C. R. Biologies 325 (2002) 547–556
© 2002 Académie des sciences / Éditions scientifiques et médicales Elsevier SAS. Tous droits réservés
S1631069102014580/FLA

Model-driven acquisition / Acquisition conduite par le modèle

Real-time MRI and articulatory coordination in speech

Didier Demolin[a]*, Sergio Hassid[b], Thierry Metens[c], Alain Soquet[a]

[a] *Laboratoire de phonologie, université libre de Bruxelles, 50, av. Franklin-Delano-Roosevelt, CP 175, B-1050 Bruxelles, Belgium*
[b] *Service ORL, hôpital Erasme, université libre de Bruxelles, 808, route de Lennik, B-1070 Bruxelles, Belgium*
[c] *Unité de résonance magnétique, hôpital Erasme, université libre de Bruxelles, 808, route de Lennik, B-1070 Bruxelles, Belgium*

Received 12 April 2002; accepted 22 April 2002

Presented by Michel Thellier

Abstract – This paper describes the real-time MRI technique and its use for the study of speech production. The two major problems, (*i*) the simultaneous recording of the MR images and the speech signal and (*ii*) the synchronisation of the images and of the speech signal, are addressed. Measurement accuracy on real-time images is evaluated by comparison with similar measurements on static MR images. ***To cite this article: D. Demolin et al., C. R. Biologies 325 (2002) 547–556.*** © 2002 Académie des sciences / Éditions scientifiques et médicales Elsevier SAS

magnetic resonance imaging / speech production

Résumé – IRM temps réel et coordination des mouvements articulatoires en production de la parole. Cet article décrit la technique d'IRM temps réel et son application à l'étude de la production de la parole. Des solutions sont proposées aux deux problèmes majeurs liés à l'étude de la parole, à savoir (*i*) l'enregistrement simultané des images MR et du signal de parole et (*ii*) la synchronisation de ces deux types de données. La précision de mesures effectuées sur les images temps réels est évaluée par comparaison avec des mesures équivalentes sur des images statiques. ***Pour citer cet article : D. Demolin et al., C. R. Biologies 325 (2002) 547–556.*** © 2002 Académie des sciences / Éditions scientifiques et médicales Elsevier SAS

imagerie par résonance magnétique / production de la parole

1. Introduction

Baer et al. [1] initially demonstrated the imaging of the vocal tract during phonation with Magnetic Resonance (MR). Since then, the acquisition speed of MR scanner has improved considerably, making fast or even real-time imaging possible [2, 3]. Using a recent MR scanner, we obtained the simultaneous MR acquisition of multi oblique slices with a 1 mm × 1 mm spatial resolution in a sole 14 s acquisition. This provided an improved generation of area functions, which is an important step in the study of the relation between vocal tract geometry and speech acoustics. Similarly, the acquisition of parallel joint slices is also possible. Sustained phonation was required during these acquisitions (14 s) [4]. Nevertheless, these experiments were restricted to the study of oral or nasal vowels, because of the low temporal resolution. Progresses made to increase the temporal resolution are limited by low signal-to-noise ratios and by susceptibility artefacts

*Correspondence and reprints.
E-mail address: ddemoli@ulb.ac.be (D. Demolin).

when fast gradient echo techniques are involved. In the present work, we adapted an ultra-fast implementation of Turbo Spin Echo (TSE) called the TSE Zoom sequence (previously named Lolo [5, 6]) to achieve a real-time continuous monitoring of the vocal tract with an actual time resolution of 4–6 images per second. In TSE Zoom imaging, all the raw data needed for the acquisition of one image are collected in a single shot. In this sequence, a single RF excitation pulse is followed by a Carr–Purcel Meiboom Gill train of 180° refocusing pulses, generating a train of spin echoes, where each echo is separately phased [7]. The TSE Zoom sequence is designed such that the initial RF excitation pulse and the subsequent 180° refocusing pulses excite perpendicular slabs, resulting into an intersecting slice, free of foldover artefacts and without compromising the spatial resolution. The combination of this feature with a rectangular field of view and a half Fourier acquisition makes a very short single shot acquisition possible. We aimed to show that this technique could be used to study the relative movements of the main articulations involved in speech production, i.e. lips, tongue, larynx, lower jaw and velum.

2. Material and method

Subjects were lying in supine position in a 1.5 T MR system equipped with fast gradients (CompactPlus, PowerTrak 6000, 20 mT m^{-1} and 100 mT m^{-1} ms^{-1} maximum amplitude and slew rate, Philips Medical Systems, Best, The Netherlands). The receiver coil was a quadrature neck coil. All subjects gave their informed consent before the experience; none of them presented counterindications to MR examinations. A total of 2 subjects were involved in the study. The subjects were fixed with solid foam cushions in order to prevent any unwanted movement of their head. MR procedures were performed in accordance with FDA and European Rules concerning the safety of individuals, including the specific absorption rate below 4 W kg^{-1} body weight. We first acquired survey images in three orthogonal planes. Then, on the basis of these survey images, a stack of 11 parallel TSE proton density-weighted images was positioned in the sagittal plane. These static views were simultaneously acquired during a 12 s continuous phonation. Each of these images covered a field of view of 250 mm × 200 mm, with a spatial resolution of 0.95 mm^2. Finally, real-time studies were performed with the TSE Zoom sequence: a single sagittal T1-weighted section of 6 mm thickness was continuously acquired during at least 20 s, at the rate of four or five images per second. This section was carefully positioned in the mid-sagittal plane on survey images and on static TSE proton-density weighted images. In a few subjects, for the sake of comparison, acquisitions at a rate of six images per second were also obtained. For the four images per second real-time acquisition, we used *TR* = 250 ms, with 98 ms acquisition and 152 ms delay between acquisition of consecutive images, to allow some magnetisation T1-recovery. In the context of real-time imaging, *TR* stands for the time period of the continuous acquisition of single shot images, i.e. the time resolution of the dynamic acquisition process. For the five images per second acquisition: *TR* = 200 ms, with 98 ms acquisition and 112 ms delay (Fig. 1 describes the *k*-space sampling of this sequence). All acquisitions were implemented with the following parameters: *TE* = 21 ms, flip angle = 60°, echo spacing = 4.4 ms, water-fat shift = 0.5 pixels, turbo spin-echo factor (number of echoes in the echo train) = 20 and 62.5% partial Fourier acquisition. The rectangular field of view in the antero-posterior and caudo-cranial directions was of 125 mm × 250 mm, with a 32 × 128 matrix, providing a spatial resolution of 3.9 mm × 1.95 mm. For the six-image-per-second acquisition, we used *TR* = 169 ms, *TE* = 15 ms and a train of 18 echoes. These images were acquired in the sagittal and transversal oblique sections and compared with four- and five-image-per-second acquisitions in the same orientations. Images were reconstructed and displayed in real time. The operator instructed the subject to initiate the speech process by counting every second, five seconds before the acquisition start. The simultaneous recording of the speaker's voice using the MR built-in intercom during the acquisition of real-time images at a rate of four and five images per second were performed with several subjects (Figs. 2 and 3).

