Towards an articulatory tongue model using 3D EMA

Ingmar Steiner[1,2], Slim Ouni[1,3]
[1]LORIA Speech Group, [2]INRIA, [3]Université Nancy 2
Firstname.Lastname@loria.fr

Overview: Within the framework of an acoustic-visual (AV) speech synthesizer [1], our aim is to integrate a tongue model for improved realism and visual intelligibility. The AV text-to-speech (TTS) synthesizer uses a bi-modal corpus of speech and high-resolution data captured from a real speaker.

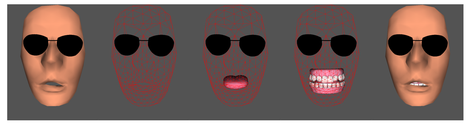
In this paper, we describe a geometric tongue model that is both simple and flexible, and which is controlled by 3D electromagnetic articulography (EMA) data through an animation interface, providing fairly realistic tongue movements and maintaining the overall design that the AV TTS is driven by speech data.

The tongue model's mesh is deformed using a skeletal animation approach; the skeletal armature is in turn controlled by mapping the positional and rotational information from the EMA coils.

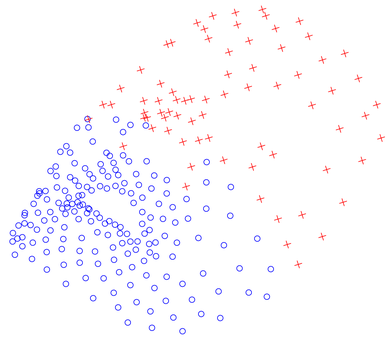
[1] A. Toutios, U. Musti, S. Ouni, V. Colotte, B. Wrobel-Dautcourt, and M.-O. Berger. Towards a true acoustic-visual speech synthesis. In *Proc. 9th International Conference on Auditory-Visual Speech Processing*, 2010.

Acoustic-visual data: setup for AV data acquisition; facial marker points are visible in camera array view on foreground display

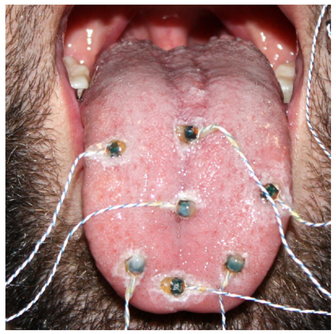
Simultaneous acoustic recordings are processed for waveform concatenation in unit-selection TTS



Skinned mesh and wireframe views of talking head with placeholder tongue and teeth



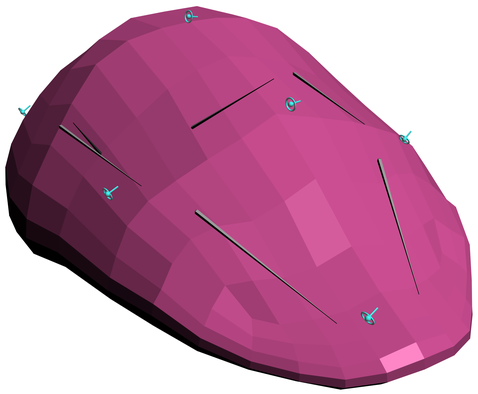
3D projection of marker points used for vertices of facial mesh (blue circles represent vertices relevant to speech animated during TTS)



EMA data: High temporal resolution, but *sparse* representation of the tongue surface

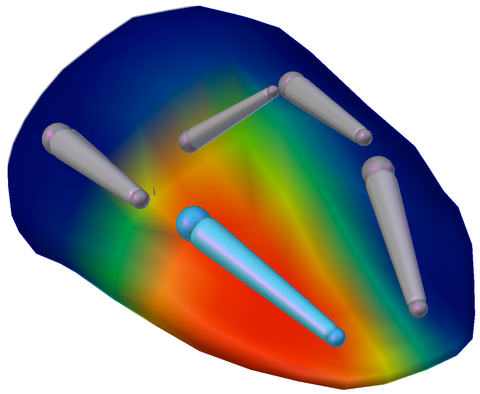
EMA coils are too few in number to be directly usable as vertex positions in tongue mesh construction, so an independent tongue model is needed

Movements and orientation of EMA coils can be mapped to control parameters of a geometric tongue model, but surface positions are insufficient for realistic animation



Tongue model: 3D rendering of tongue mesh, superimposed with the embedded controlling armature (shown as six gray staves)

EMA coils (different layout than in previous box) rendered as blue shapes, showing their orientation; these are mapped to the armature components to control tongue mesh deformation



Heat map visualizing the influence of one armature component (hightlighted in blue) over mesh vertices during deformation; the underlying vertex weights are automatically assigned

Translation and rotation of the armature's components are correspondingly applied to the mesh during animation

Two example poses of the armature deforming the tongue mesh into a bunched (left) and retroflex (right) configuration

During rendering and animation, the armature is hidden, only the deformed tongue mesh is visible

Advanced features such as volume preservation, soft body dynamics, collision detection, etc. are available, but not yet implemented

Conclusion and future work: the resulting tongue model can be animated in real time using EMA trajectories, providing realistic kinematics.

This tongue model can be integrated into the AV TTS synthesizer using a suitable 3D engine

Synthesis of new utterances is possible either by

- generating the trajectories at synthesis time from a corpus of EMA data (directly or via statistical models); or
- storing EMA trajectories with the acoustic-visual unit data, using an offline inversion process

Once integration is complete, an evaluation study can be carried to assess the overall performance and contribution of the tongue model