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

Julie Fontecave, Frédéric Berthommier

ICP, Institut de la Communication Parlée
INPG, 46 Av. Félix Viallet, 38031 Grenoble cedex1, France

fonte@icp.inpg.fr, bertho@icp.inpg.fr

Abstract. *Recently, we have proposed a new technique for facilitating the extraction of vocal tract contours from complete sequences of large existing cineradiographic databases. The articulators (tongue, tongue tip, velum, lips, etc.) are processed independently before being combined to reconstruct the whole vocal tract. Applied to one sequence of the ATR database, Laval43, the method allows us to estimate the shape of the complete vocal tract and the corresponding mid-sagittal sections. These are compatible with standard articulatory synthesis models. The formant trajectories are synthesized using the transfer functions calculated from the estimated area. A comparison between estimated and original formants is carried out. Then, by introducing the 2-subbands amplitude modulation extracted from the original audio signal, the synthesis of intelligible speech is realized and spectral distances are evaluated.*

1. Introduction

In this article, we attempt to develop a speech synthesis process driven by vocal tract contours acquired from cineradiographic data, using a new semi-automatic method (Fontecave & Berthommier, 2005, 2006).

Classically, X-ray films are a reference technique to study speech production since it combines a dynamic view of the entire vocal tract and a good temporal resolution (about 50 im/sec). In this framework, Munhall et al. (1995) have compiled the ATR "X-ray film database for Speech Research" from films contributed by Perkell, Stevens and Rochette in order to make them available for the speech research community. Thus, we could benefit of the Laval43 sequence and propose a method which facilitates the long and laborious step of manual contouring. For each articulator, the method associates the manual marking of anchor points on a limited number of key images randomly chosen within a sequence and then performs the automatic reconstruction of contours and movements. At first, each articulator is treated independently with specific parameters. Then they are combined together to obtain the shape of the vocal tract. The resulting geometrical model is defined by degrees of freedom marked on the vocal tract (x or y coordinates of anchor points). However, we show (Fontecave & Berthommier, 2006) that we are able to estimate sagittal distances corresponding to the contours and we calculate the associated area functions.

The generation of area functions from measurements of the sagittal sections (Heinz & Stevens, 1965) is a critical step for applying an acoustical model based on the

vocal tract geometry. Acoustic transfer functions are computed from these area functions using an electrical analog of the vocal tract (Badin & Fant, 1984). A comparison between original and estimated formants positions is carried out, at first for the vowels pronounced in the Laval43 sequence. The synthesis of full spectra is also realized using a whitened source signal and an amplitude modulation estimated from the audio signal. Thus, extracted contours, mapped onto vocal tract transfer functions and combined with the source signal, yield a speech signal which is spectrally compared to the original recording.

Extraction method details, area functions generation and articulatory synthesis are presented in the following sections.

2. From cineradiographic data to vocal tract area functions

This semi-automatic method is based on an adaptation of the "retro-marking" algorithm (Berthommier, 2004) initially applied for the estimation of the mouth opening parameters from video data recorded without the traditional blue chroma key. This has been transposed for extracting the tongue contour in the "Wioland" database (Fontecave et al., 2005) and then extended for the treatments of other articulators and other databases (Fontecave et al., 2006), including the sequence "Laval43" from the ATR cineradiographic database (Munhall et al., 1995). The "X-ray film database for Speech Research" is the largest one available, with 25 different films, 55 minutes and nearly 100000 images. The sequence Laval43 recorded in 1974 is composed of 4043 images (720*480 pixels) of the vocal tract, from a video film (sentences read by a male native speaker of Canadian French) recorded at 50 im/sec.

2.1. Semi-automatic extraction method for the complete vocal tract

First, the manual step aims to describe with a few degrees of freedom and for a small set of key images chosen randomly among the whole database, the position of anchor points related to the contour of the target articulator (tongue, tongue tip, velum or lips). A majority of these points are placed at the intersection between lines and visible contours, and some are free (eg., the tongue tip). This step defines the geometrical features and consists in a marking process. Motion and adjacent images are taken into account to facilitate the identification of theses features, barely visible in many cases on the static key image. The choice of the intersecting lines is made to miss no data. For each key image, a sketch of the articulator is then obtained by connecting the points (Fig. 1a).

