening

C) Coronal stack in low-resolution

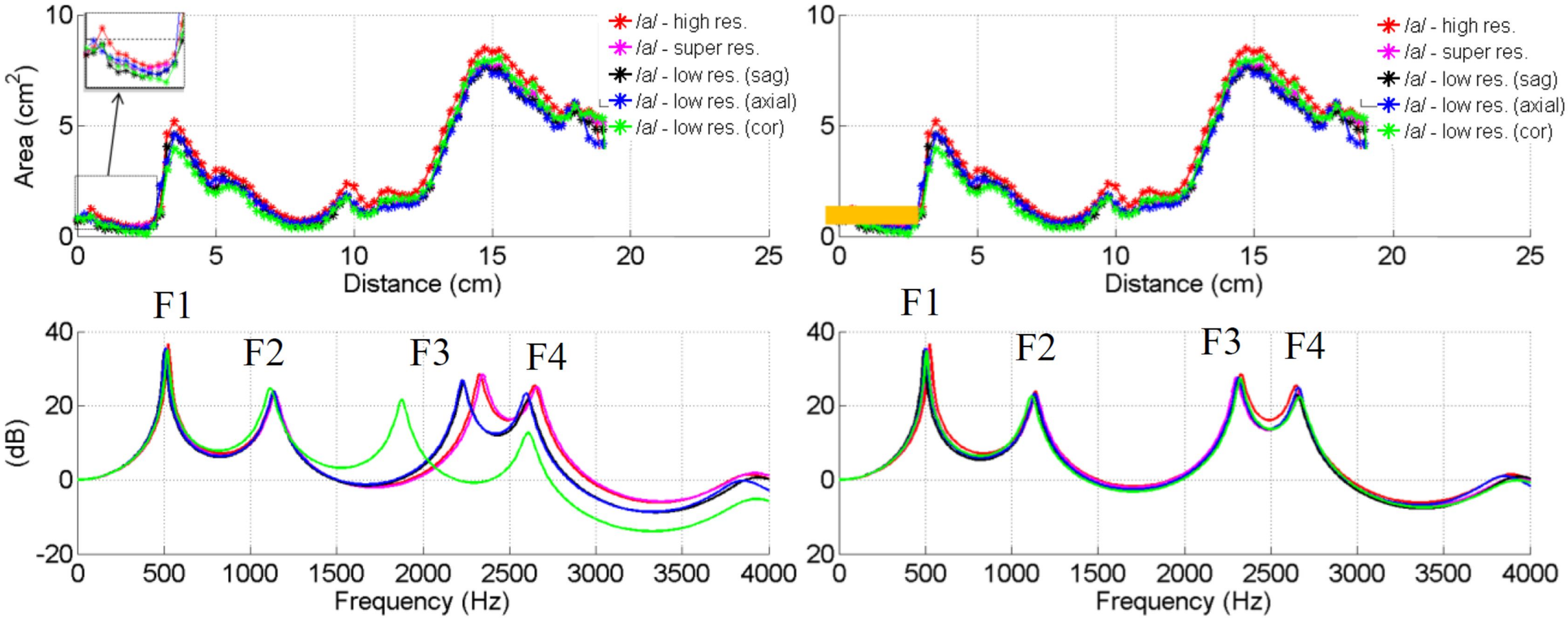
Narrower pharynx

D) Super-resolution volume

Above features are improved here

Area function

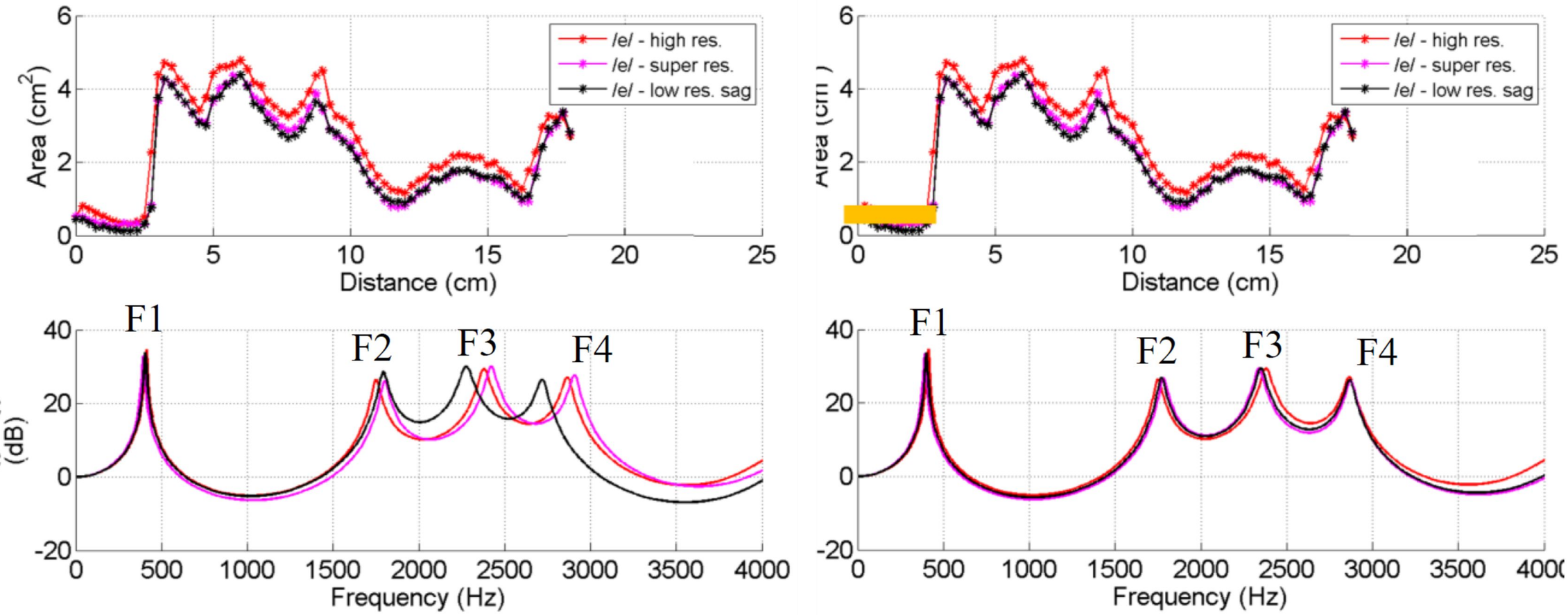
Acoustic response



(a)

