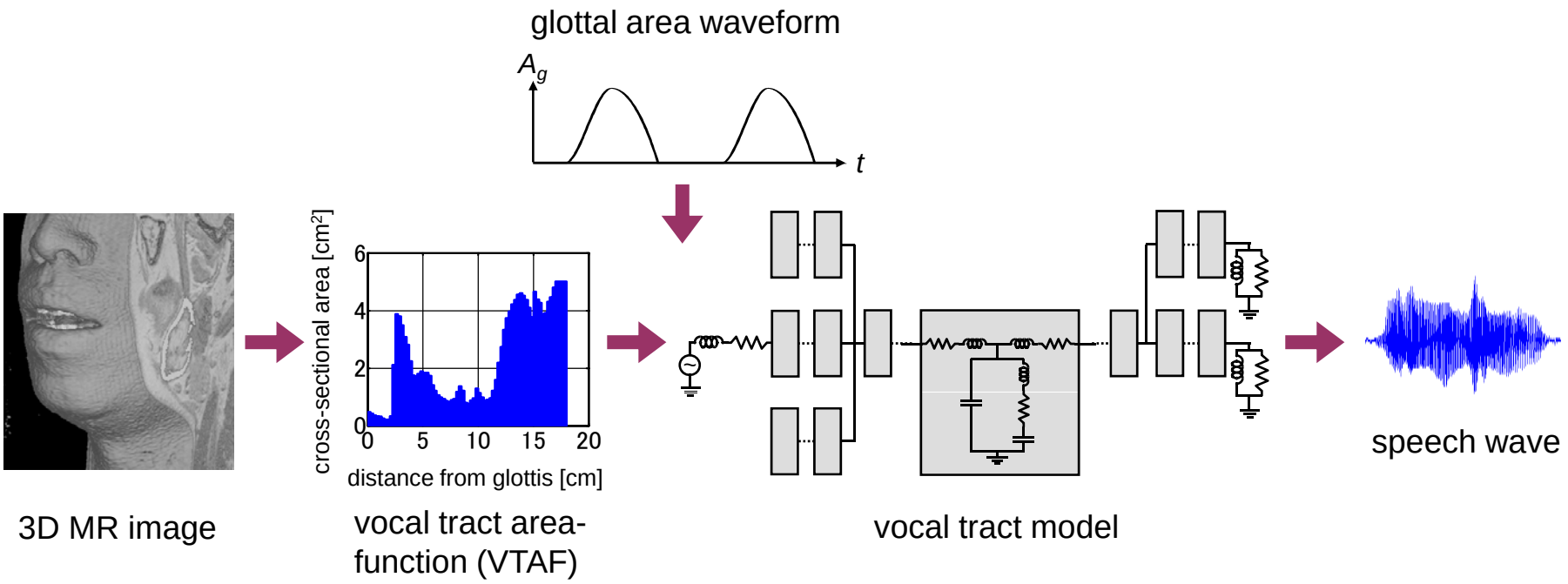
Objective

- MRI-based articulatory synthesis system.
  - articulatory model: modified Maeda's (1982) time-domain synthesizer.



schematic diagram of proposed synthesis system.

Background

- There is increased demand for controlling voice quality of synthetic speech.
- Maeda's synthesizer (1982)
  - simulates acoustic wave propagation in the vocal tract in the time domain.
  - input: VTAFs and glottal area waveform.
  - allows to control voice quality of synthetic speech by varying VTAFs and glottal parameters.
- Recent improvement in MR image quality
  - allows to obtain precise VTAFs.
  - Synchronized sampling method (Masaki *et al*., 1999).
  - Usage of a bone-conduction speaker (Nota *et al*., in press).

