# Sound synchronization and motion compensated reconstruction for speech Cine MRI.

Pierre-André Vuissoz[1,2], Freddy Odille[1,2], Yves Laprie[3,4], Emmanuel Vincent[3,5], and Jacques Felblinger[6,7]
*[1]Imagerie Adaptative Diagnostique et Interventionnelle, Université de Lorraine, Nancy, France, [2]U947, INSERM, Nancy, France, [3]LORIA, Université de Lorraine, Nancy, France, [4]LORIA, CNRS, Nancy, France, [5]LORIA, INRIA, Nancy, France, [6]University Hospital Nancy, Nancy, France, [7]CIC-IT 1433, INSERM, Nancy, France*

**TARGET AUDIENCE:** MR scientists and physicians interested in dynamic high resolution images of speech.

**PURPOSE:** Dynamic imaging of the vocal tact is important for modeling speech through the acoustic-articulatory relation. The average duration of each sound is about 80ms. Movements of each articulator, in particular the tongue, should be captured with sufficient precision. Current clinical techniques use X-ray video fluoroscopy which involves ionizing radiation. Real-time MRI allows direct recording of speech motion [1] but is intrinsically limited in terms of resolution and SNR. Synchronization of MRI with an acoustic device is possible [2] but requires motion of vocal system to be highly reproducible. In this work we propose an optimized setup for achieving dynamic MRI of speech with high spatial and temporal resolution based on a combination of: an MR-compatible acoustic device allowing simultaneous recording of speech during MRI; and a retrospectively gated, motion-compensated image reconstruction that can deal with the variability of the subject repeating the same sentence over the acquisition.

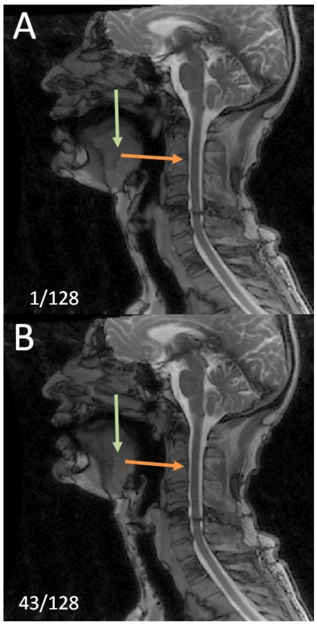

Fig 1. : Two temporal position (A) 1, (B) 43 of the 128 images cine loop of sagittal slice along the vocal tract with TM

**METHODS:** Data acquisition: Dynamic MRI data were acquired at 3T (Signa HDxt, GE Healthcare, Milwaukee, WI) using an ungated balanced SSFP sequence (one sagittal slice, 256x256 matrix, TR/TE = 3.9/1.7 ms, 5 mm slice thickness, 45° flip angle, FOV 30 cm, 65 temporal phases). A list of 10 sentences was carefully selected for the dynamic imaging protocol to be short and yet yield a good coverage of the tongue movements in French language [3]. For each dynamic acquisition the subject was asked to repeat one of the 10 sentences until the sequence stopped. Acoustic signals were acquired using an optical microphone (FOMRI III, Optoacoustics, Yehuda Israel). The scanner's acquisition window signal was also recorded with the device to allow synchronization of MR events and acoustic signals.

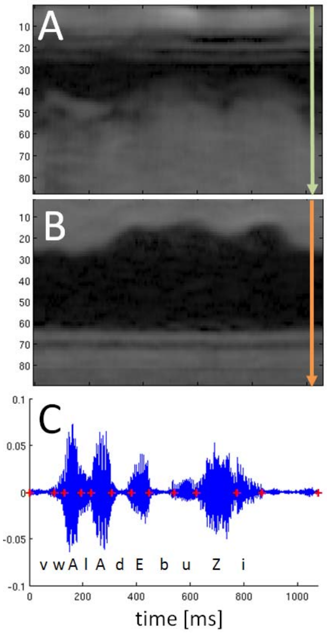

Fig 2. : Time Motion display of the cine loop for (A) hard palate to tongue dorsum and (B) tongue back to pharyngeal wall, (C) sound record used for the motion compensated reconstruction.

Acoustic signal processing: Microphone recordings were first denoised [4] to eliminate gradient noise. Then they were phonetically segmented by hand in order to annotate the beginning of each phone within the sound record. An acoustic phase signal was then created in order to indicate the temporal position within the sentence. In order to account for the variability of repeated phone timings, the phase signal was adjusted using a piecewise linear scaling based on the manual segmentation.

Reconstruction: Cine images from each template sentence were reconstructed using cine-GRICS [5]. Here cine-GRICS used the acoustic phase signal to reconstruct the images by a motion-compensated slice-window approach. The window width was 80 ms and motion correlated signals were the relative phase distance to the key frame position and the squared distance. 128 key frames were used.

**RESULTS:** In Figure 1 two images of the subject pronouncing "Voilà des bougies" show the absence of motion blurring or artifact in the reconstructed cine loop. In Figure 2 the distances between tongue dorsum and hard palate on the one hand, and between the tongue back and the pharyngeal wall on the other hand are shown, along with the template sentence signal.

**DISCUSSION**: A potential issue with balanced-SSFP is related to the banding artifacts and the strong $B_0$ gradient at the air tissue interface especially in the tongue, but the signal void in the vocal tract is still clearly distinguished in the dynamic movie.

**CONCLUSION:** Each cine loop enables the delineation of the vocal tract with sufficient spatial and temporal resolution enabling the acquisition of a personalized speech model within an MR examination of half an hour.

**REFERENCES:** [1] Narayanan et al. J. Acoust. Soc. Am. 115(4):1771 (2004). [2] Frauenrath et al. Act. Acus. 94(1) 148 (2008). [3] Maeda S. Actes X JEP p152, Grenoble, Mai 1979. [4] Ozerov et al. IEEE TASLP 20(4) :1118 (2012). [5] Vuissoz et al. JMRI 35 :340 (2012).