Area function

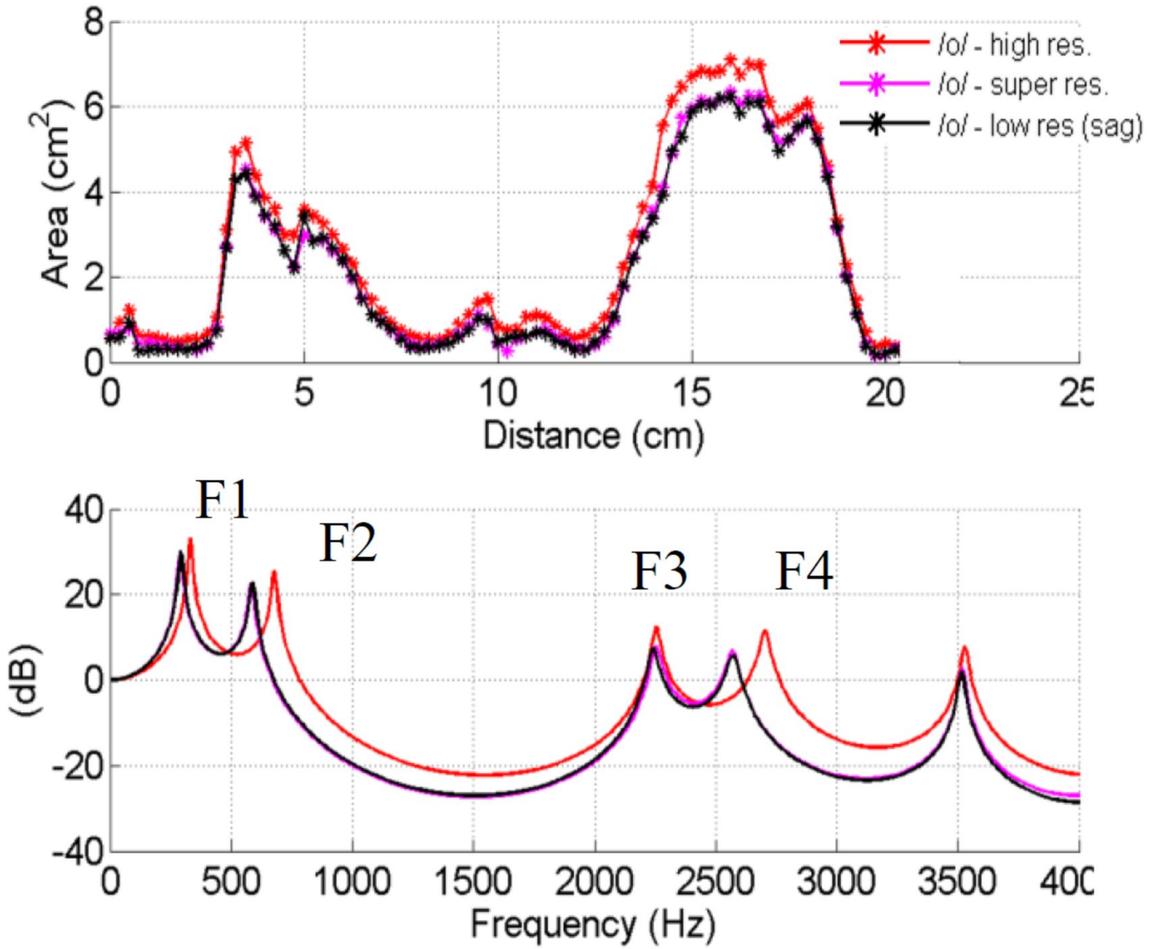