Aim

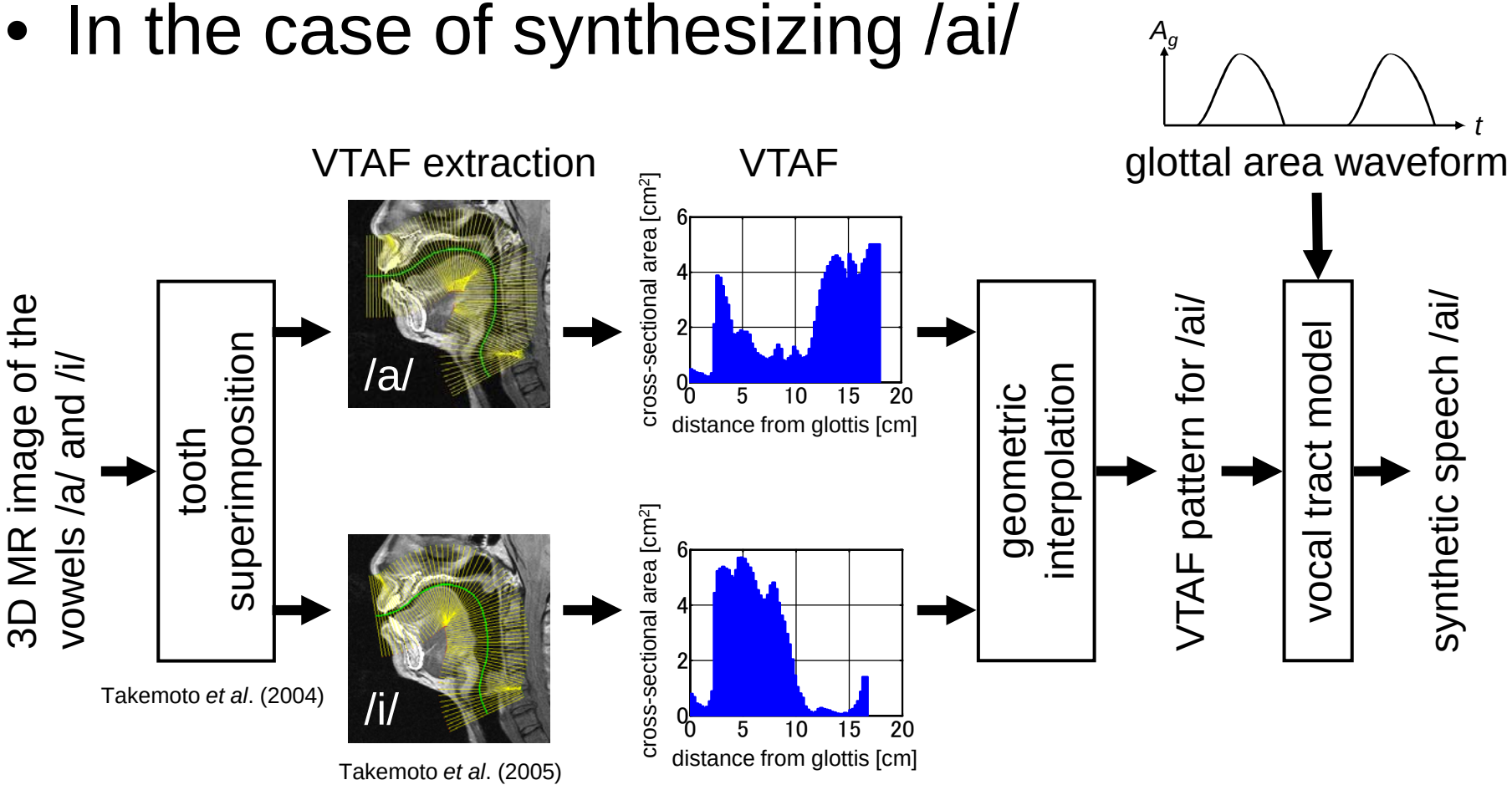
- Implementation of an MRI-based Maeda's synthesizer.

- Advantages of articulatory synthesis
  - Small data size.
  - Smooth concatenation of phonemes.
  - Voice quality can be controlled by changing the vocal tract shape.
  - The synthesizer is definitely useful for speech science research.

