ON-LINE VISUALIZATION OF SPEECH ORGANS USING MRI: A 3D APPROACH TO SPEECH ARTICULATION MODELING

G. Ye. Kedrova[i], N. V. Anisimov[ii], Y. A. Pirogov[ii]
Moscow State Lomonosov University
[i]Philological Faculty, [ii]Educational-Research Center of Magnetic Tomography and Spectroscopy
119992 Moscow, GSP-2, Leninskije Gory, MGU, 1st Humanities
Russian Federation
kedr@philol.msu.ru

ABSTRACT
A three-dimensional on-line visualization of the vocal tract during speech production was performed based on MRI data obtained from a female speaker producing the six Russian vowels. These images were collected using original method of 3D MRI-scanning where the starting moments of MRI processes enabled co-operative activities from a patient's side via a special remote-control device. A stroboscopic method of data acquisition was used to reconstruct real articulatory processes for each speech stimulus. Data show clear real-time movements of the lips, tongue, underjaw and mandible, as well as velum and facial surfaces. Thus collected animated data could be exposed for improved teaching and learning foreign languages' (in our case, the Russian language) technology, as well as for speech synthesis based on a physiologically relevant articulatory model. Sample movies and data analysis strategies are presented.

KEY WORDS
CALL, Simulation, Modeling, MRI, speech articulation.

1. Introduction

At present, the need for 3D articulatory data from various languages is motivated by further development of speech production models (a faithful 3D articulatory speech synthesis), as well as by meeting the most demanding challenges of innovative CALL applications and other open questions of speech rehabilitation activities. Meeting this needs a MRI investigation of the model articulatory patterns in various languages has been developing at a quick pace and has written since its early stage a long and rather successful history.

Early research in the field dates from the mid-60s of the previous century. From that time onwards a rich pool of empirical data on different types of vocals and consonants (primarily fricatives, approximants, laterals, retroflex and nasals) in French (P. Badin), Swedish (O. Engwall), German (P. Hoole, B. Kröger), British and American variants of the English language (M. Tiede, S. Narayanan) as well as in some non-Indo-European languages (K. Honda, S. Maeda) is available in the literature.

Recently another ambitious target - 3D modeling of the speech articulation in dynamics (articulatory synthesis) has determined the main trend in the area.

Until now the state of the art in articulatory 3D modeling was roughly dominated by two main productive schemes. One is using faster and more accurate MRI to measure the vocal tract at different points along the direction of the air stream (e.g. sagittal, coronal, coronal oblique and transversal cuts) while producing speech signals; the measurements' results evolving into the vocal tract's shape reconstruction relying upon this data [1]. This idea has been very soon estimated to be very productive and a series of experimental research conducted in the same strain were not long in coming, resulting thus in a very realistic 3D model of speech articulation processes – a 3D articulatory synthesis called "Virtual Talking Head" [2].

Currently a series of assessments is under way to test its validity and effectiveness. A preliminary evaluation of the contribution of tongue display (exposed in Talking Head avatar) to speech understanding in various degrees of noisy conditions has presented very promising results, thus giving the authors more stimuli to use the augmented speech capabilities of the virtual talking head for applications in the domains of (1) speech therapy for speech retarded children, (2) perception and production rehabilitation of hearing impaired children, and (3) pronunciation training for second language learners [3].

A completely different approach to 3D articulatory modeling deals with core principles of classical parametric speech synthesis. Most elaborated three-dimensional articulatory model of the vocal tract for Multimodal Speech Synthesis based on parametric descriptions of both the acoustic and visual speech modalities, i.e. in a text-to-speech framework, has been developed at the Department of Speech, Music and Hearing of The Swedish Royal Institute of Technology

(Kungl Tekniska Högskolan – KTH). The model, described in [4], consists of vocal and nasal tract walls, lips, teeth and tongue, represented as visually distinct articulators by different colours resembling the ones in a natural human vocal tract. The internal part of the model includes meshes of the tongue, palate, jaw and the vocal tract walls based on the analysis of three-dimensional Magnetic Resonance Imaging (MRI) data of a reference subject [5]. Using statistical analysis, six articulatory parameters were defined to control the tongue shape. As it was crucial for the proposed application that articulations and articulatory movements were natural and that the timing between the facial and tongue movements was correct, simultaneous measurements of the face (with optical motion tracking of reflective markers) and tongue (with electromagnetic articulography) movements [6] have been used to train the two models in a coherent way.

