Synthesis of running speech for studying the mechanisms of speech production : the case of fricatives

Benjamin Elie and Yves Laprie

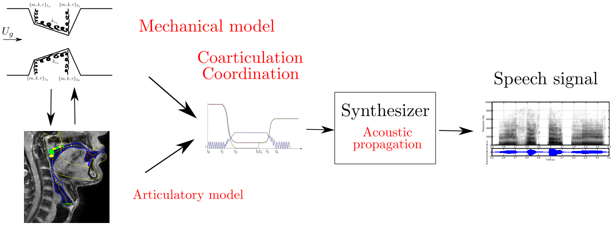
LORIA, INRIA/CNRS, Nancy
https://members.loria.fr/BElie/

December, 2017

Speech synthesis (utterances), **complete** and **realistic**, based on purely acoustical model

Example of an articulatory synthesizer

Phonatory source



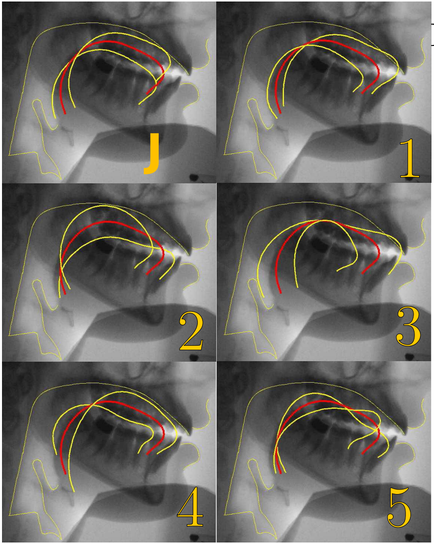
- Realistic acoustics
- Articulatory control

Vocal tract deformation

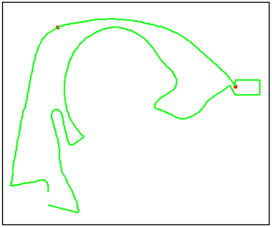
Applications: Medicine, audiovisual, language learning, text-to-speech...

Plan

Tongue modes

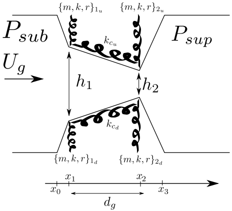


First mandible mode
and
5 first tongue modes



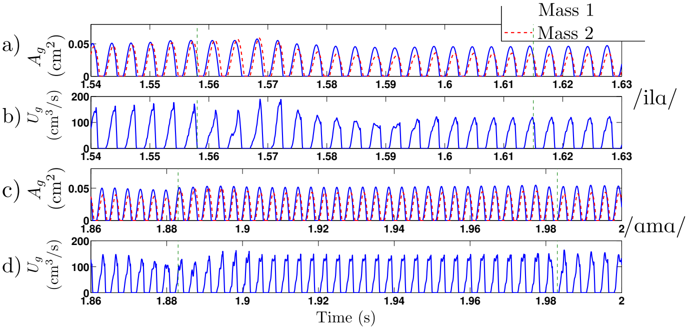
Complete model

Self-oscillating model of the vocal folds



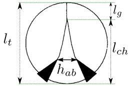
Lous *et al.* (1998)

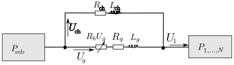
$$\mathbf{M\ddot{y}} + \mathbf{R\dot{y}} + \mathbf{Ky} = f(P_{sup}, P_{sub}, \theta_{geom})$$



Modified glottis model

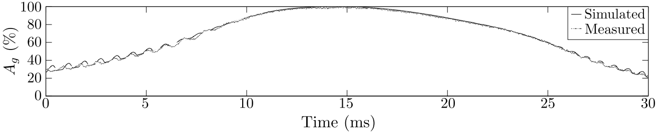
Partial glottal closure

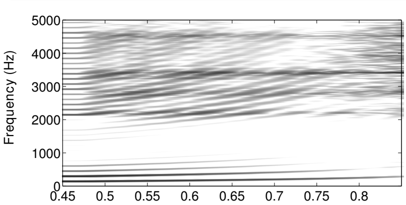


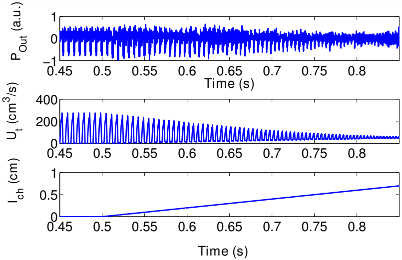


$$P_{sub} - P_1 = \Delta P_{close} + \frac{\partial}{\partial t}\left(L_1 U_{ch} + R_1 U_{ch}\right)$$

cf. Birkholz *et al.*, Interspeech 2011 or Elie and Laprie, ICASSP 2016

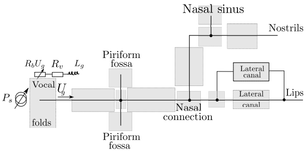






Waveguide network paradigm for speech synthesis

Modeling the vocal tract as a waveguide network[1]



$$\begin{bmatrix} \mathbf{f}^{(1)} \\ \mathbf{f}^{(2)} \\ \vdots \\ \mathbf{f}^{(\mathcal{N})} \end{bmatrix} = \begin{bmatrix} \mathbf{Z}^{(1)} & \mathbf{C}^T_{(1,2)} & \cdots & \mathbf{C}^T_{(1,\mathcal{N})} \\ \mathbf{C}_{(1,2)} & \mathbf{Z}^{(2)} & & \\ \vdots & & \ddots & \\ \mathbf{C}_{(1,\mathcal{N})} & & & \mathbf{Z}^{(\mathcal{N})} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{u}^{(1)} \\ \mathbf{u}^{(2)} \\ \vdots \\ \mathbf{u}^{(\mathcal{N})} \end{bmatrix}$$