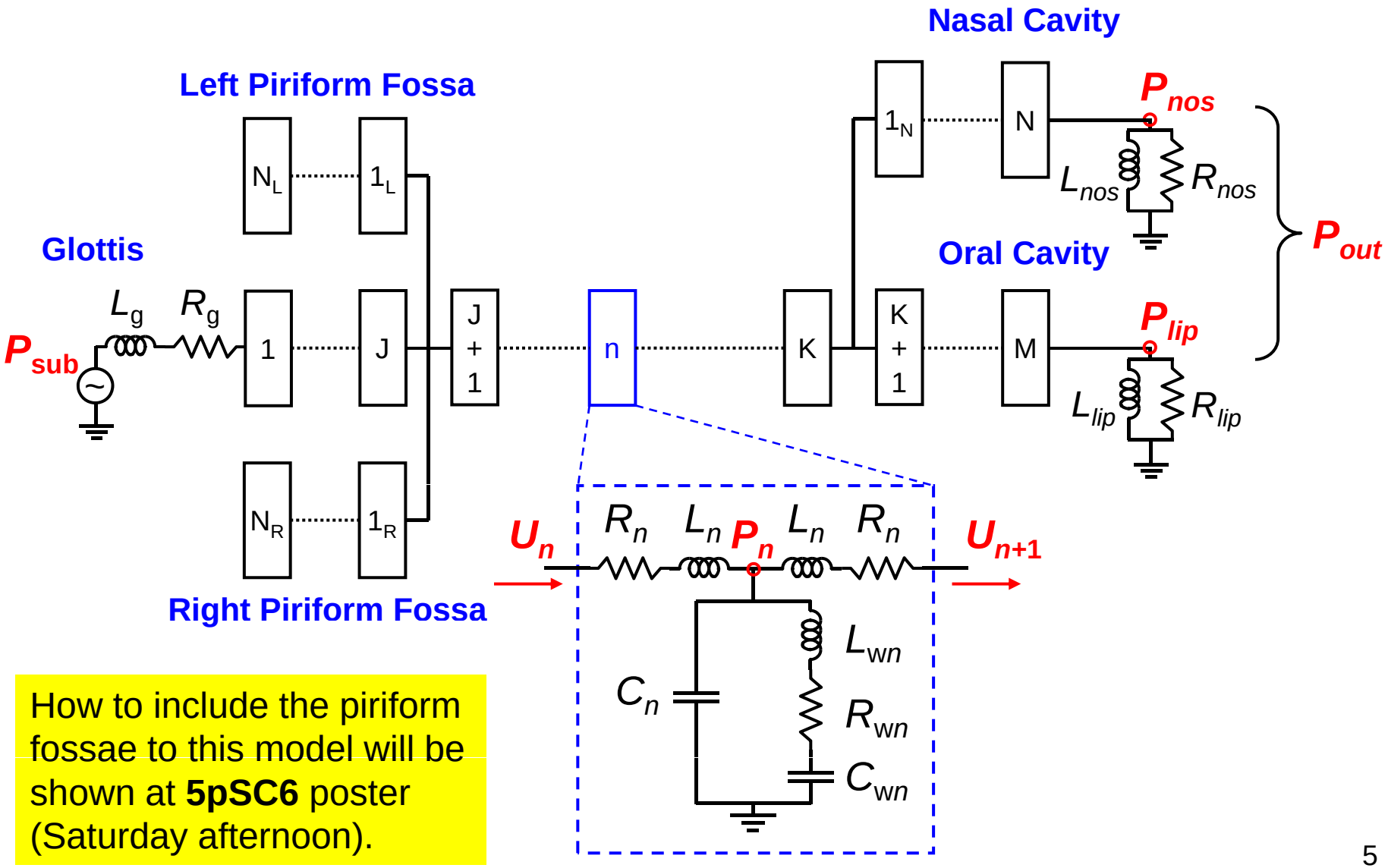
System Flowchart

Vocal Tract Areafunction
→ VTAF

- In the case of synthesizing /ai/



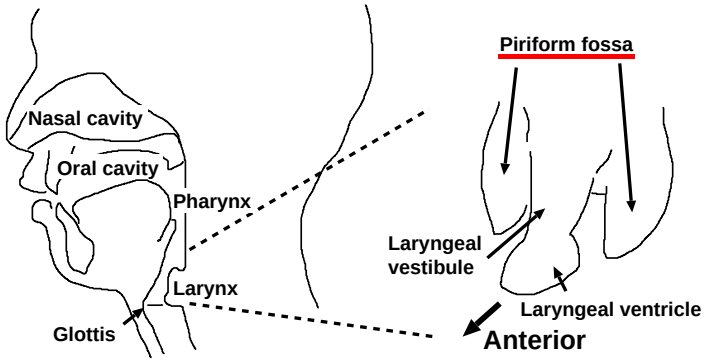
Vocal Tract Model (see also 5pSC6)