The main hope of the authors was that a realistic 3D model of the tongue, made visible in the frame of a synthetic face can be of use in pronunciation training to provide visual feedback to hearing-impaired children. The future ambition of this research group as proclaimed in [4] is to create a tutor that can be engaged in many aspects of language learning from detailed pronunciation to conversational training. Later on such virtual language teacher named Ville was successfully created for teaching and learning Swedish as foreign language. Ville is designed to present language-specific distinctive features in a meaningful contrastive situation. As is described in [7], he can detect and give feedback on pronunciation errors, and has many challenging exercises that are used in order to raise the student's awareness of particular perceptual differences between their first and second languages, or to teach new vocabulary.

One of the most challenging capacity of the approach is that according to the authors' point of view it enables a flexible architecture that allows to create new characters either by adopting a static wireframe model and specifying the required deformation parameters for that model, or by sculpting and reshaping an already parameterized model.

Thus, one could state that after a certain period of intensive internal development the technology of constructing 3D models of articulatory processes based upon MRI data has evolved and even resulted in some efficient practical applications, at least for several languages (i.a. French and Swedish).

However, unlike the latest global trends in experimental and applied phonetics, the MRI investigation of speech production mechanisms in Russia is still at its outset. There is a considerable deficiency of authentic experimental research and/or reliable data for Russian articulatory patterns. Only few pilot research based upon 2D on-line MR-imaging techniques are yet available either for the vocal or consonant system of the Russian language.

It should be also mentioned that until now there is no 3D MRI techniques applications targeted to the studies of the Russian vocalic articulation. Our current research aims to fill this gap.

2. Preliminary work

The research in question forms an integral part of a broader magnetic resonance imaging investigation of the complete inventory of articulatory motor patterns representative for contemporary Russian language pronunciation practice we are busy with. According to our main targets special interest was focused upon the Russian vocalic system, since recent investigation has argued that it was primarily vowel categories that might have distinct language-specific acoustic and/or articulatory motor patterns [8], and therefore could be considered as a key factor managing any language basic articulations.

Up to now any investigation of the Russian speech articulatory patterns was based on 2D MRI data. Our research within this experimental trend has led to the construction of empirical database encompassing complete inventory of Russian vocalic and consonants articulations along with some basic co-articulation patterns. This database was based on 2D on-line MR-images of the sagittal cut of articulatory tract enlarged with audio- and video recordings taken from native speakers of Russian with standard pronunciation and without any perceptible articulation disease. The 2D MRI investigation of the Russian sound patterns (phonemes) was based upon admitted procedures and techniques which were expanded with several new original methods offered and successfully tested by the Russian team of experimentalists [9].

Thus, all of our speaking subjects were required to repeat experimental stimuli at their own pace as many times as possible during acquisitions of MR images. The experimenter instructed the subject to initiate the speech process by counting every second, a couple seconds before the MRI acquisition starts. Meanwhile we've arranged simultaneous audio recordings taken via a microphone LifeVideo(tm) fixed on a receiver's coil close to the speaker's mouth. As this recording was strongly dominated by the MR scanning machine noise, it was impossible to label reliably and in details the speech sequences, therefore a parallel recording of the starting points of MRI sequences was also previewed. Both recordings were presented as a two-channel oscillogram (Figure 1), which enabled more precise timing of an MR image with a particular phase of phonation, as well as with any phase of pausal period in the future procedures of images' identification. Irrespective of the MRI sessions extra control audio and video recordings of the same speech data from the same reference subject producing

speech in the same position were realized in a professional record's studio environment.
All types of the data obtained in each experimental session are disposed on Figure 2.

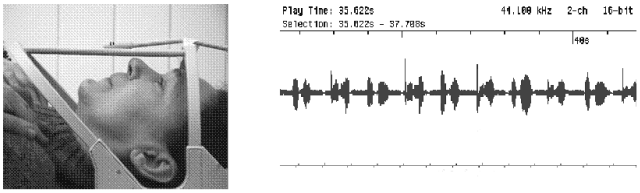

Figure 1. Position of a reference subject in MRI experiment. To the right: two-channel oscillogram displaying phonation and starting points of MRI sequences.

The speaking subject was asked to produce a series of Russian vowel phonemes [a:], [o:], [u:], [i:], [e:], [ɪ:], repeating each stimulus several times during a session of MR image acquisition. A reference subject (a native Russian language male speaker born and grown up in Moscow) has been reproducing each vowel phoneme up to 33 times in every experimental session, the aggregate total of relevant MR-images being 768 items from each speaker. The whole data set of MR images collected in all experimental MRI acquisitions was identified and ascribed to each phase of a phoneme realization.
During our investigation we've done several sessions of MR image acquisition, namely: three separate experimental sessions with the time gap of one month and one year, but using the same reference subject producing the same language stimuli through all MRI acquisitions.