Frication noise generation

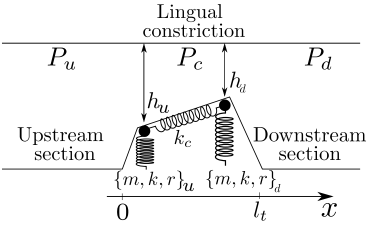
Pressure source is activated when the Reynolds number Re is above the threshold Re_c:

$$P_{n_i} = \max \left\{ 0, \xi w \left(Re^2 - Re_c^2 \right) \frac{U_{DC}^3}{a_{i-1}^{3/2}} \right\}, \quad Re \propto \frac{U_{DC}}{a_c}$$

[1] Elie and Laprie, *Speech Comm.*, 2016

Other oscillator: the alveolar trill

Self-oscillation model of the tongue tip[2]



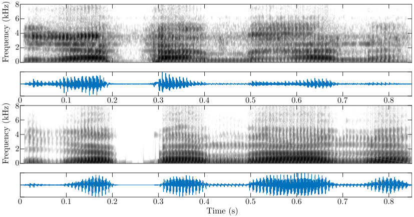
Alveolar trills

- Two-mass model, similar to the VF
- Included in the waveguide network, can be used with realistic VT geometries
- Possibility to consider the incomplete occlusion during contacts

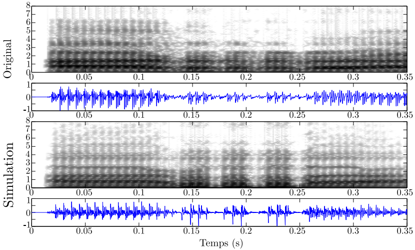
[2] Elie and Laprie, JASA, nov. 2017

A few examples

Il a pas mal (/i.la.pa.ma.lə/)



"ara" (/ara/)

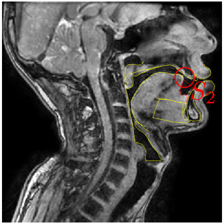


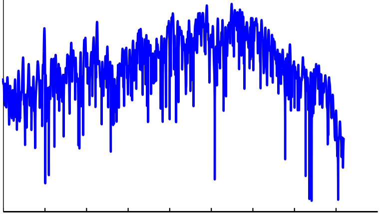
- Reproduction of acoustic features
- Access to quantities not accessible experimentally
- Control of the input articulatory/phonatory parameters

Plan

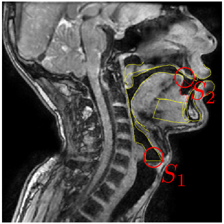
Different sources

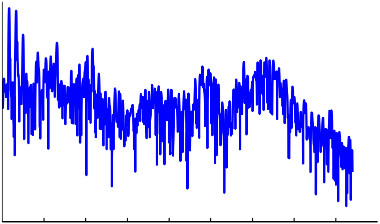
Voiceless Fricative



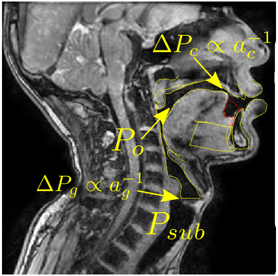


Voiced Fricative





Condition of noise source generation



$$P_{sub} \simeq \Delta P_g + \Delta P_c$$

$$P_o \simeq P_{sub} - \Delta P_g$$

At the glottal level

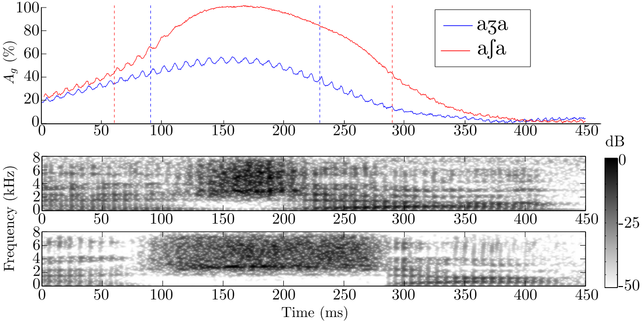
- sufficiently high airflow $\rightarrow$ the glottis should be open
- if voiced fricatives, glottis not totally abducted

At the supraglottal level

- narrow constriction
- high ΔP_c $\rightarrow$ high P_o $\rightarrow$ open glottis

Continuous coordination

Glottal opening in a VFV sequence

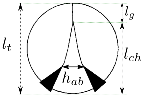


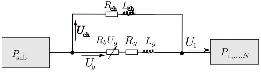
- a low-frequency component (partial abduction of the glottis)
- a high-frequency component (oscillation of the vocal folds)

→ What is the acoustic impact of the partial abduction of the vocal folds ?