3. Simultaneous rtMRI and speech signal acquisition

In order to transform rtMRI into an adequate tool for the study of speech production, the simultaneous recording of the speech sound produced by the subject and the MR images is of course essential. The recording of speech sound produced inside the machine during image acquisition undergoes two major problems. First, no metallic part can be approached to the head of the subject for both safety reasons and in order to avoid images artefacts. Second, the acquisition process by itself produces a large amount of noise. The noise properties depend of the acquisition sequence. In the case of the rtMRI described in this paper, the noise is structured in succession of higher and lower noise intensities, corresponding to successive image acquisition.

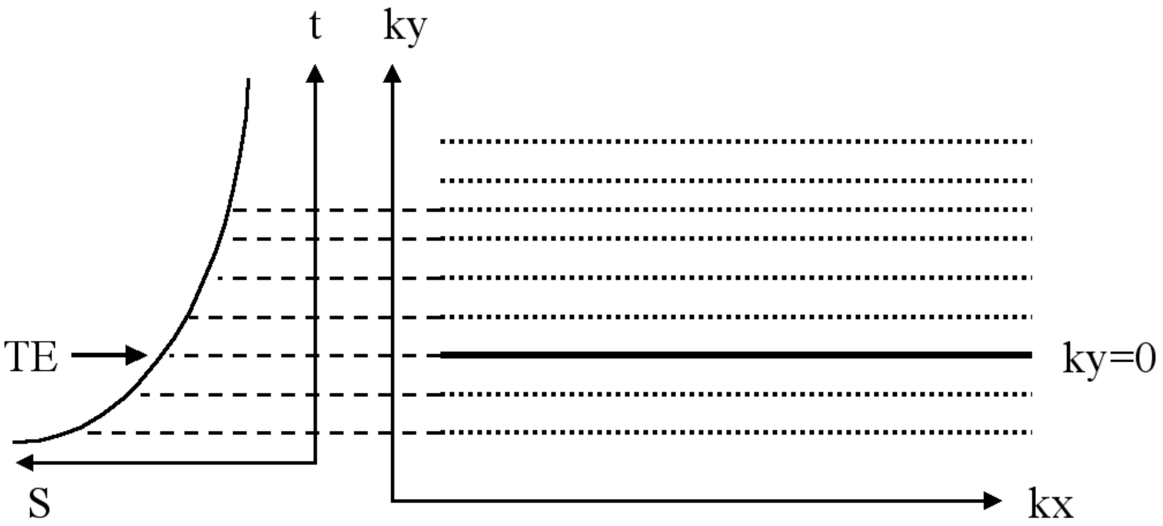

Fig. 1. *k*-space sampling with a single-shot half Fourier TSE acquisition. The frequency encoding is performed along k_x and the phase encoding along k_y. The two first echoes are acquired with negative k_y values, the third echo is sampled at the time of the k_y-encoding and provides the effective TE of the sequence, above $k_y = 0$, a few of the remaining 17 echoes are shown. The total duration of the sequence was 186 ms. Only 62.5% of the k-space was acquired (20 profiles corresponding to the 20 echoes of the train); the missing profiles were reconstructed by Fourier symmetry. Finally, the reconstructed image contains 32 phase-encoding steps and 128 frequency-encoded samples.