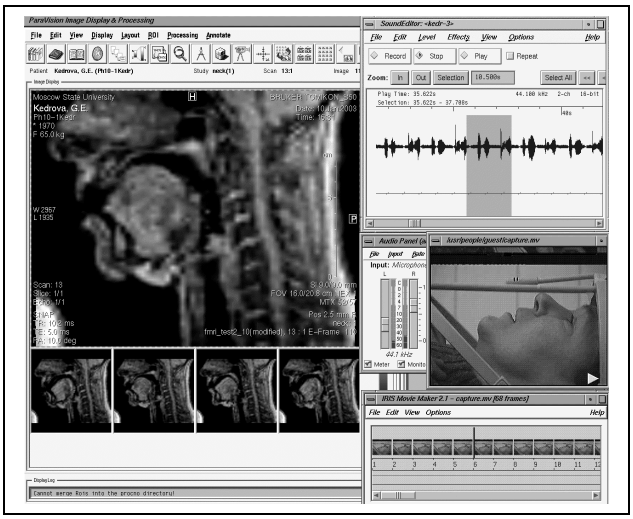

Figure 2. **Experimental data of MRI acquisitions' interface, post-MRI session control audio and video recordings incorporated.**

It is worth mentioning that in all experimental sessions we've observed the highest degree of image matching within each speaker's various performances of a particular vocal and consonantal phoneme under investigation, though certain dependencies of a latter one from a vocal context were also observed (typical vocal articulations are disposed on Figure 3).

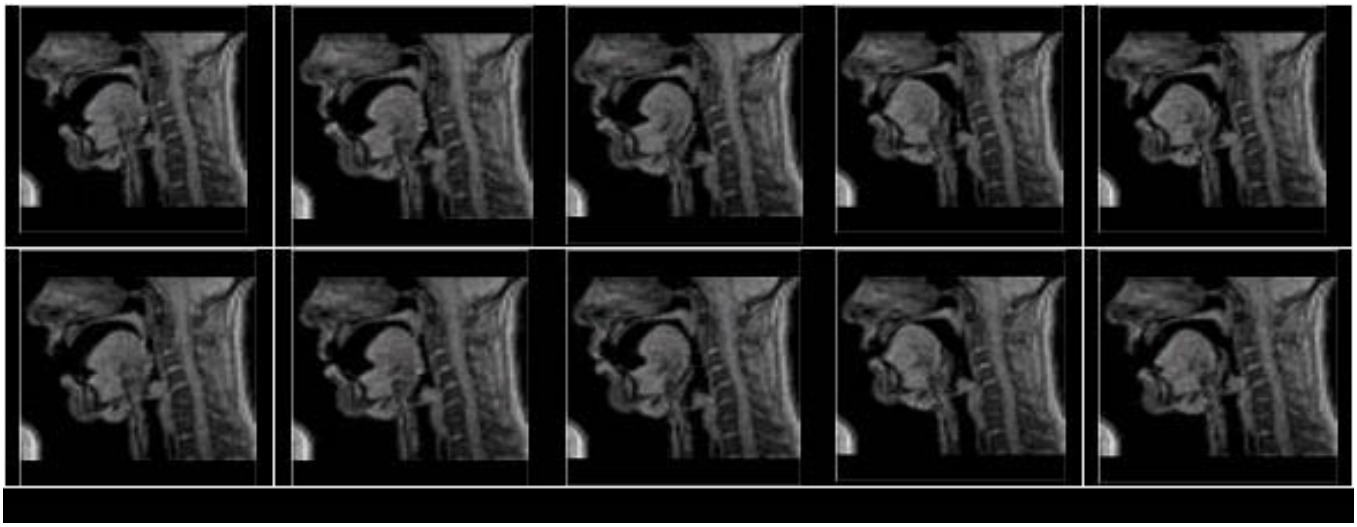

Figure 3. Experimental data of MRI acquisitions of the Russian vowels' articulation (female speaker, first and second experimental sessions). Vowel phoneme sequence is (from left to right): [a: o: u: i: e:].

Thus, we considered this results as another evidence in support of the linguistic concept of a phoneme defined as a specific psychomotor complex – "sound of language", that has been first hypothesized by the Russian linguist of Polish origin J. Bodouin de Courtenay [10], and later has been elaborated in more details by V. Bogoroditsky and

L. Scherba. V. Bogoroditsky insisted on objective reality of a notion "sound of a language" and defined it as a "psychomotor complex formed in the early childhood via association of contiguity" [11]. This primary knowledge was used as a keystone for 3D MRI experiments as well as for further processing of the collected data.