Acoustic model of fricative production

Incomplete closure of the glottis

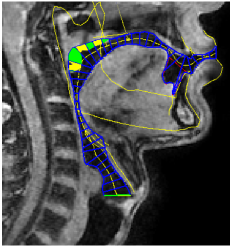


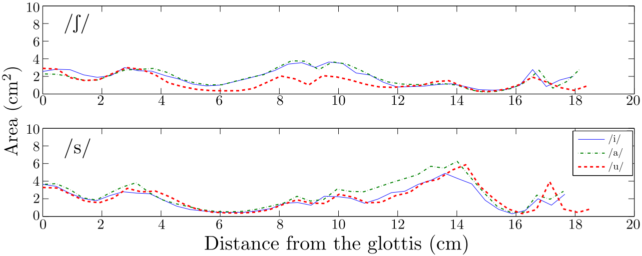


$$P_{sub} - P_1 = \Delta P_{close} + \frac{\partial}{\partial t}\left(L_1 U_{ch} + R_1 U_{ch}\right)$$

cf. Birkholz *et al.*, Interspeech 2011 or Elie and Laprie, ICASSP 2016

A set of area functions extracted from static MRI





→ Simulation of fricatives for different degrees of glottal abduction D_{ab}

Acoustic features

Voicing quotient (VQ)

Quantify the amount of voicing

$$VQ = \frac{\text{Energy of the periodic component}}{\text{Energy of the mix signal}}$$

$VQ = 0 \rightarrow$ voiceless signal, $VQ = 100\% \rightarrow$ purely voiced signal

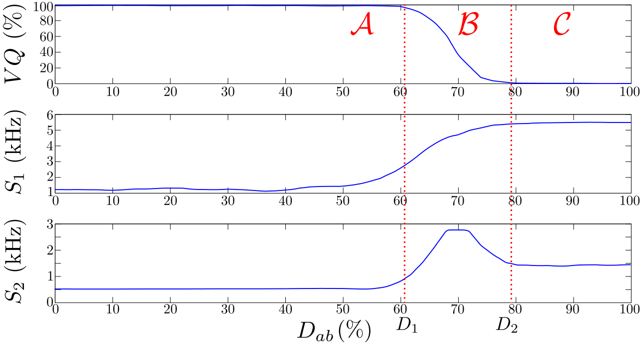
Spectral centroid (S_1)

Balance between low and high frequency components
low $S_1 \rightarrow$ mainly low frequency, high $S_2 \rightarrow$ mainly high frequency

Spectral spread (S_2)

Variance of the spectral distribution
low $S_2 \rightarrow$ narrow band spectrum, high $S_2 \rightarrow$ broad band spectrum

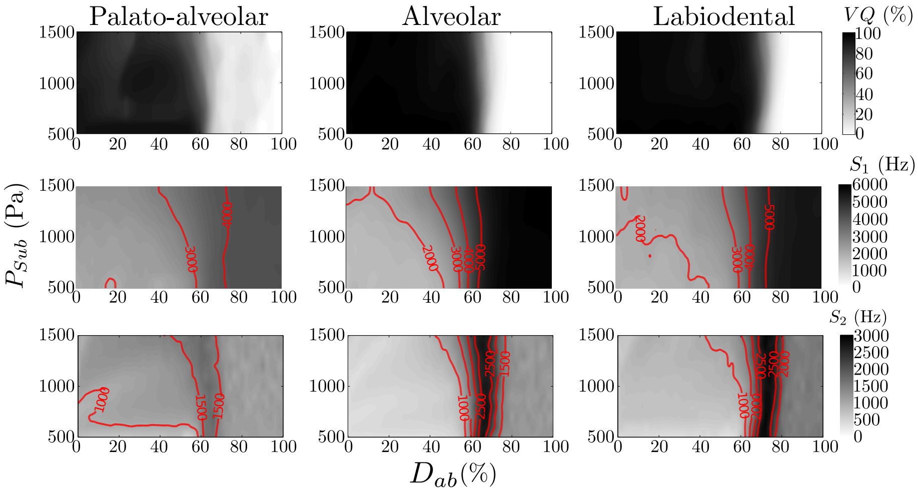
Typical examples



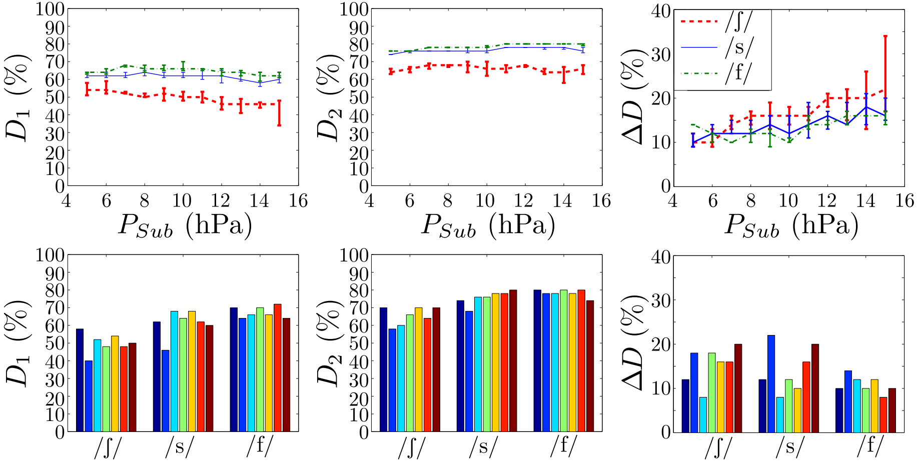
3 regimes of production:

- $\mathcal{A}$ ($D_{ab} < D_1$): low frication noise
- $\mathcal{B}$ ($D_1 < D_{ab} < D_2$): frication noise and voice have similar energy
- $\mathcal{C}$ ($D_{ab} > D_2$): voiceless signal

Acoustic features as a function of P_{sub} (vowel context: /a/)