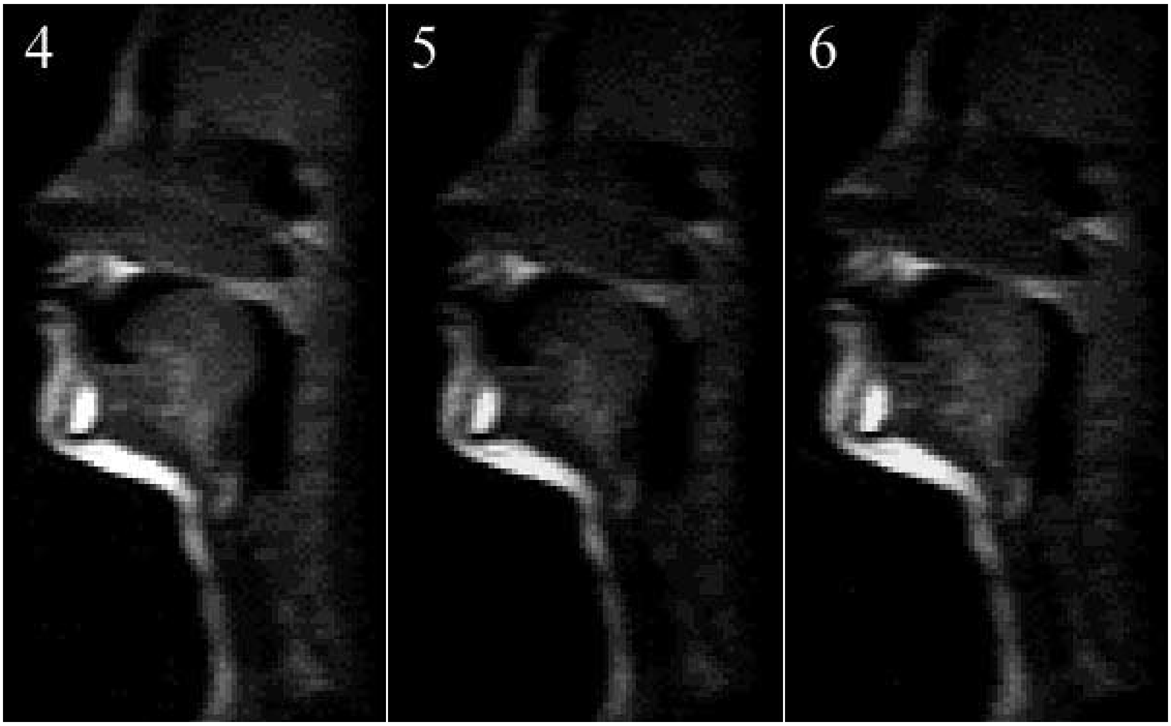

Fig. 2. TSE zoom images acquired respectively (left to right) at a rate of four, five and six images per second. The image quality degrades with increasing temporal resolution, but remains interpretable.

3.1. Recording of the speech signal

The quality of the signal that reaches the intercom is of insufficient quality to allow the experimenter to observe the speech signal. The main reason is that the sound is recorded far from the lips, with a pneumatic system.

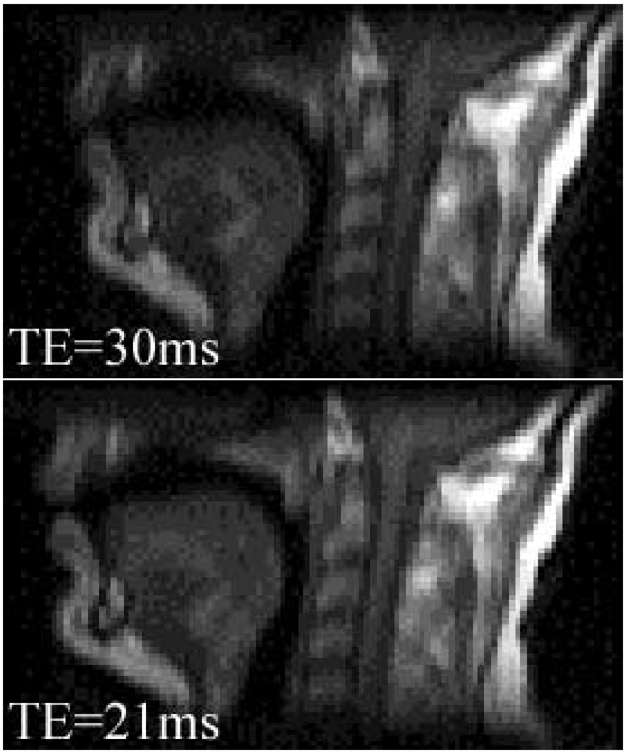

Fig. 3. TSE zoom images with *TE* = 30 ms (top, slow gradient system) and *TE* = 21 ms (bottom, fast gradient system of this study). The image acquired with the fast gradient system are less blurred and provide a better depiction of the air–tongue and air–velum interfaces, because it is less sensitive to susceptibility artefacts.

For this study, we have used an optical microphone [8]. The principle of the optical microphone is as follows. A first optical fibre carries an incident light produced by a light source. The light is then reflected by a diaphragm and transmitted through a second optical fibre to a photo detector. When the diaphragm moves according to the ambient sound, it modulates the amount of light reaching the photo detector.

This kind of microphone has two major advantages. First, it does not contain any metallic part; it can therefore be placed close to the lips without any danger or artefact. Second, the signal is carried with optical fibres; this allows the electronic and the recording device to be placed outside the acquisition room, close to the experimentalist.

Fig. 4 shows the recording of the two logatoms /apa/ and /ata/ with the optical microphone during a five-image-per-second acquisition. The figure shows the acoustic signal, the corresponding narrow band spectrogram and segmentation realised on the spectrogram.

Solutions are under investigation to further improve signal-to-noise ratio with signal processing techniques.

3.2. Synchronisation

In order to synchronise the acoustical signal to the rtMR images, we have configured the Philips MR machine to produce a TTL signal at the beginning of each image acquisition. A dedicated hardware shapes this pulse so as it can be recorded simultaneously with the acoustic signal. Fig. 5 shows the sequence of images, the recorded pulses and the acoustic signal.

4. rtMRI versus static MRI measurement precision

The objective of this first experiment is to evaluate the measurement precision that can be achieved with the rtMR images compared to the classical static MR images currently used for the analysis of speech production.

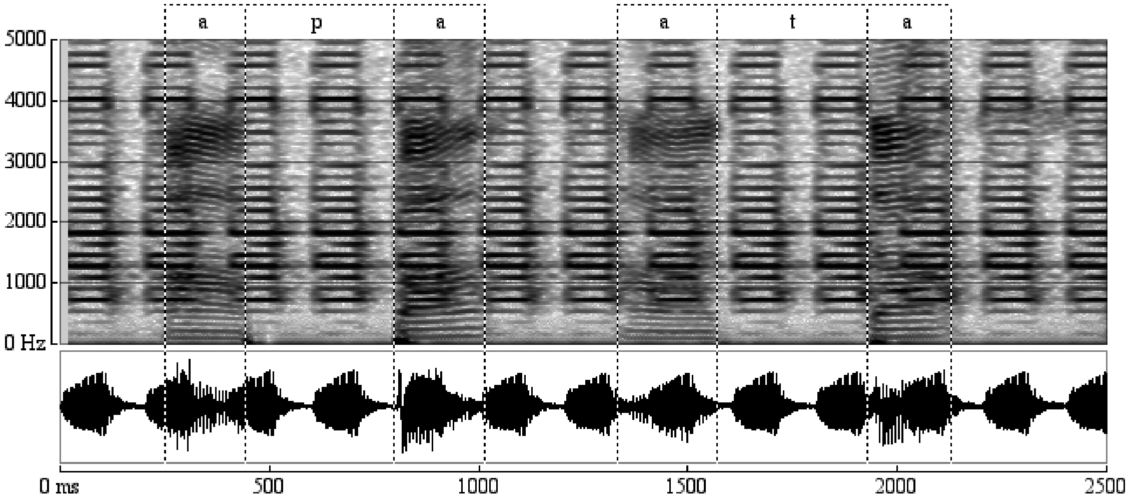

Fig. 4. Acoustic signal, corresponding narrow band spectrogram and segmentation realised on the spectrogram of a sequence of two logatoms /apa/ and /ata/.