The following step is the automatic indexing of the video sequence according to these key images. For each frame, the index of the nearest key image is assigned. The similarity measure is the Euclidean distance between the lowest frequency DCT (Discrete Cosine Transform) components of the images. These video components are calculated on a reduced frame, chosen to focus on the concerned articulator and to remove some artifacts. Then an association is built between the frames and the geometrical information available for the key images only, and the geometrical marking of the original sequence is thus realized. This first estimation is affected by quantization effects and some indexing errors. The reconstruction error is reduced by complementary subsequent treatments (temporal filtering and multi-indexing).

Considering each articulator independent, this complete process is applied to each articulator of the vocal tract, with appropriate parameters. The original images are

first framed and cut out so as to only include the considered element for the whole sequence and to avoid interferences. Then the features of the method, such as the number of key images, the points and degrees of freedom or the number of DCT components used for indexing, are decided independently for each element. Frames and marking points chosen for various articulators are shown Figure 1a. Let remark that the tip is considered as an independent articulator and its estimate is merged by substitution in the global tongue contour estimate (Fontecave & Berthommier, 2006).

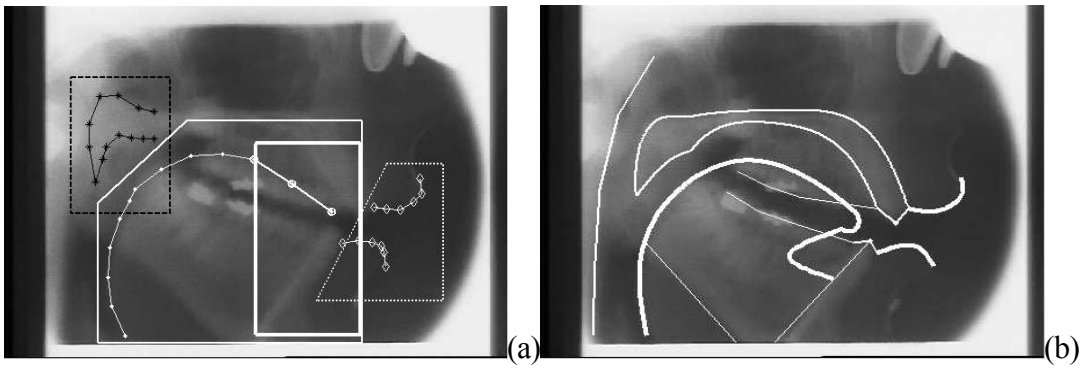
(a) (b)

Figure 1. (a) Frames and degrees of freedom defined for various articulators (from left to right: velum, tongue, tongue tip and lips) (b) Contours of the whole vocal tract in Laval43.

To complete the shape of the vocal tract, the rigid parts (palate, pharynx) are also marked using a few key images. The full contour of the vocal tract is reconstructed by combining with a set of appropriate spline interpolations the contours of each articulator (Fig. 1b and http://www.icp.inpg.fr/~bertho/m2p/issp06/Vtract-laval43.html).

2.2. From contours to mid-sagittal sections and area functions

For each image, starting from the contours described above, the mid-sagittal distances are measured along a grid. This is not the usual semi-polar grid which is adopted. The lines of our grid were initialized perpendicular to the palate or to the pharynx. Then the orientation of these lines has been slightly corrected in average along the midline in the mid-sagittal view to be as orthogonal as possible to the tongue as well. Finally, the classical correction image by image has been carried out using the Yehia's method (2002). Along the vocal tract, 28 mid-sagittal distances are measured: 26 thanks to the lines defined from the low pharynx to the alveolar zone, 1 between the upper and lower front teeth and 1 between the lips.