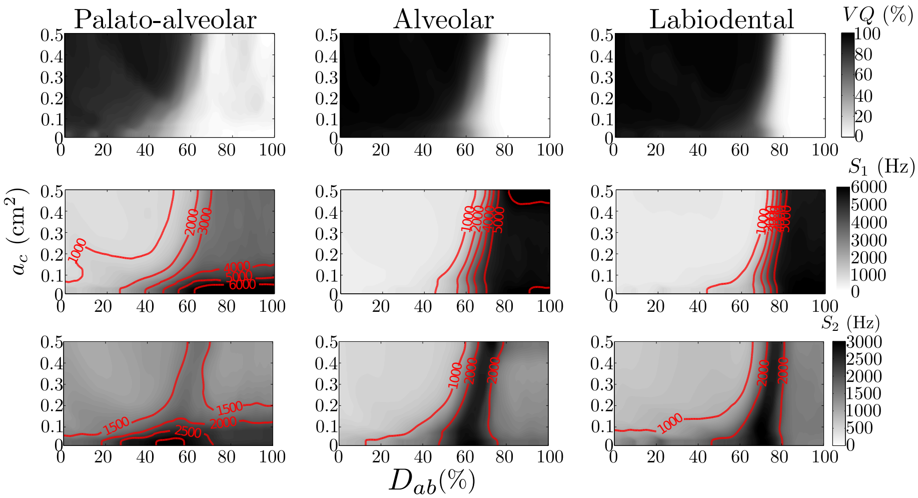


Minimal lengths as a function of P_{sub}

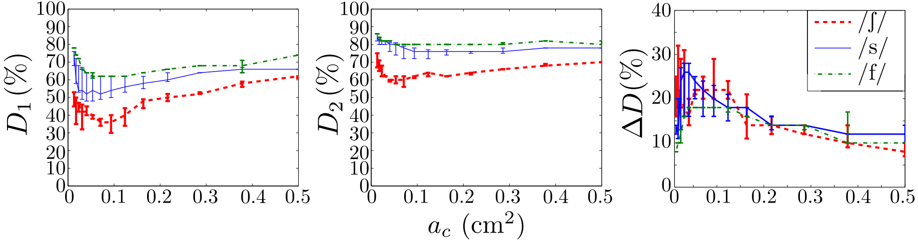


- P_{sub} modifies D_1 and D_2: $D \searrow$ when $P_{sub} \nearrow$
- P_{sub} modifies ΔD: $\Delta D \nearrow$ when $P_{sub} \nearrow$

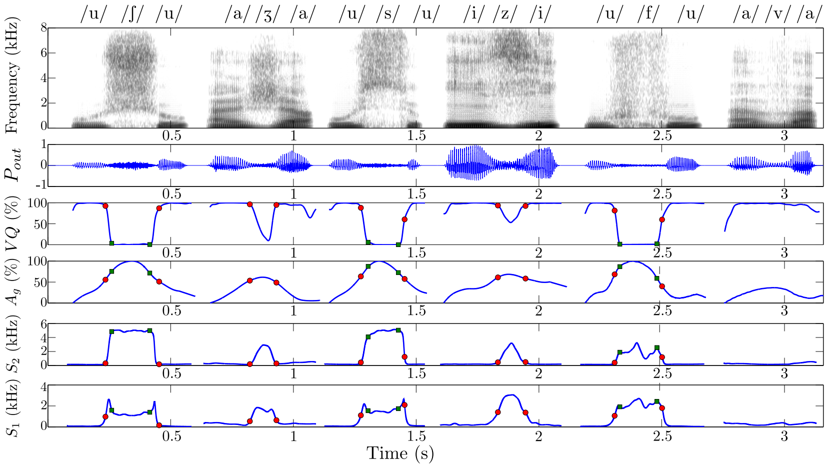
Acoustic features as a function of a_c (vowel context: /a/)



Minimal lengths as a function of a_c



Experiments confirm the observations



Possible strategies for fricative production: hypothesis

Voiceless fricatives

- $\mathcal{A} \rightarrow \mathcal{B} \rightarrow$ sustained $\mathcal{C} \rightarrow \mathcal{B} \rightarrow \mathcal{A}$: easy ($\mathcal{C}$ is stable)

$\rightarrow$ voiceless fricatives are longer to maximize the ratio $\mathcal{C}/\mathcal{B}$

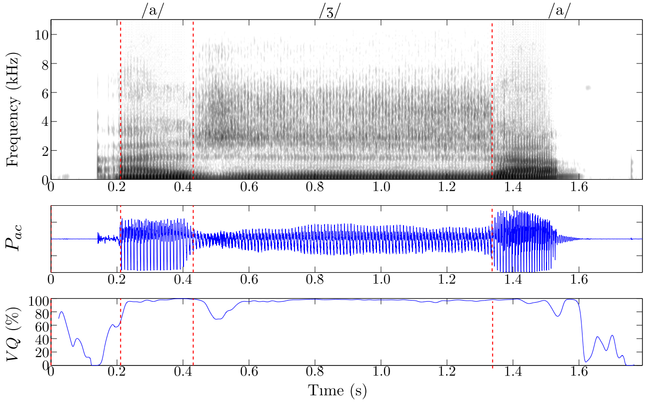
Voiced fricatives

- $\mathcal{A} \rightarrow$ sustained $\mathcal{B} \rightarrow \mathcal{A}$: risky ($\mathcal{B}$ too unstable)
- $\mathcal{A} \rightarrow \mathcal{A}/\mathcal{B}$ boundary $\rightarrow \mathcal{A}$: favors voicing
- Very short $\mathcal{A} \rightarrow \mathcal{B} \rightarrow \mathcal{A}$ or $\mathcal{A} \rightarrow \mathcal{B} \rightarrow \mathcal{C} \rightarrow \mathcal{B} \rightarrow \mathcal{A}$ sequence: maximize proportion of $\mathcal{B}$ over the fricative segment