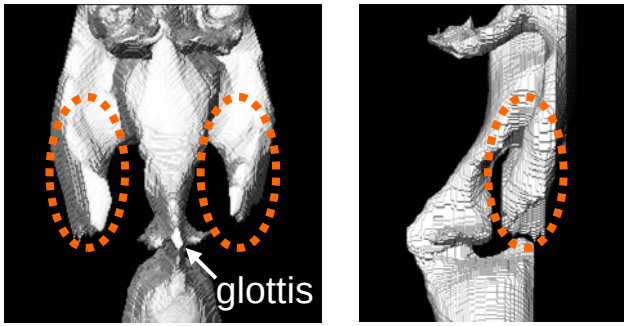


How to include the piriform fossae to this model will be shown at **5pSC6** poster (Saturday afternoon).

Piriform Fossa

- The piriform fossae are a pair of bilateral cavities located behind the laryngeal tube.
- Dang & Honda (1997)
  - The fossae cause anti-resonances in speech spectra at the frequency region from 4 to 5 kHz.
  - The area functions of the fossae are different among speakers.





front view lateral view

Takemoto *et al*. (2006)

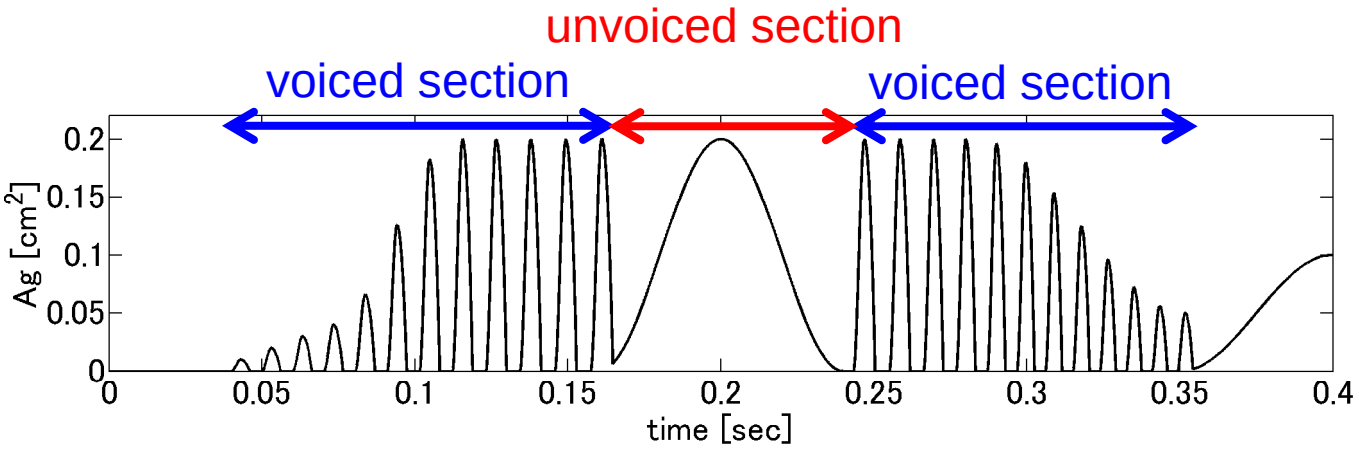
Glottal Wave

- Voiced section
  - Rosenberg wave

$$f(t)=\begin{cases} a\left\{3\left(\frac{t}{\tau_1}\right)^2-2\left(\frac{t}{\tau_1}\right)^3\right\} & 0\le t\le \tau_1 \\ a\left\{1-\left(\frac{t-\tau_1}{\tau_2}\right)^2\right\} & \tau_1\le t\le \tau_1+\tau_2 \end{cases}$$

- Unvoiced section
  - The glottis opens.