In order to study the relations between the geometry of the vocal tract and the acoustics, the midsagittal view is transformed into an area function, which is the input to an electrical analog of the vocal tract. The model used to convert midsagittal distances into area functions is based on the original transformation defined by Heinz & Stevens (1965), which is $A(x) = \alpha(x)d(x)^{\beta(x)}$, where d is the midsagittal distance, A the cross-sectional area, x the position along the vocal tract mid-line, α and β the parameters of the transformation. Commonly, the authors adapt the value of the transformation parameters to the speaker and the position along the vocal tract mid-line. Since we have no means to estimate these parameters, we use the ones defined by Soquet et al. (2002) for a male speaker.

The glottis position is necessary to evaluate the vocal tract length, but this is not visible in Laval43, and we fix it along the mid-line. The pixels/cm ratio is evaluated at 38

pixels for 1 cm. Thanks to those estimations, the mean vocal tract length is 17.5 cm and an area function is available along the vocal tract for each image.

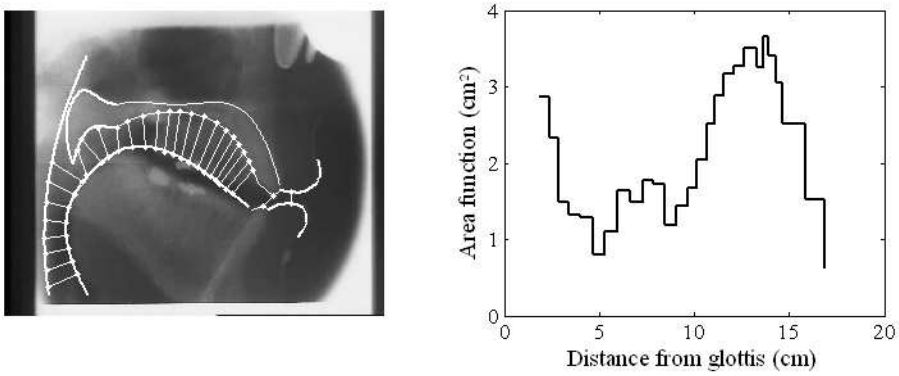

Figure 2. Midsagittal sections and area function calculated for one frame of Laval43

3. From the geometry of the vocal tract to its acoustics

The main purpose of this paper is to apply an acoustical model using extracted geometrical data. As in existing articulatory synthesis models (Maeda, 1982, Narayanan & Alwan, 2000 - among others), the "source-filter" concept of speech production is applied: the linear acoustic behavior of the vocal tract filter can be modeled as an electrical transmission line model assuming planar wave propagation. For each frame, a transfer function associated to the area function is calculated with a harmonic simulation using an electrical analog of the vocal tract. This includes losses and boundary conditions, as proposed by Badin & Fant (1984), so that the viscosity and heat conditions losses, wall impedances and radiation are taken into account.

From these synthesized transfer functions, formants positions are extracted for the whole sequence. A comparison of these estimated formants with the original ones is first carried out. In a second time, we attempt to synthesize a complete speech signal. In the "source-filter" framework, we propose a new decomposition of the source into two separable components: a whitened source signal (the so called excitation source) and a 2-subbands amplitude modulation. An evaluation is carried out on LPC spectrums, and not only on formants positions, in order to take into account the effect of this decomposition in the comparison study.

4.1. Comparison of formants positions

Spectral transfer functions are synthesized for the whole sequence; for each frame, the 3 first formants (F_1, F_2 and F_3) are extracted from these functions with a peak detection algorithm. They are compared to reference formants, evaluated from the audio signal using Praat (Boersma & Weenink). Comparisons are made in terms of correlation c and mean difference μ between the estimation and the reference. μ is normalized in respect to the reference frequencies, as defined below (p corresponds to the frame number).

$$\mu(F_i) = \frac{1}{N}\sum_{p=1}^{N}\frac{\left|F_i(p) - F_{i,ref}(p)\right|}{F_{i,ref}(p)}$$

The speech frames are selected thanks to an amplitude threshold in the low frequency domain ([0-500Hz]). It allows us to consider results for the sequence, with or without silence. To focus on the vowels of the corpus, we have selected about 200 non-consecutive frames from the speech signal corresponding to French vowels [a], [i], [e], [u], [y], [o], [ɛ], [ø], [ɔ].