$\rightarrow$ voiced fricatives are shorter to avoid instability

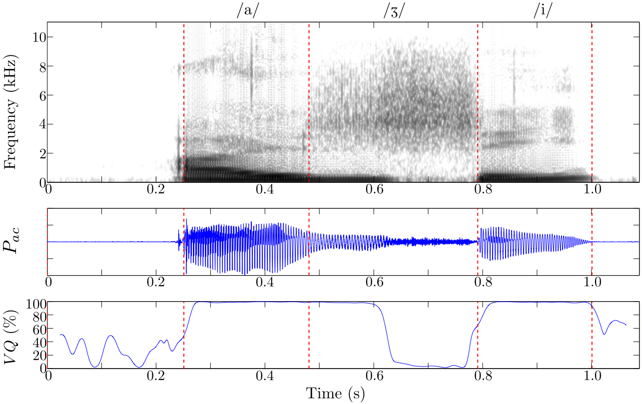
What if voiced fricatives are exaggeratedly longer ?

Speakers usually prefer sustaining regime $\mathcal{A}$ for longer fricatives (Ex. 1)



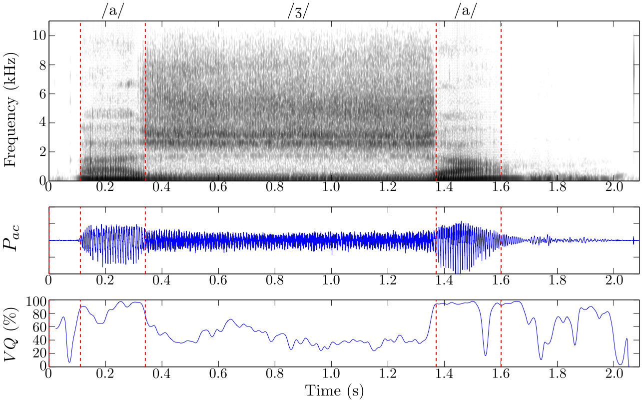
What if voiced fricatives are exaggeratedly longer ?

There may be some "devoicing" incidents (Ex. 2)



What if voiced fricatives are exaggeratedly longer ?

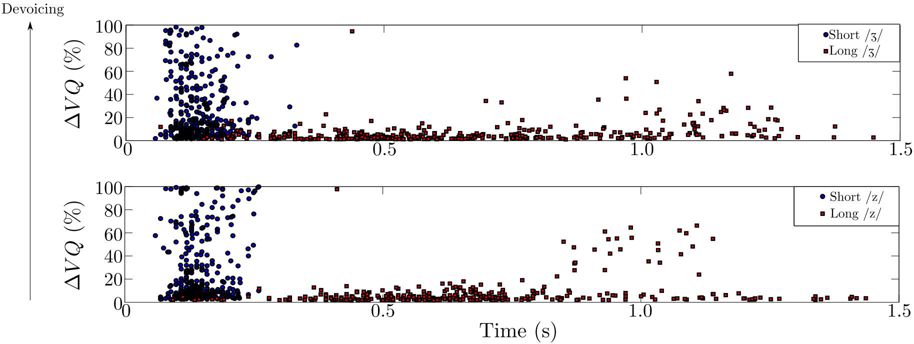
But a (very) few speakers sustains $\mathcal{B}$! (Ex. 3, study in progress)



First results

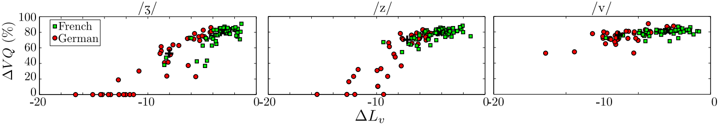
Corpus of 15 speakers (/VFV/ pseudowords)

Short fricatives *vs.* long fricatives

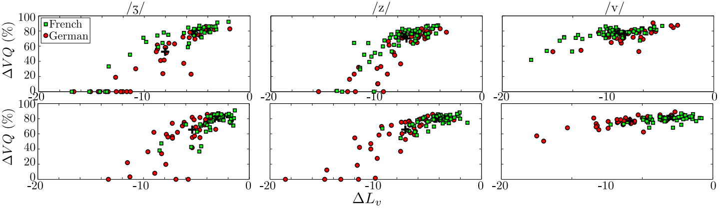


Another investigation in progress: Influence of language

German speakers devoice their fricatives more than French speakers



Learners of both German and French include these differences in the learning process



IFCASL Database: Fauth *et al.* Designing a bilingual speech corpus for French and German language learners: a two-step process. In LREC-9th Language Resources and Evaluation Conference, 2014

Plan

Some conclusions on the production of fricatives

Simulations have evidenced the role of the glottal opening in fricatives

- It controls regimes of production
- The simultaneous presence of noise and voicing is unstable

→ Several articulatory strategies for producing voiced fricatives

Possible reasons for using different strategies

- Only physiological
- Phonological context
- Contextual (sociolinguistic, prosodic. . .)

Future investigations

- Check speaker variability
- Influence of language
- Role in prosody

→ Integration into running speech synthesis

References of our works

Articulatory synthesis

- Elie B., and Laprie Y. "Extension of the single-matrix formulation of the vocal tract: consideration of bilateral channels and connection of self-oscillating models of the vocal folds with a glottal chink". Speech Comm. 82, pp. 85–96 (2016).
- Elie B., and Laprie Y. "Copy-synthesis of phrase-level utterances". EUSIPCO, Budapest, pp 868–872 (2016).
- Elie B., and Laprie Y. "A glottal chink model for the synthesis of voiced fricatives". ICASSP, Shanghai, pp 5240–5244 (2016).