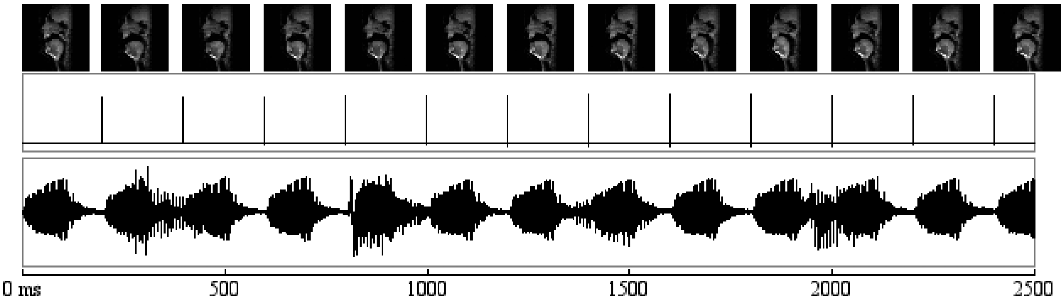

Fig. 5. Sequence of images, the recorded pulses and the acoustic signal during the acquisition of a sequence of two logatoms /apa/ and /ata/.

4.1. Material

The static views were provided by the acquisition of 11 TSE sagittal images that were simultaneously acquired during a 12 s continuous phonation. Each of these images covered a field of view of 250 mm × 200 mm. The contrast of these images was merely influenced by the proton density of the tissues. A remarkable feature of TSE images is that they are free of susceptibility artefacts, unlike other fast image acquisition techniques, like the so-called Echo Planar images [9] and gradient-echo images [10].

The real-time MRI sequence has been described here above. For this experiment, we used a single sagittal T1-weighted section of 6 mm thickness, which was continuously acquired during 20 s, at the rate of four images per second. The sections were carefully positioned in the mid-sagittal plane on survey images acquired in three orthogonal planes.

4.2. Measurements

Fig. 6 presents the images of one repetition of the sequence /ieaou/ pronounced by the male speaker. The duration of this sequence is 3.2 s, which gives a total of 16 images. It can be seen that the different articulators involved are well imaged (lips, tongue, hard palate, velum and larynx). The back of the pharynx is less contrasted. This can be explained by the fact that it is located close to the edge of the coil, where its detection sensitivity becomes poor. The articulatory configurations of the different vowels can be observed respectively on image 16 for /i/, 19 for /e/, 24 for /a/, 27 for /o/ and 31 for /u/.

In order to measure the deformation of the vocal tract, a measurement grid has been adjusted to the speakers. The grid has been adjusted on the rtMRI data, so that each grid line is in first approximation orthogonal to the fixed articulators (back of the pharynx and hard palate) and to the approximated flux line (see Fig. 7). The same grid has then been placed in equivalent place and orientation on the static MRI data. The grid is speaker-specific and is designed to allow the measurement of the lip opening (grid line 1), larynx position (grid line 14), and a set of 12 grid lines distributed along the vocal tract.

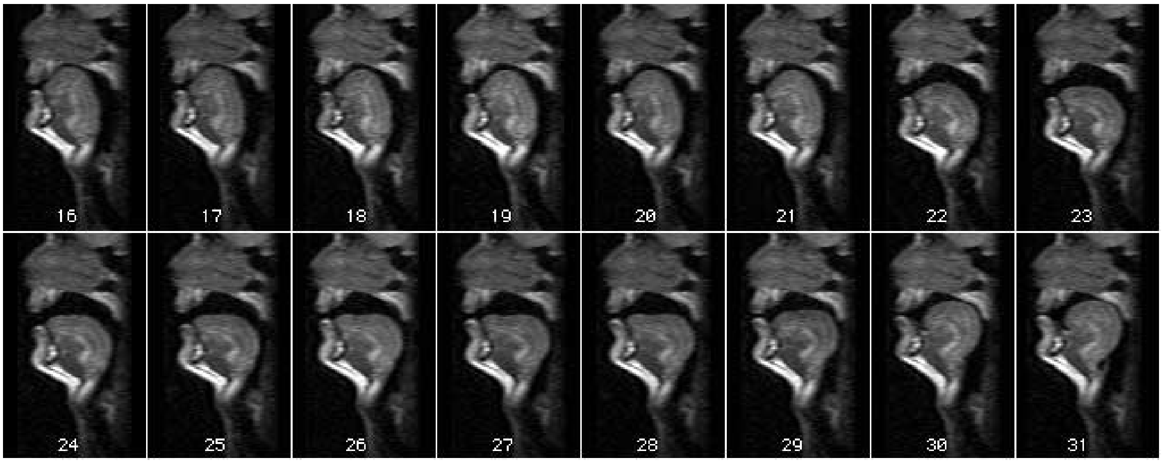

Fig. 6. Real-time acquisition of one repetition of the sequence /ieaou/ pronounced by the male speaker. The 16 images represent a total of 3.2 s, with one image every 200 ms.