Results are summarized in Table 1. The linear correlation between reference formants and estimated ones is improved when the observation is restricted to speech frames and even more for vowels frames. The mean difference is also better for these selected frames. The correlation coefficient indicates that the F_3 estimation is not satisfactory so that the other measures are not significant. On the contrary, for F_1 and F_2, we observe a high correlation, and the distance between original and estimated formants is evaluated at about 16% for the sequence without silence.

	Correlation c			Mean difference μ (%)		
	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

Table 1. Comparison between reference formants from the original signal and formants estimated from the geometry of the vocal tract

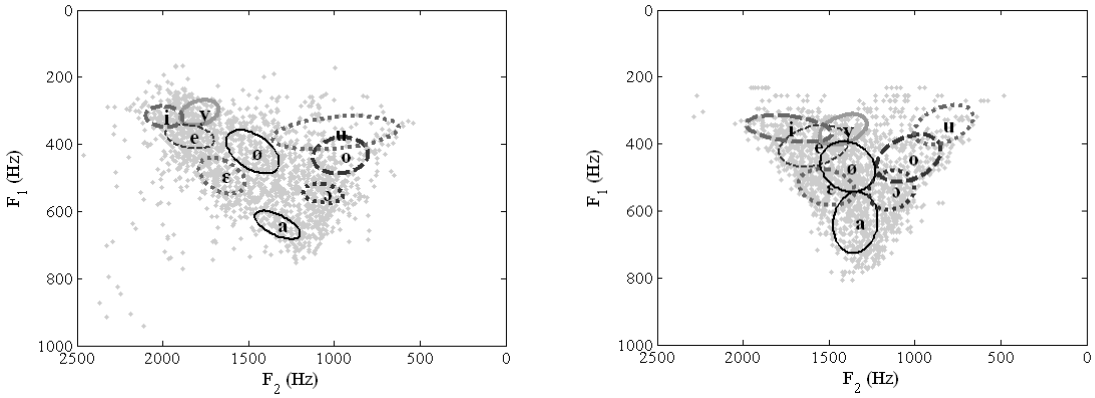

Figure 3. Formant positions plotted for all speech frames (in gray) and dispersion ellipses for selected vowels: (a) Formants from the original audio signal (using Pratt) - (b) Formants estimated from the vocal tract contours.

F_1-F_2 representations in figure 3 show the speech frames (gray points) and the dispersion ellipses (one standard deviation) for the selected frames of vowels. The mean values of formants are correct for each vowel class. The overlapping is higher for the estimated formants (Fig. 3b). According to the F_1-F_2 Euclidean distance between vowels and the class centroïds, the classification rate is 72% for the reference and 50% for the estimation.

4.2. Synthesized signal

The audio synthesis is based on 3 components: an excitation source, an amplitude modulation and a frequency modulation. Classically, the transfer function calculated from area function is applied as a linear filter on a "source" signal. Here, we decompose this source signal into two components estimated from the original audio signal. The excitation source is obtained by whitening the original signal: we take the cosine part of the Hilbert transform and apply a complementary LPC inverse filtering. The amplitude information is introduced in the source by reassigning the subband amplitudes, evaluated in the original signal at 30 frames/s, to the previous excitation source. A filterbank designed for 2 subbands cuts the frequency domain at 1000Hz, between F_1 and F_2.

To evaluate the similarity between the original and the synthesized signals, we use a spectral distance calculated with the LPC spectrums scaled in dB. For one frame n, considering 2 LPC spectrums S and S' (in dB):

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

where f is the number of frequency bins taken into account in the distance.
We limit our distance evaluation to the [0-3.5 KHz] bandwidth, corresponding approximately to the 3 first formants. The spectral distance D, referred for the comparison of signals, is the mean value of d.