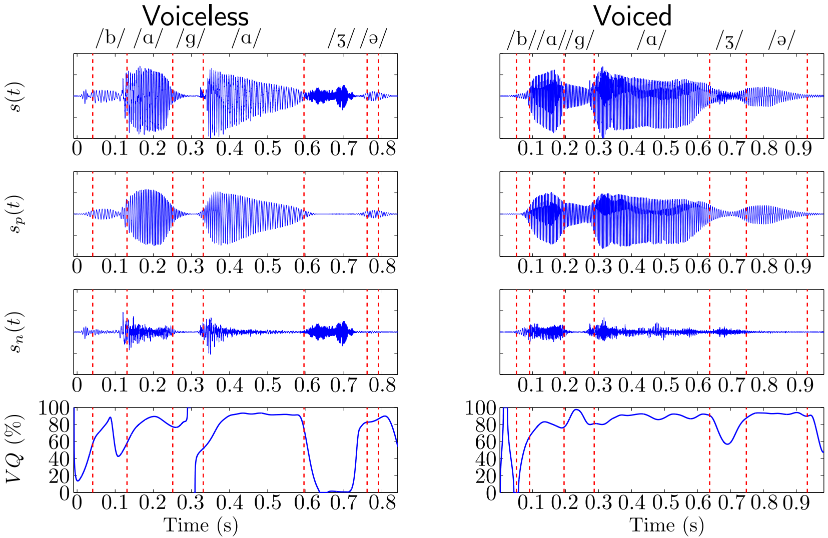
Production of fricatives

- Elie B., and Laprie Y. "Acoustic impact of the gradual glottal abduction degree on the production of fricatives: A numerical study ". J. of the Acoustical Society of America 142(3), pp. 1303–1317 (2017).
- Elie B., and Laprie Y. "Glottal opening and strategies of production of fricatives". Interspeech, Stockholm, pp. 206–209 (2017).
- Ghosh, Sucheta, et al. "L1-L2 Interference: The case of final devoicing of French voiced fricatives in final position by German learners." Interspeech (2016).

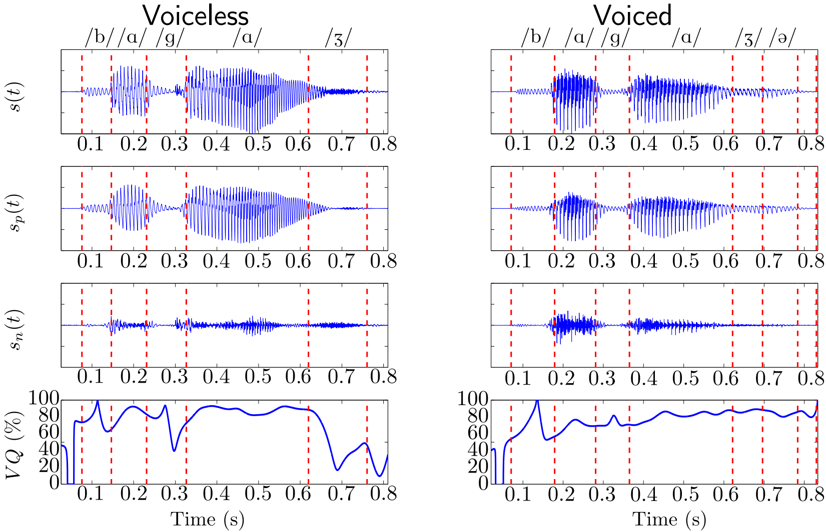
Trill production

- Elie B., and Laprie Y. "Simulating alveolar trills using a two-mass model of the tongue tip". J. of the Acoustical Society of America 142(5), pp. 3245–3256 (2017).

Example of French native speakers uttering final voiced fricatives

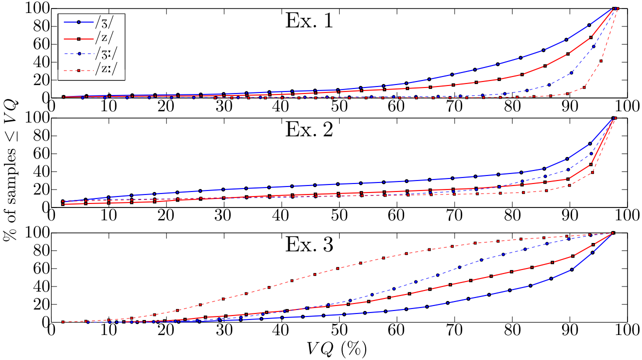


Example of French native speakers uttering final voiced fricatives



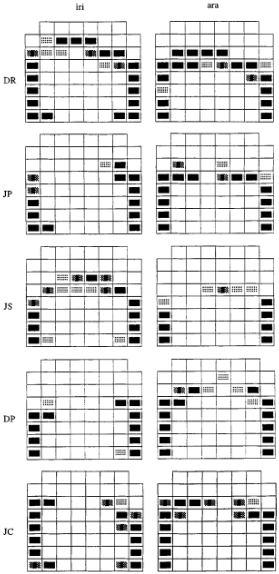
What if voiced fricatives are exaggeratedly longer ?