Acoustic response



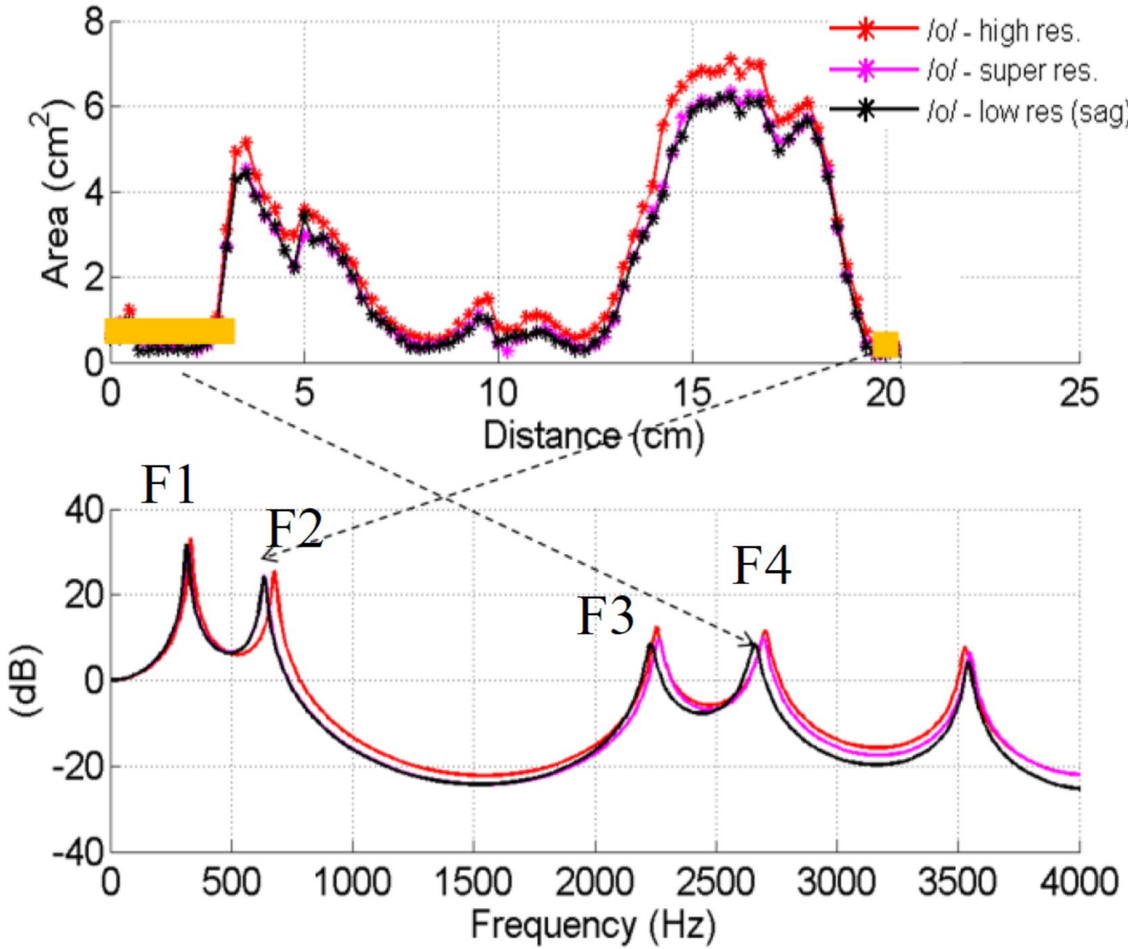
(b)