The distance D is calculated between the LPC spectrums of :

- the original signal (S_0)
- the synthesized signal (S_{i1}) obtained by filtering the whitened and amplitude modulated source with the LPC spectrums of the original signal
- the synthesized signal (S_{f1}) obtained by filtering the whitened and amplitude modulated source with the transfer functions
- the synthesized signal (S_{f2}) obtained by filtering the source, whitened but not modulated, with the transfer functions
- the synthesized signal (S_{fd1}) obtained by filtering the whitened and amplitude modulated source with time-shifted transfer functions

	S_{f1}	S_{f2}	S_{fd1}
S_0	6.27 dB	8.44 dB	6.84 dB
S_{i1}	5.27 dB	8.95 dB	5.99 dB

Table 2. Average spectral distance between original and synthesized LPC spectrums for the whole sequence

(S_0) and (S_{i1}) are the reference signals. The difference between the whole sequence, the speech frames or the selected vowel frames is not significant. We observe that the distance D is lower between original and synthesized signals when the synthesis incorporates the amplitude modulation component. The signal is also more intelligible in this condition. By shifting the transfer functions of a few frames (S_{fd1}), formants are not correctly estimated, the distance D is just a little higher than for (S_{f1}), but the resultant signal is not intelligible at all. In this case the amplitude modulation is not synchronous with the frequency modulation.

The Figure 4 shows, for a short sequence of the corpus, the formants estimations superimposed on the LPC spectrogram of the original signal. Despite some handovers, the formants trajectories are well synthesized.

Four vowel frames are shown in figure 5, in which we compare the LPC spectrums of original and synthesized signals (the synthesis is realized with amplitude modulation). We observe a good agreement between the original and the estimated spectrums when the comparison is limited to 3.5 KHz.

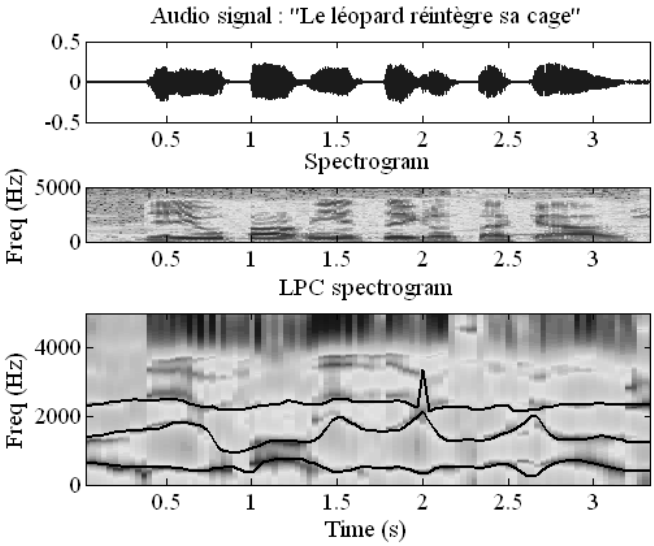

Figure 4. Spectrogram of the original signal and estimated formants (in black thick lines)