Cumulative histograms



Occurrence of LP contacts: some answers

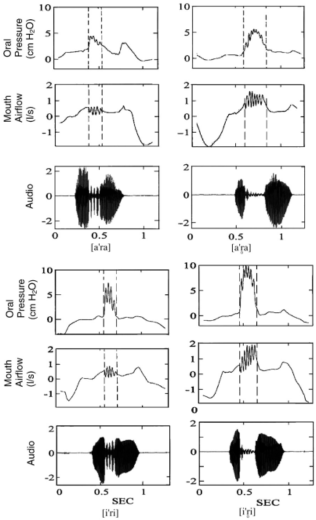
Data from Recasens and Pallarès



- Variability across speakers: some almost never make LP contacts
- Variability intra-speaker

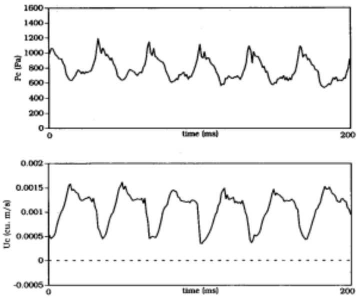
Air flow measurements

Data from Solé and McGowan



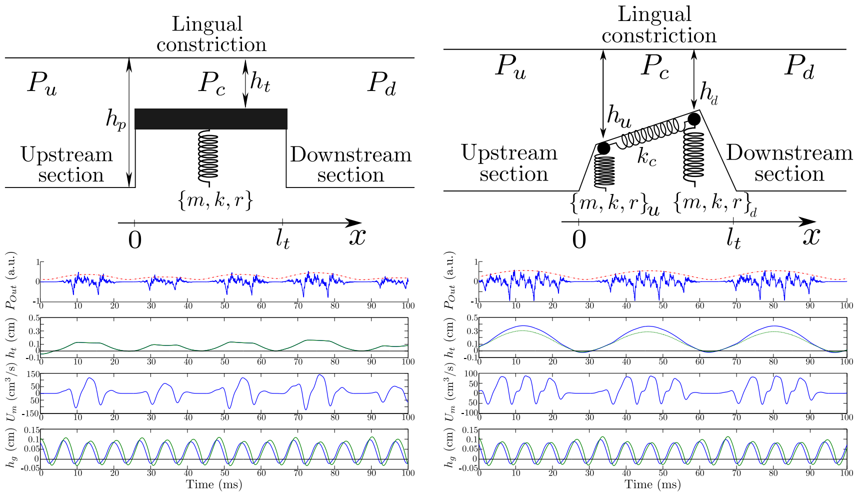
- DC component of the airflow: incomplete closure of the vocal tract ?
- are there LP contacts ?

From McGowan, on voiceless trills:



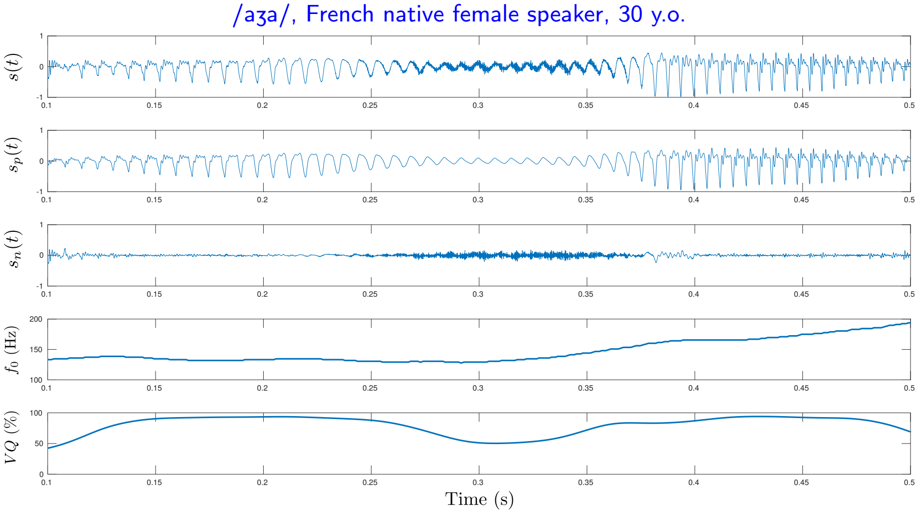
→ Needs more data

Comparison single mass and two-mass models



Effect on perception

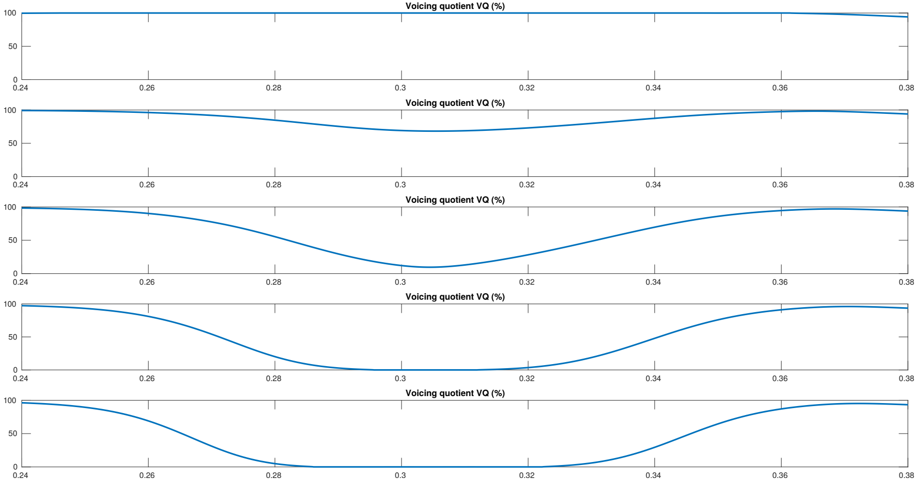
Example of natural utterance



Effect on perception

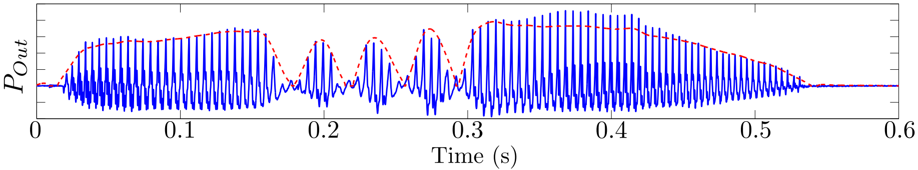
Virtual modification of *VQ*

/aʒa/for decreasing voicing



Example of alveolar trill

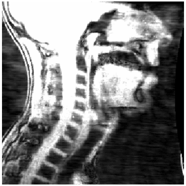
Trills in /ara/ context

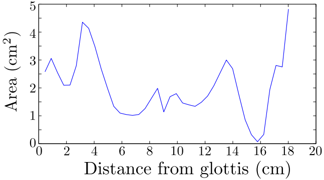


Questions:

- Can we model LP contacts and incomplete closure of the VT ?
- What are the articulatory/phonatory conditions that favor the self-oscillation of the tongue tip ?

Data taken from cineMRI acquisitions

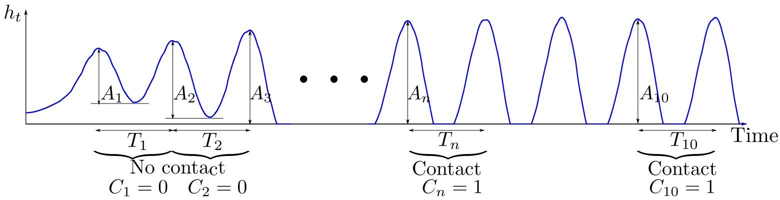




Investigation of the impact of various model parameters

- Mass of the tongue tip m_1
- Equilibrium position h_0
- Lateral ratio r_l $(= \frac{\text{open area during contact}}{\text{initial area at rest}})$
- Glottal abduction degree D_{ab}

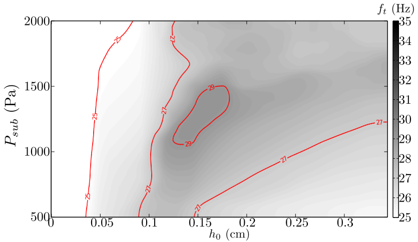
Studied features

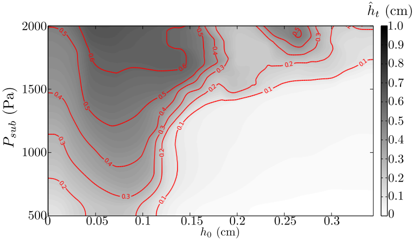


Investigation of the impact of various model parameters

- Trill frequency $f_t = \frac{1}{\hat{T}}$
- Trill amplitude $\hat{h}_t = \hat{A}$
- Contact ratio $C_r = 100 \times \frac{1}{N_{per}} \sum_{n=1}^{N_{per}} C_n$

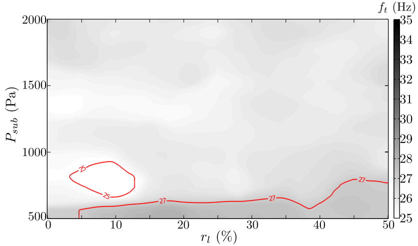
Effect of the equilibrium position

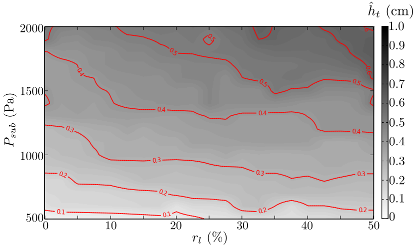




- Max of amplitude for 0.5 mm$<h_0<1$ mm
- No oscillation for $h_0>1.5$ mm, if $P_{sub}<1500$ Pa

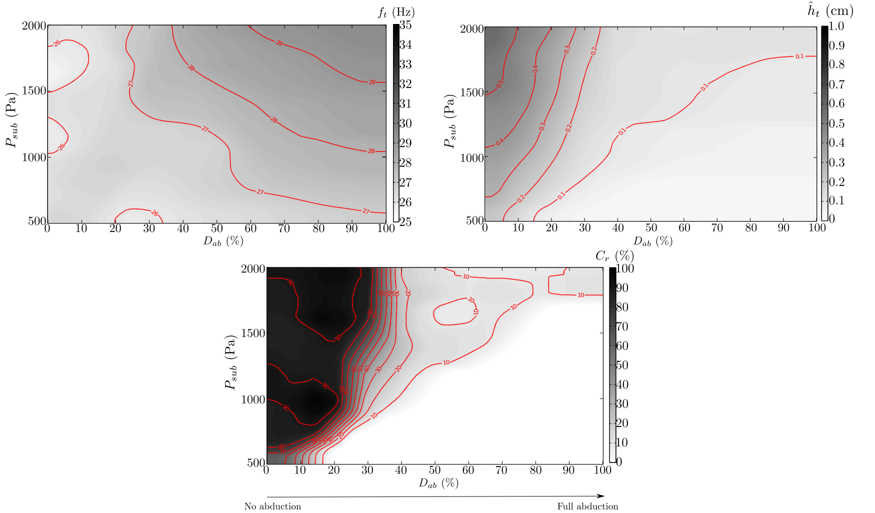
Effect of the lateral ratio





- Slight rise of the trill amplitude with lateralization
- Yet, limited impact of the incomplete closure on the trill properties

Effect of the glottal abduction degree



Glottal abduction decreases the trill amplitude