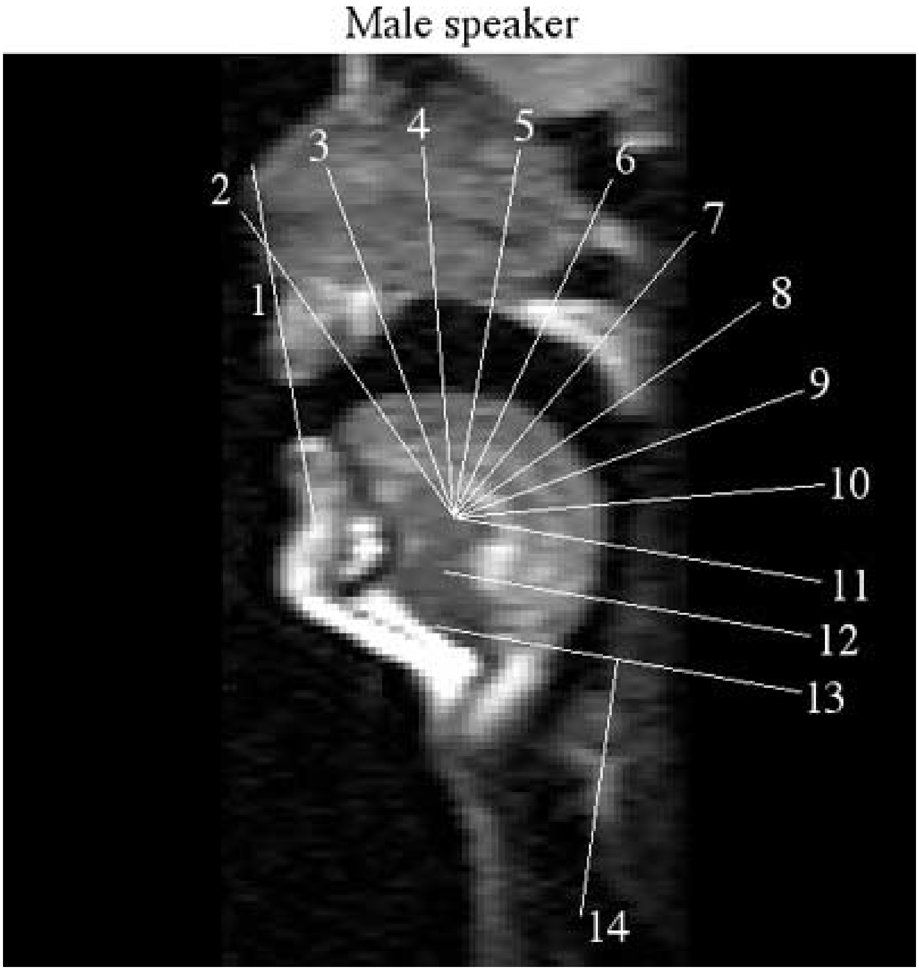

Fig. 7. Position of the measurement grid on the male speaker.

Fig. 8 shows a comparison of the sagittal distances for the vowels /i, e, a, o/ for the male speaker measured along the first 13 grid lines on a real-time acquisition and on a static acquisition [4]. For the real-time acquisition, an image located in the stable part of each vowel has been used. For the static acquisition, the medio-sagittal image of each vowel was selected. It can be seen from the results that sagittal distances are comparable for the two sets of measurements. The differences are always inferior to 5 mm, with an average of 2.0 mm. This result is acceptable regarding the fact that (1) the two acquisitions were made at a three-year interval, (2) that the tasks of the subject were different (the vowels were either produce in a sequence or sustained for 12 s), and (3) the pixel size for the real-time sequence is 3.9 mm × 1.95 mm and for the static sequence 0.98 mm × 0.98 mm.

5. Measurement of articulatory movements

In order to study articulatory movements during speech production activities, it is possible first to study every image at successive time frames independently, and then to recover the evolution in time of the structure of interest. This approach has two main drawbacks. First, as the measures are realised on successive images, it is therefore difficult to segment every image in a coherent way. Second, structures moving rapidly (like the tongue tip) produce images difficult to segment and measure. To avoid those problems, it is possible to study the data in planes obtained by sampling a segment of interest on the image (the *Y* axis) for every image in time (the *X* axis). Fig. 9 shows the result of this procedure for the 14 segments corresponding to the grid lines of Fig. 7, for the male speaker repeating the sequence /ieaou/; the beginning and end of one repetition of the sequence /ieaou/ is labelled respectively by the cursors **a** and **b**.

It can be observed that: (*i*) the movement along each grid line appears clearly, (*ii*) the problem of low contrast between the back of the pharynx and the vocal tract depends on the size of the speaker – given a width of the imaging window of 125 mm, it was easier to

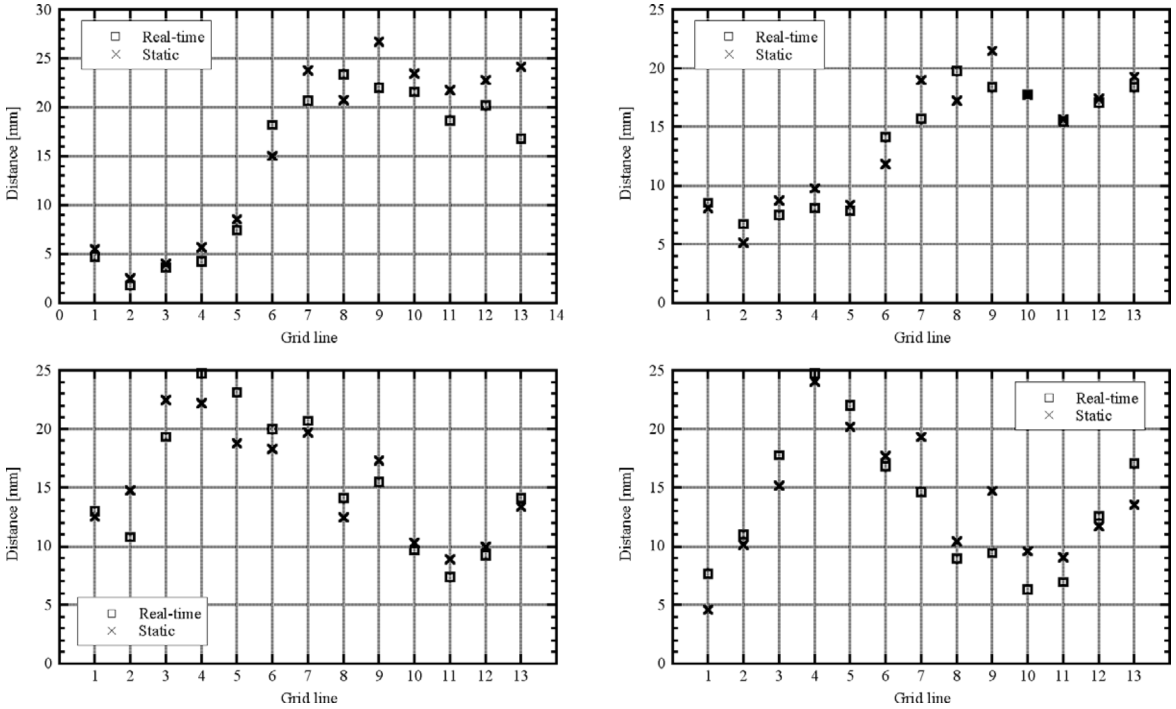

Fig. 8. Comparison of the midsagittal distances obtained with the measurement grid on four vowels in the real-time acquisition (squares) and static acquisition (crosses) for the male speaker.