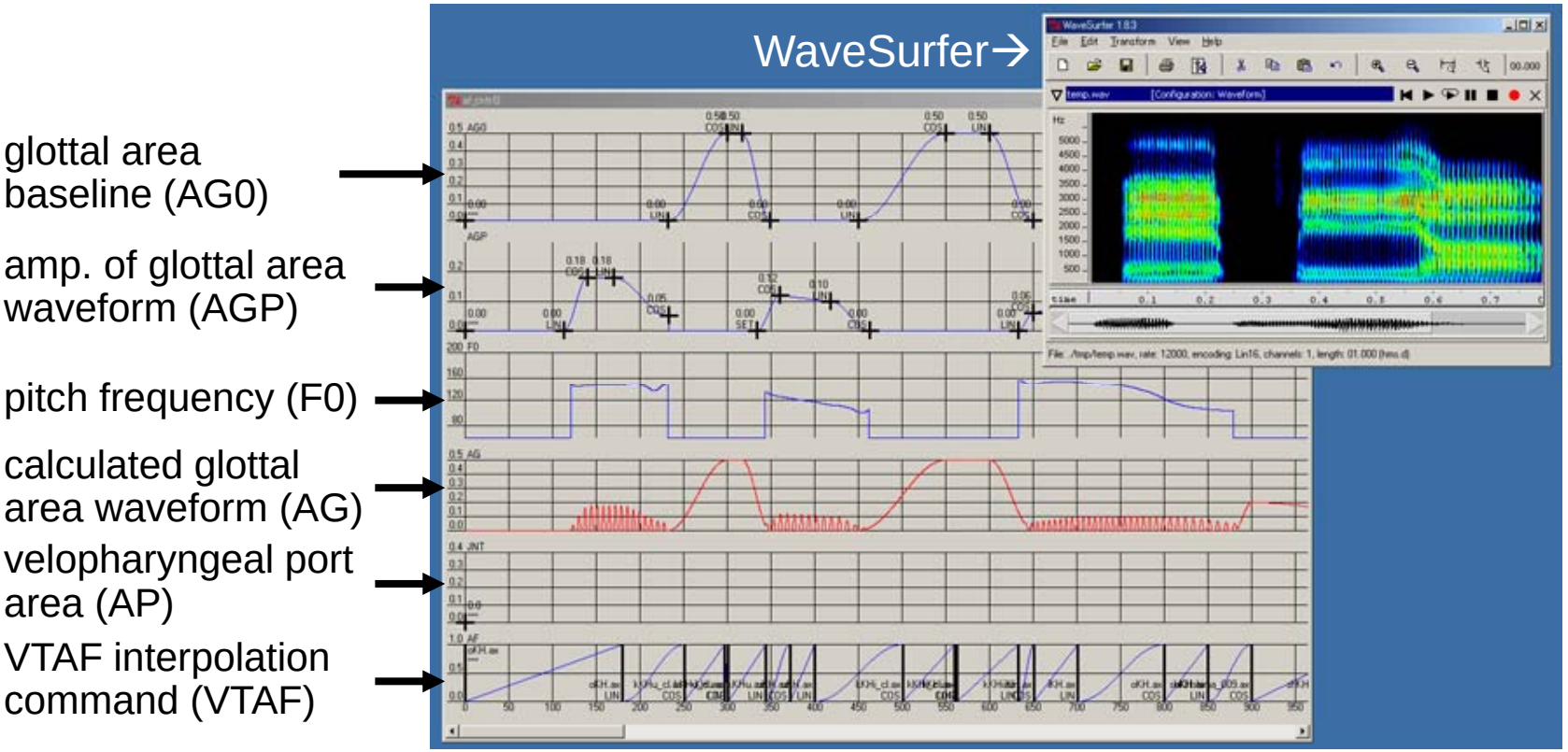
➡ A dipole pressure noise source is inserted at the constriction of VTAF.