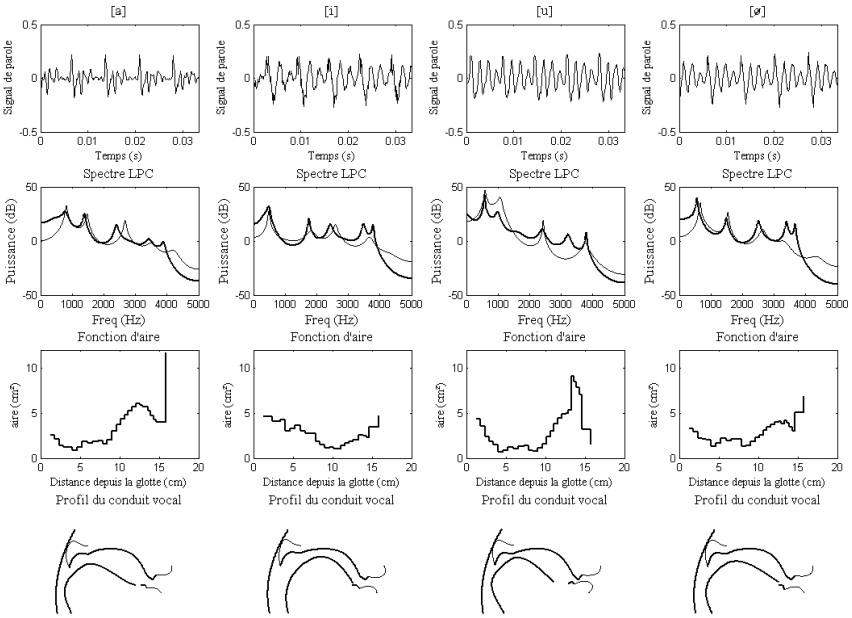

Figure 5. The LPC spectrum (thick line) of the original signal is compared to the one estimated from the area function (thin line). From left to right, the columns correspond to 4 frames of the French vowels [a], [i], [u] and [ø]. Speech signal, area functions and vocal tract profiles are also shown.

4. Conclusion

Experimental tests of speech synthesis from cineradiographic data of the vocal tract have been presented. This is a manner to evaluate the quality of our measurements. Thanks to the retro-marking method based on low frequency DCT video parameters, the extraction of geometry and movements of the vocal tract in the complete sequence of Laval43 has lead to mid-sagittal sections and area functions measurements. This knowledge allows an estimation of the formants trajectories, distant of 16% from the reference. To synthesize a complete speech signal, excitation source and amplitude modulation are necessary in addition to the spectral filtering. Since they cannot be estimated from the vocal tract measurements, we infer them from the original audio signal. Thus, the whitened source is modulated in amplitude according to the original amplitude observed in the audio signal decomposed in 2 subbands. This source signal is then filtered by the estimated vocal tract transfer functions, and the resultant signal is somewhat distorted but intelligible. A perception test is in progress to evaluate better the quality of this synthetic speech.

Acknowledgments: We thank K. Munhall and B. Burt for providing a DVD with a copy of the ATR "X-Ray films 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.

References

Badin, P. and Fant, G. Notes on vocal tract computation. *Speech Transmission Laboratory - Quarterly Progress Status Report*, Stockholm, 2-3, pages 53-108, 1984.

Berthommier, F. Characterization and extraction of mouth opening parameters available for audiovisual speech enhancement. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, vol. 3, pages 789-792, 2004.

Boersma, P. and Weenink, D. Praat : doing phonetics by computer (Version 4.3.14) [Computer program]. Retrieved May 6, 2005, from http://www.praat.org/.

Fontecave, J. and Berthommier, F. Quasi-automatic extraction method of tongue movement from a large existing speech cineradiographic database. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 1081-1084, 2005.

Fontecave, J. and Berthommier, F. Semi-Automatic Extraction of Vocal Tract Movements from Cineradiographic Data. In *Proceedings of the International Conference on Spoken Language Processing*, 2006.

Heinz, J.M. and Stevens, K.N. On the relations between lateral cineradiographs area functions and acoustic spectra of speech. In *Proceedings of the 5th International Congress of Acoustics,* Liège, Paper A44, 1965.

Maeda, S. A digital simulation method of the vocal-tract system. *Speech Communication*, 1, pages 199-229.

Munhall, K.G., Vatikiotis-Bateson, E. and Tohkura, Y. X-ray Film database for speech research. *The Journal of the Acoustical Society of America*, 98, pages 1222-1224, 1995.

Narayanan, S. and Alwan, A. Noise Source Models for Fricative Consonants. *IEEE Transactions on Speech and Audio Processing*, 8(2), pages 328-344, 2000.

Soquet, A., Lecuit, V., Metens, T. and Demolin, D. Mid-sagittal cut to area function transformations : Direct measurements of mid-sagittal distance and area with MRI. *Speech Communication*, 36, pages 168-180, 2002.

Yehia, H.C. A study on the speech acoustic-to-articulatory mapping using morphological constraints. *PhD Thesis,* Nagoya University, Graduate School of Engineering, 2002.