The main difference between 2D and 3D MRI experiments dealt with special techniques of the 3D MR images' acquisition elaborated by the authors [12] and successfully tested in our further experimental MRI sessions.

3. Materials and Methods

The 3D MRI experiments were realized on a 0.5 T MR system (Tomikon S50 "Bruker"). MR signal was received by quadrature neck coil. The speaking subject was lying in supine position on the horizontal patient couch with his head placed inside the receiver coil of the MRI unit.

The main peculiarities of the current MRI techniques dealt with providing additional capabilities for the patient (speaking subject) to control and arbitrary direct the whole process of MR scanning. Thus, our method enabled voluntary interruption of scanning processes by the patient at any moment of the experimental session. For this purpose the speaking subject was equipped with the button switch connected through cable with the MRI unit. The subject could thus intentionally synchronize every launching of MRI series with a certain phase of the speech production processes. Primary testing of this method has proved its high efficiency in such applications of MR investigation as abdomen research and some provisional articulatory organs studies [9].

MR scanning activity was executed on sagittal cut to a field of view 220*240 mm with the slice thickness of 2 mm and total number of slices – 95 units. Aggregated thickness of the whole dataset – 170 mm. By interpolation the initial 3D matrix was transformed into a new 3D matrix 128*128*128 enabling generation of MR images in three orthogonal projections. The central cut of the sagittal projections was taken as a benchmark for exact localization of coronal and axial cuts (Figure 4).

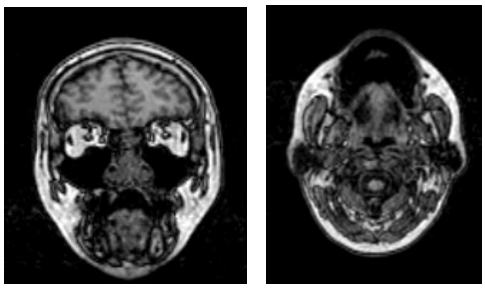

Figure 4. Correlation of sagittal, coronal and axial cuts on MR images taken on-line during speech articulation activity.

Alternatively, the coronal and axial cuts were used as benchmarks for identification and localization of the sagittal projections (Figure 5).

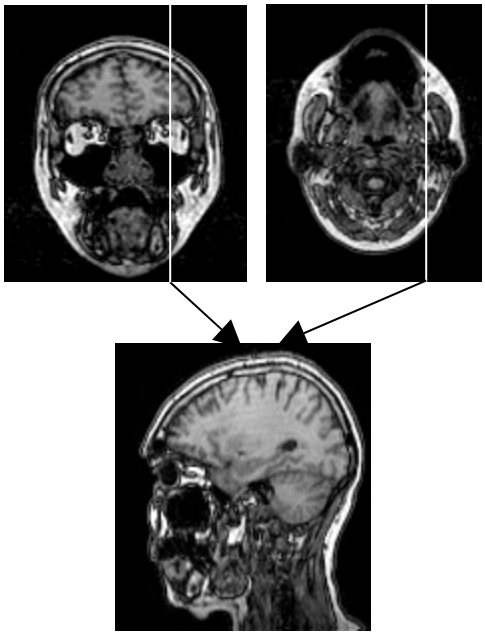

Figure 5. Correlation of the coronal, axial and sagittal cuts on MR images taken on-line during speech articulation activity.

4. Results

The scanning data presented in a matrix was used as a main basic source for re-constructing 3D representations of a scanned object; whereas editing certain elements of the data matrix provided additional capabilities for modification of the 3D model. All types of the 3D models reconstructed in each experimental session for every speech stimulus are disposed on Figure 6.