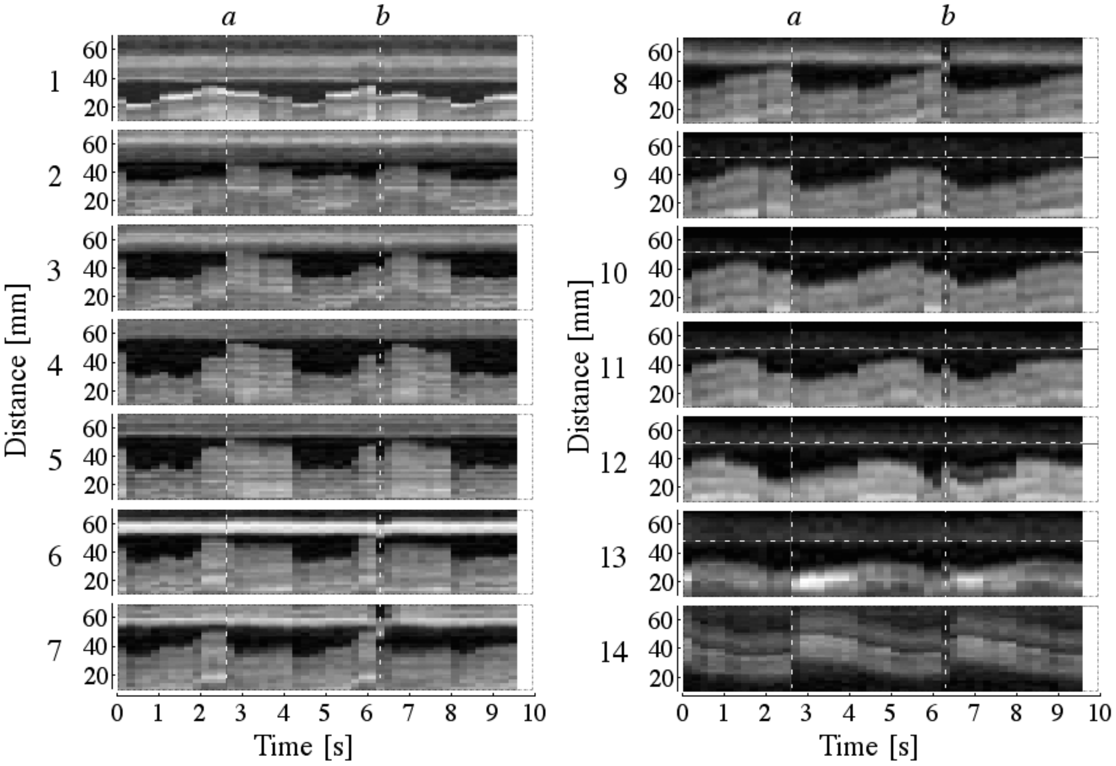

Fig. 9. Position of the measurement grid on the male speaker and analysis of the movement in time along each grid line.

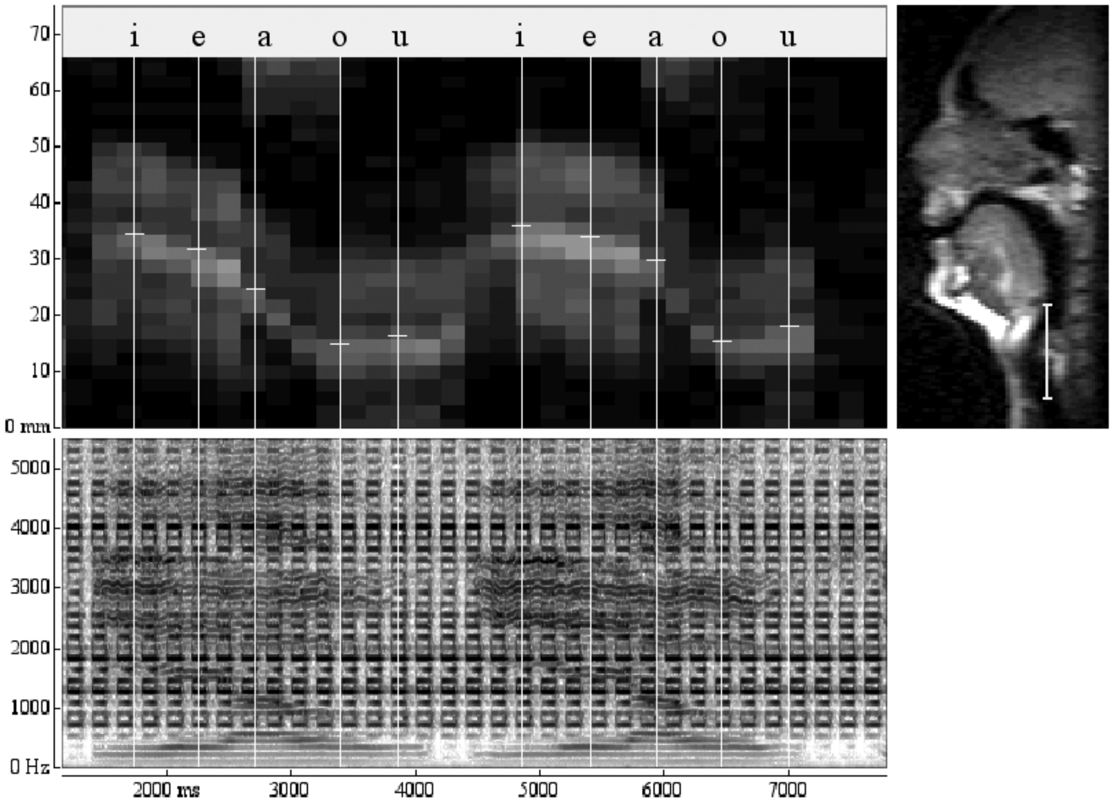

Fig. 10. Vertical position of the larynx (in mm) measured for two repetitions of the sequence /ieaou/ pronounced by the male speaker.