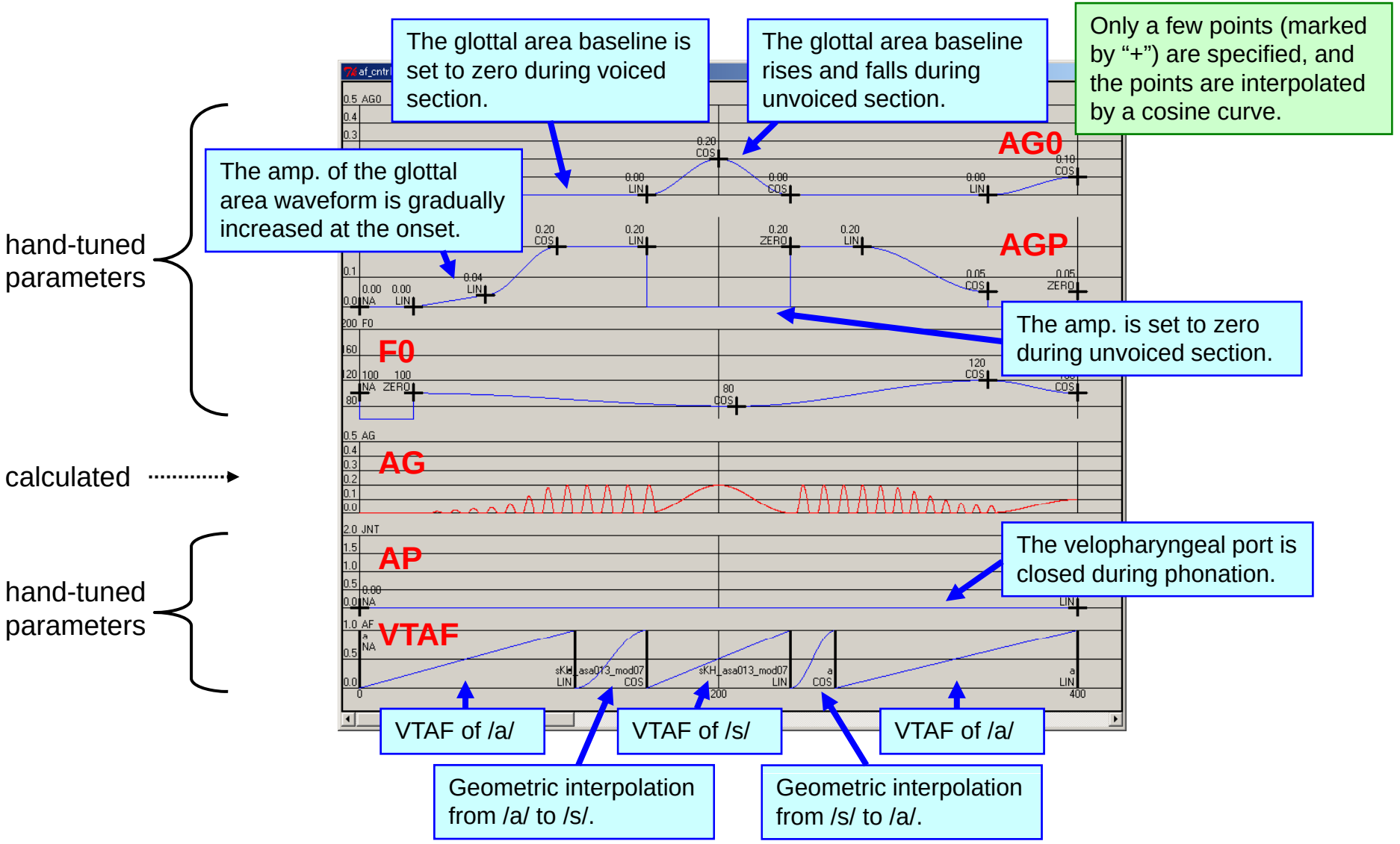
glottal area waveform for /asa/.

GUI Interface

- To control synthesis parameters and timing.



Example of Parameters for */asa/*



Data Size

- Vocal tract area-functions (VTAFs)
  - Basically one VTAF per phoneme.
  - One VTAF
    - Main vocal tract: 44 sections
    - Piriform fossa: 1 section x 2 (left and right)
    - Nasal cavity: 38 sections
    - One section is specified by its cross-sectional area and length.
- Glottal area waveform
  - The number of control points is on the same order as the number of phonemes.

Altering Voice Quality

- By varying VTAFs
  - Based on analysis-by-synthesis method.
  - No MR image of the target speaker is needed.
- By varying glottal waveform
  - Changing the parameters of the glottal waveform model.
  - Lowpass filtering the glottal waveform (Klatt and Klatt, 1990).
  - Superimposition of noise on the glottal waveform (Klatt and Klatt, 1990).

Conclusion

- An MRI-based articulatory speech synthesis system based on Maeda's model (1982) was proposed.
- The synthesizer can
  - produce speech from a small data set.
  - concatenate phonemes smoothly.
  - alter voice quality by changing VTAFs and glottal parameters.
- The parameters are set by hand so far.
  - Techniques used in text-to-speech can be applied to set the parameters automatically.

This work is supported in part by NICT, SCOPE-R (No. 042107002), and MEXT Grant-in-And for Scientific Research (No. 16700199) of Japan.