Dynamic imaging of the vocal tract using a cine-MRI sequence: Protocol optimization and evaluation

G. Gilbert[1,2], J. Nissenbaum[3], and G. Beaudoin[1]
[1]Department of Radiology, Centre Hospitalier de l'Université de Montréal, Montreal, QC, Canada, [2]MR Clinical Science, Philips Healthcare, Cleveland, OH, United States, [3]Department of Languages, Literatures and Linguistics, Syracuse University, Syracuse, NY, United States

Introduction

Dynamic magnetic resonance imaging is increasingly used as a non-invasive tool to visualize vocal organs in linguistics and phonetics studies [1-4]. In general, two different approaches are used to perform dynamic imaging of the vocal tract, one relying on a cine-MRI sequence [1-2] and the other on a real-time acquisition [3-4]. While the use of a real-time sequence allows for a dynamic series to be acquired over a single repetition of the elocution task, this approach is intrinsically limited in terms of spatial and temporal resolutions in comparison to a cine-MRI sequence, for which imaging is performed over a large number of the same dynamic pattern. However, the cine-MRI method can suffer from synchronization and motion errors that can lead to effective spatial and temporal resolutions below the theoretical values. Accordingly, the aims of this study are to develop an optimized protocol for the dynamic imaging of the vocal tract using a cine-MRI sequence and to evaluate the synchronization accuracy that can be reached using this approach.

Materials and Methods

All experiments were performed on a well-trained healthy volunteer using a clinical 3T system and a 16-channel neurovascular head and neck coil. Dynamic cine-MRI was performed using an unspoiled fast gradient-echo sequence with parameters: TR = 3.46 ms, TE = 1.78 ms, echo train length = 3, α = 5°, SENSE factor = 1.7, NSA = 2, FoV = 320 mm x 320 mm, in-plane resolution = 1 mm x 1 mm. Slice thickness was varied between 6 mm and 26 mm, with an increment of 4 mm, in order to evaluate the best compromise between signal-to-noise ratio (SNR) and through-slice blurring. A single mid-sagittal slice was acquired.

Triggering of the dynamic sequence was performed using the internal physiology simulation tool of the MR system. A simulated RR interval of 4 seconds was used, with imaging performed over the first half of the interval, while the second half acted as a rest period. A total of 192 dynamic phases were acquired over the two-second imaging period, for a theoretical temporal resolution of 10.4 ms. Total acquisition time (128 shots/speech tasks) was 8:32. Synchronization between the imaging sequence and the speech task was performed through an external TTL signal that was sent by the MR system at the beginning of each simulated R wave. This TTL signal was used to trigger an auditory stimulus, which the subject had to repeat aloud.

Figure 1: Representative frames from a dynamic series.

During each two-second acquisition period, the subject pronounced six syllables (two repetitions of [mid meid mæd]). Wide band spectrograms of the speech output were used to identify important phonetic landmarks. The onsets and offsets of each of the six vowels provided 12 landmarks, which were easily identifiable as they defined temporal regions of periodic energy characteristic of vowels. The temporal location of each landmark was then compared with the start of the two-second scan period (also easily identifiable in the spectrogram). The intervals between each scan onset and the corresponding phonetic landmarks provided the basis for gauging the synchronization accuracy.

Results and Discussion

Fig. 1 presents three representative frames from a dynamic series, for a slice thickness of 18 mm. It can be assessed that despite the variations that occurred between the speech synchronization and the MR signal acquisition (described below), the methodology yielded highly informative images in which it is possible not only to see the major speech articulators but also their dynamic interaction during speech production. Many of the anatomical structures involved in speech are visible in the mid-sagittal view: the lips, the tongue, the velum, and several structures of the larynx: the vocal folds, as well as the cricoid and arytenoid cartilages.

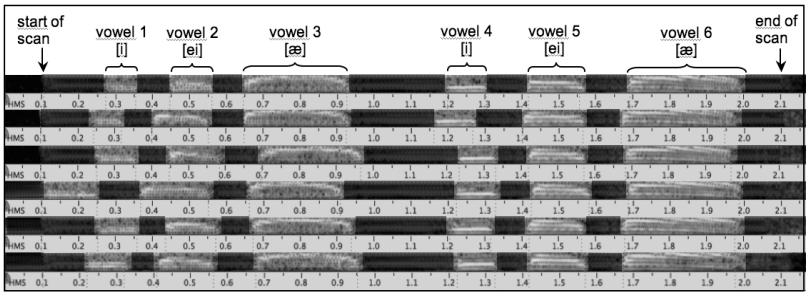

Figure 2: Six representative audio signals showing temporal alignment of repetitions.

A visual comparison of dynamic cine-MRI images acquired with different slice thicknesses highlighted that there was visually limited improvement in the apparent noise level for slice thicknesses of 18 mm and more. Accordingly, a slice thickness of 18 mm was deemed optimal, providing a high SNR, while still preventing too much blurring of the significant structures of the vocal tract.

Tab. 1 shows the average values, for a single cine-MRI acquisition, of the temporal intervals between the beginning of each scanning period and the onset or offset of each of the six vowels. This table reveals two relevant results. First, the intervals (in particular of the first two syllables) are characterized by a fairly high degree of variation. For instance, the standard deviation of the onset of syllable one is 94 ms. Second, by the offset of the third vowel, the variability is markedly lower, with standard deviations ranging between 23 and 35 ms. These basic characteristics were evident for the other sessions that were analyzed as well. Fig. 2 shows the spectrograms for six representative scans. The figure makes clear that, while the subject began the task with fairly poor synchronization, it became more accurate during the course of the speech task.

Conclusion

The results presented in this abstract illustrate that dynamic imaging of the vocal tract using a cine-MRI sequence can achieve a high spatial resolution along with a relatively high temporal fidelity (~ 30 ms). The high variability in synchronization for the first vowels may be a result of the two second rest time in between scans, during which there is correspondingly no auditory stimulus. This pattern suggests that in future studies it may prove beneficial to start with an auditory stimulus and not to have the subject begin the speech task until several hundreds of milliseconds into the scan.

	Vowel 1 [i]		Vowel 2 [ei]		Vowel 3 [ae]		Vowel 4 [i]		Vowel 5 [ei]		Vowel 6 [ae]	
	On	Off	On	Off	On	Off	On	Off	On	Off	On	Off
Mean Int. (ms)	107	220	311	451	554	823	1093	1185	1300	1449	1572	1851
Std. Dev. (ms)	94	61	44	26	25	23	30	31	28	27	26	35

Table 1: Average temporal intervals (in ms) between start of scan and phonetic landmarks.

References

[1] Parthasarathy V et al. *J. Acoust. Soc. Am*. 2007; 121: 491-504.
[2] Takemoto H et al. *J. Acoust Soc. Am*. 2006; 119: 1037-1049.
[3] Kim Y-C et al. *Magn. Res. Med*. 2009; 61 : 1434-1440.
[4] Byrd D et al. *J. of Phonetics* 2009; 37 : 97-11.