Articulatory synthesis driven by geometrical contours of the vocal tract extracted from cineradiographic data

Julie Fontecave & Frédéric Berthommier
Institut de la Communication Parlée, CNRS UMR 5009, INP Grenoble, Univ. Stendhal, France
E-mail : fonte, bertho@icp.inpg.fr

Introduction

1/ **X-ray films** = a reference technique to study speech production (dynamic view of the entire vocal tract + good temporal resolution (around 60 im/sec))
Compilation of the ATR "X-ray film database for Speech Research" (Munhall et al., 1995) including the **Laval43** sequence
2/ Manual extraction too long and laborious in this context and weak results of existing automatic extraction methods
➢ **Semi-automatic method** (introducing human expertise) = a limited manual step + an automatic extraction > Reconstruction of the entire vocal tract movements

Phonetic evaluation of the validity of the extracted contours, using an acoustic model:
Are the measurements enough precise to recover temporal and spectral features of the original speech signal ?

From cineradiographic data to geometrical contours of the vocal tract

A semi-automatic extraction method applied on Laval43 sequence

Principle of the method

1/ Manual processing (marking) of a limited number of **key images**
- Definition of geometrical features = degrees of freedom
- Description of the considered articulator position

2/ Automatic indexing of the full database according to these key images
- Each image is assigned by the index of the nearest key image

Similarity measure = Euclidean distance on the video features (low frequency **DCT components** of each image) calculated on resized, centered and framed images to focus on the concerned articulator and remove artifacts

- Geometrical marking of the full sequence ***(retro-marking)***

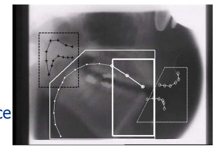
Association via the indexing between the frames and the geometrical information available for the key images only

3/ Post-processing treatments to restore the continuity
- Temporal filtering
- Averaging of neighboring configurations (multi-indexing)

Separate extraction for articulators

A specific treatment applied for each articulator
➢ same process but parameters adapted

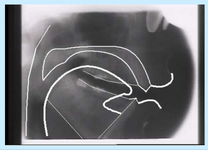
1/ original images framed and cut out
- include (only) the considered articulator for the sequence
- avoid interferences



2/ choice of the features of the method (number of key images, points and degrees of freedom, number of DCT components for the indexing…)

Reconstruction of the complete vocal tract

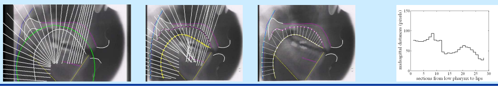
➢ combination of the various contours
- Articulators marked independently (tongue, tip, velum, lips, jaws)
- Rigid parts (palate, pharynx) also marked
- Points interpolation with spline smoothing specific for each articulator



From contours to mid-sagittal sections and area functions

From contours to sections

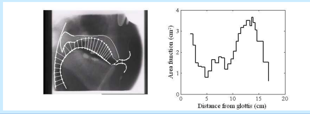
Grid along the vocal tract to measure mid-sagittal distances
- Lines orthogonal to the palate and to the tongue in average on the sequence
- Correction image by image (Yehia, 2002)



28 mid-sagittal distances :
> 26 thanks to the lines
> 1 between front teeth
> 1 between lips

From sections to area functions

- $\alpha\beta$ Model (Heinz & Stevens, 1965) $A(x) = \alpha(x) d(x)^{\beta(x)}$
- Parameters α et β elaborated for a male speaker (Soquet et al., 2002)
- Estimation of the glottis position (not visible)

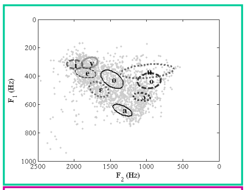


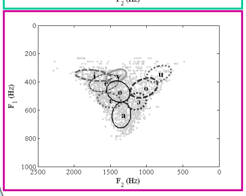
From the geometry of the vocal tract to its acoustics: Articulatory Synthesis

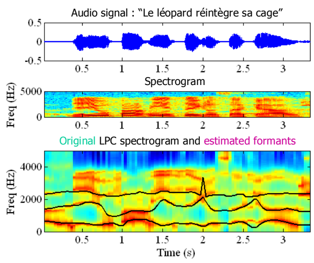
Simulation of the acoustics in the vocal tract, starting from the extracted geometrical data and using the "source-filter" speech production concept
filter = transfer function associated to the area function calculated with an electrical analog of the vocal tract (Badin & Fant, 1984)

Formants comparison

Estimated formants: extracted from transfer functions of the vocal tract
Reference formants: extracted from the audio signal (Praat)







	Correlation			Mean difference (%)		
	Seq. with silence	Seq. without silence	Vowels*	Seq. with silence	Seq. without silence	Vowels*
F_1	0.08	0.66	**0.86**	26	18	**12**
F_2	0.35	0.61	**0.75**	19	14	**13**
F_3	0.06	0.07	**0.22**	12	7	8

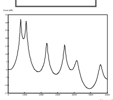
Vowels*: [a], [i], [e], [u], [y], [o], [ɛ], [ø], [ɔ]

Speech synthesis

Source → Amplitude modulation → Filter → Synthesized signal

Source: original signal whitened with a Hilbert filter and a LPC inverse filtering

Amplitude modulation: 2 subband amplitudes
low frequencies (0-1 KHz) / high frequencies
➢ Consonants "recovering"



Intelligible speech

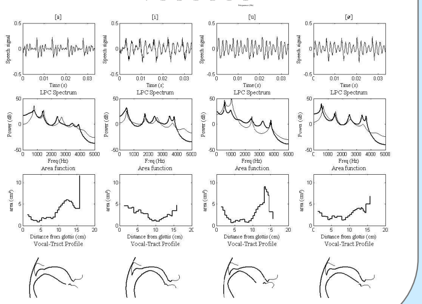
Spectral distance on the LPC spectrums
➢ to evaluate the similarity between original and synthesized signals

$$d(S,S') = \frac{1}{p} \sum_{k=1}^{p} \left| S(k) - S'(k) \right|$$

p = nb of frequency bins taken into account (0-3.5 KHz)

Reference \ Estimation	$S_{m,f}$	S_f	$S_{m,fd}$
S_o	6.27 dB	8.44 dB	6.84 dB
$S_{m,i}$	5.27 dB	8.95 dB	5.99 dB

o = original
m = amplitude modulation
i = filtering with the LPC spectrums of the original signal
f = filtering with the transfer functions
fd = filtering with time-shifted transfer functions



Conclusions and Prospects

- Extraction of geometry and movements of the vocal tract using our semi-automatic method leads to intelligible speech synthesis.
- A perception test is in progress to evaluate the quality of this resulting speech.
- The tongue contact events are analyzed and related to production of consonants (especially thanks to the tongue tip tracking).

Acknowledgments: We thank K. Munhall and B. Burt for providing a DVD with a copy of the ATR "X-Ray film database for Speech Research", including the Laval43 sequence.
We thank D. Beautemps, P. Badin and A. Serrurier for their help as for the use of the acoustic model.