cover entirely the vocal tract of the female than the one of the male speaker –, (*iii*) in the hard palate region, the distances are quite stable during each vowel and move rapidly into the configuration of the next vowel. On the contrary, in the soft palate region and in the pharynx, the deformation appears to be more gradual, except for the transition from /e/ to /a/, which involves larger modifications of the overall shape of the vocal tract.

5.1. Larynx movements

Fig. 10 shows that the larynx goes downward during the production of the sequence /ieaou/ for the male speaker. The position for the five vowels is given in Table 1.

5.2. Coordination of speech movements

This second example shows the use of the real-time MRI for the study of coordination between different articulators involved in speech production. Fig. 11 shows the co-evolution of movement of the velum and the larynx height during the production of a sequence of intervocalic consonants pronounced by a male speaker.

6. Discussion

The subsecond MRI TSE Zoom technique allows exploring in real time the movement of articulators involved during normal speech production.

A remarkable feature of TSE images is that they are relatively free of susceptibility artefacts, in comparison with MR images provided by other fast acquisition techniques, like Echo Planar [9] and gradient-echo images [10]. These artefacts are especially deleterious for the imaging quality of the air–tissue limit, which is of utmost importance in the context of the vocal tract

Table 1. Vertical position of the larynx (in mm) measured for seven repetitions of the sequence /ieaou/ pronounced by the male speaker (lower values meaning lower position).

Vowel	/i/	/e/	/a/	/o/	/u/
Larynx height (mm)	32.7 (2.1)	31.8 (2.7)	27.3 (2.1)	16.4 (1.9)	17.5 (0.5)

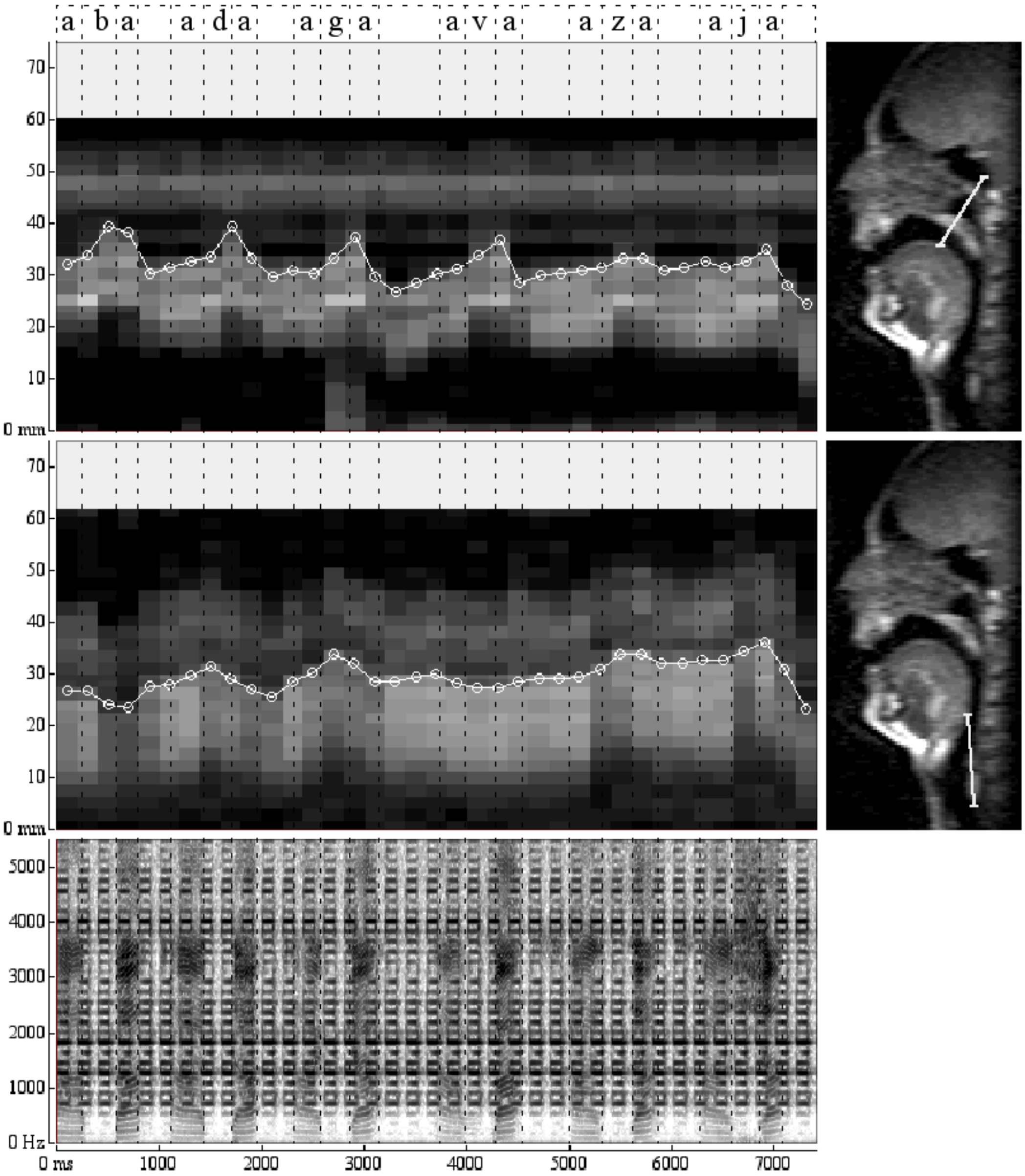

Fig. 11. Movement of the velum and larynx height during the production of a sequence of logatoms (/aba/, /ada/, /aga/, /ava/, /aza/, /aja/).