Area function

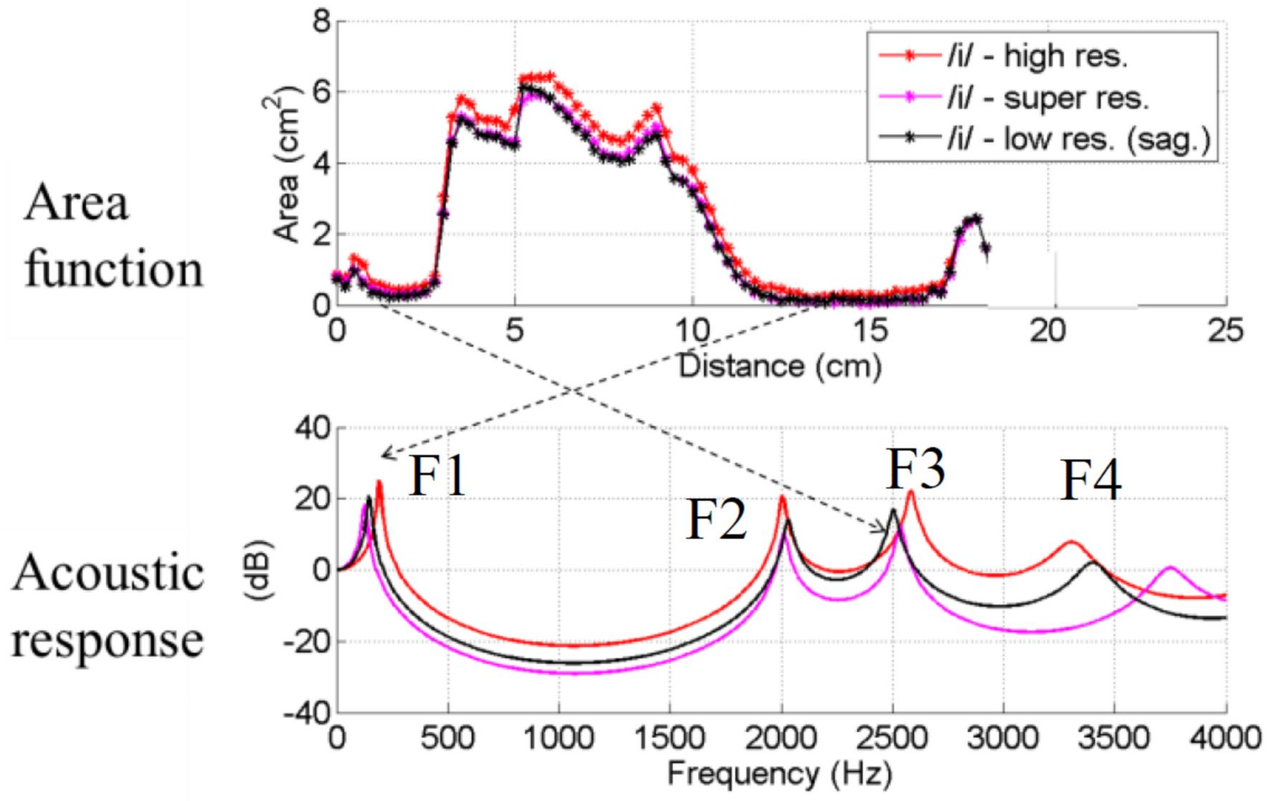
Acoustic response



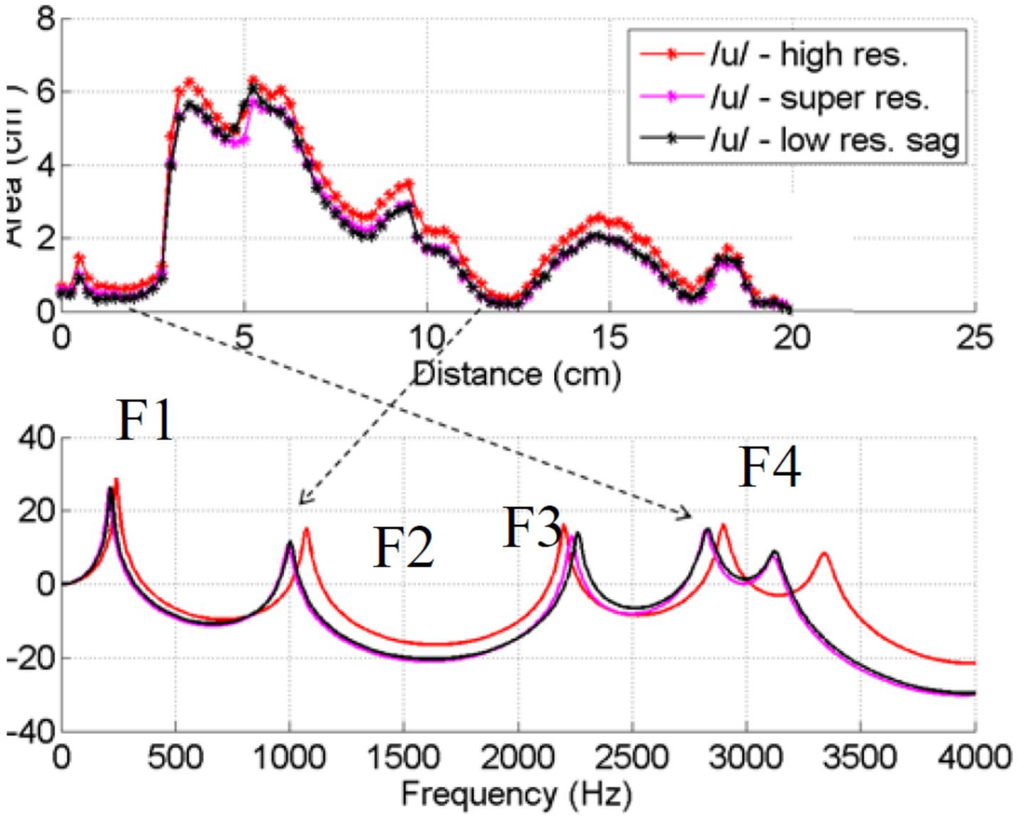
(a)



(b)


Area function
Acoustic response
/i/ - high res.
/i/ - super res.
/i/ - low res. (sag.)
Area (cm²)
Distance (cm)
F1
F2
F3
F4
(dB)
Frequency (Hz)

(a)


/u/ - high res.
/u/ - super res.
/u/ - low res. sag
Distance (cm)
F1
F2
F3
F4
Frequency (Hz)

(b)