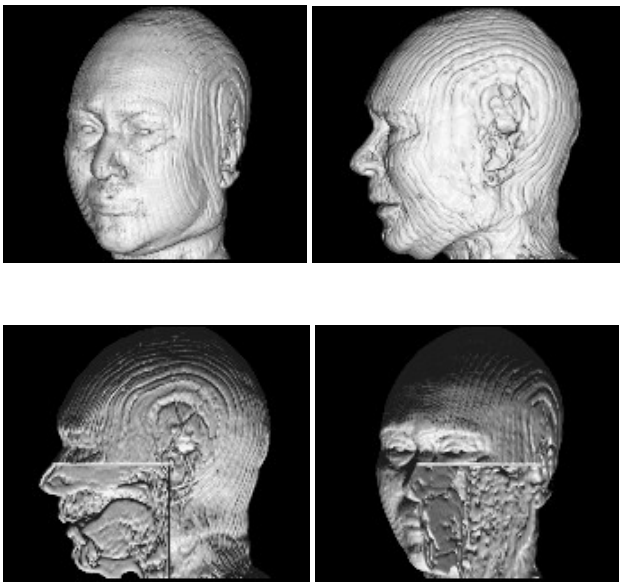

Figure 6. 3D models of speaking subject's activity reconstructed upon MR images taken on-line during speech production processes.

On the next stage of our investigation the 3D reconstructions of the dynamic processes of the speech activities were imitated through animated simulation of articulatory processes for all experimental stimuli. Every simulation was synchronized with the corresponding audio signal recorded in the record studio environment (in this experiment the subject strove to imitate under the control of researcher his phonation patterns inside the tomography apparatus). The aggregated database of static MR images as well as dynamic simulations synchronized with corresponding speech signals were introduced to teachers and learners of Russian as a foreign language at the appropriate Department of the Philological Faculty of our University as learning and training e-resource. An example of the learning material samples is presented in the Annex. Testing and assessment of the data implementation will follow in the future.

5. Conclusion

The methodological and technological approach elaborated in current investigation has proven its validity and could be recommended for implementation. Preliminary results of our research make a significant contribution for further development of 3D modeling techniques based on MRI data, as well as elaboration of efficient computer simulation technology for visualization of human articulatory strategies. Thus collected animated data should be recommended for improvement of methodology and practice in teaching and learning foreign languages (in our case, the Russian language) as a substantial support. Our data could be also used in those speech synthesis systems that pretend to be based on physiologically relevant articulatory models. In the future work we are planning to test tentative advantage of implementation of the elaborated technology into computer-supported language learning.

References

[1] D. Demolin, T. Metens, A. Soquet, Three-dimensional measurement of the vocal tract by MRI, *Proc. ICSLP-1996*, Philadelphia, PA, 1996, 272-275.

[2] P. Badin, A. Serrurier, Three-dimensional modeling of speech organs: Articulatory data and models, *Transactions on Technical Committee of Psychological and Physiological Acoustics, Acoustical Society of Japan, Kanazawa, Japan, 36*(5), 2006, 421-426.

[3] Y. Tarabalka, P. Badin, F. Elisei, G. Bailly, Can you "read tongue movements"? Evaluation of the contribution of tongue display to speech understanding, *Conférence Internationale sur l'Accessibilité et les systèmes de suppléance aux personnes en situation de Handicaps (ASSISTH)*, Toulouse, France, 2007, (accepted).

[4] O. Engwall, P. Wik, J. Beskow, G. Granström, Design strategies for a virtual language tutor *Proc of ICSLP-2004*, Jeju Island, Korea, 2004, vol. III: 1693-1696.

[5] O. Engwall, Combining MRI, EMA & EPG in a threedimensional tongue model, *Speech Communication*, *vol. 41/2-3*, 2003, 303-329.

[6] J. Beskow, O. Engwall, and B. Granström, Resynthesis of facial and intraoral motion from simultaneous measurements, *Proc of the 15th ICPhS,* Barcelona, Spain, 2003, 431-434.

[7] Hjalmarsson, A., Wik, P., & Brusk, J. Dealing with DEAL: a dialogue system for conversation training, *Proceedings of SigDial*, Antwerp, Belgium, 2007, 132-135.

[8] L. Koenig, *Towards a physical definition of vowel systems of languages* (Yngve, V.H., Wasik, Z. (eds.). Hard-science linguistics, Continuum, 2004, 49-66.

[9] Н. Анисимов, Г. Кедрова, Л. Захаров, Ю. Пирогов, МРТ-визуализация процессов артикуляции при порождении речи, *II Евразийский конгресс по медицинской физике и инженерии, Медицинская физика – 2005,* Москва, 2005, 239-240.

[10] Бодуэн де Куртенэ И.А. Курс грамматики русского языка. Ч. 1. Фонетика. Варшава, 1887.

[11] В.А. Богородицкий, Очерки по языковедению и русскому языку. М., Едиториал УРСС, М., 2004.

[12] Н. Анисимов, Г. Кедрова, Ю. Пирогов, Магнитно-резонансное сканирование, управляемое пациентом, *Научная сессия МИФИ-2008, том 3*, Москва, 2008, 124-125.

Annex