imaging. TSE zoom imaging provides a conspicuous delineation of the air–tissue limit, enabling a satisfactorily visualisation of the relative movements of the lips, tongue, larynx, lower jaw and velum. For the same reason, static TSE images provided yet an improved generation of area functions. Compared to our previous implementation of the TSE zoom sequence on a slower gradient system [6, 11], the echo time (TE) used in the present study was reduced by 9 ms, resulting in an improved image quality. Indeed, with a shorter TE, the T2 effects in the image contrast decrease and the susceptibility artefacts are further reduced. The availability of fast switching gradients enables the use of echo trains involving the same number of echoes, but with shorter echo spacing and thus shorter total train duration. This is an important feature, because it reduces the image blurring inherent to long echo train sequences [7].

When comparing images acquired with increasing time resolution, the image quality degrades, mainly because the T1 relaxation period between successive pulses shortens dramatically and with it the signal-to-noise ratio.

In order to validate the technique, we have compared the midsagittal distances measured on a measurement grid on both static MR images and real-time MR images of the same speaker pronouncing the same vowels. The results of this comparison show that the real-time technique gives accurate and reliable information on the position of the articulators involved in speech production.

The simultaneous acquisition of the speech sound remains a problem, given the high intensity of the noise during acquisition. We currently use the sound at the input of the intercom; this does not provide a quality sufficient for accurate segmentation of the speech signal. Solutions are under investigation to improve signal-to-noise ratio and to reduce noise level with signal processing. Real-time MRI allows studying the dynamics of vocal tract deformation in any plan. This permits the collection of new data and gives new perspectives to study co-articulation processes, even if the speed of image acquisition remains below compared to X-ray fluoroscopy.

Subsecond MRI technique allows to explore movement of articulators involved during normal speech production in real time. In order to validate the technique, we have compared the midsagittal distances measured on a measurement grid on both static MR images and real-time MR images of the same speaker pronouncing the same vowels. The results of this comparison show that the real-time technique gives accurate and reliable information about the position of the articulators involved in speech production.

The simultaneous acquisition of the speech sound remains a problem, given the high intensity of the noise during acquisition. We currently use the sound at the input of the intercom; this does not provide a quality sufficient for accurate segmentation of the speech signal. Solutions are under investigation to improve signal-to-noise ratio and to reduce noise level with signal processing.

Real-time MRI can be compared with other techniques:

– cineradiography provides a higher number of images per second and a much sharper image resolution, but is limited to sagittal projection and is dangerous for health;

– electro-magnetography and microbeam allow tracking of the fleshpoint usually located in the front cavity of the vocal tract and in the mid-sagittal plane; the study of the movements in the pharynx is, for example, not possible;

– dynamic MRI relies on numerous repetition of the same sequence to reconstruct the impression of movements in time [12]. Moreover, the alignment of a speech signal with the so-obtained images has to be handled with care, given the fact that the reconstructed sequence does not correspond to any individual production [13].

Real-time MRI allows studying the dynamics of vocal tract deformation in any plan. This permits collection of new data and provides new perspectives to study co-articulation processes, even if the speed of image acquisition is still a bit slow.

Acknowledgements. This research was supported by the Convention ARC ‘Dynamique des systèmes phonologiques’ 98-02, n°226.

References

[1] T. Baer, J.C. Gore, L.C. Gracco, P.W. Nye, Analysis of vocal tract shape and dimensions using magnetic resonance imaging: vowels, J. Acoust. Soc. Am. 90 (2) (1991) 799–828.

[2] M.A. Crary, I.M. Kotzur, J. Gauger, M. Gorham, S. Burton, Dynamic magnetic resonance imaging in the study of vocal tract configuration, J. Voice 10 (4) (1996) 378–388.

[3] B.H. Story, I.R. Titze, E.A. Hoffman, Vocal tract area functions for an adult female speaker based on volumetric imaging, J. Acoust. Soc. Am. 104 (1) (1998) 471–487.

[4] D. Demolin, T. Metens, A. Soquet, Three-dimensional measurements of the vocal tract by MRI, Proc. ICSLP-96, Philadelphia, USA, 1996, pp. 272–275.

[5] J. Van Vaals, G. Van Yperen, A. Hoogenboom, M. Duivestijn, Local Look (LOLO): Zoom Fluoroscopy of a Moving Target, Proc. 1st SMR Meeting, Dallas, 1994, p. 38.

[6] T. Metens, D. Demolin, M. George, V. Lecuit, A. Soquet, H. Raeymaekers, C. Matos, Ultra Fast subsecond Lolo TSE for continuous real-time imaging of articulators movements involved in speech production, ISMR 5th Meeting, 12–18 April 1997, Vancouver BC, Canada, 1997, p. 1832.

[7] M. Vlaardingerbroek, J. den Boer, Magnetic Resonance Imaging, Chapter 3, 2nd edition, Springer Verlag, Berlin, 1996.

[8] P. Branderud, Personal Communication, 2000.

[9] P. Mansfield, J. Phys C 10 (1997) L55.

[10] A.J. Haase, Magn. Res. 67 (1986) 258.

[11] D. Demolin, M. George, V. Lecuit, T. Metens, A. Soquet, H. Raeymaekers, Coarticulation and articulatory compensations studied by dynamic MRI, Proc. Eurospeech 97, Rhodes, Greece, 1997, pp. 31–34.

[12] A.K. Foldvik, U. Kristiansen, J. Kværness, A time-evolving three-dimensional vocal tract model by means of magnetic resonance imaging (MRI), Proc. Eurospeech 93 (1993) 557–558.

[13] C.H. Shadle, M. Mohammad, J.N. Carter, P.J.B. Jackson, Multi-planar dynamic magnetic resonance imaging: new tools for speech research, Proc. ICPhS-99, San Francisco, USA, 1999, 623–626.