Audiovisual to area and length functions inversion of human tract

Benjamin Elie, Yves Laprie

AUDIOVISUAL TO AREA AND LENGTH FUNCTIONS INVERSION OF HUMAN VOCAL TRACT

Benjamin Elie, Yves Laprie

LORIA, INRIA / CNRS / Université de Lorraine, Nancy, France

ABSTRACT

This paper proposes a multimodal approach to estimate the area function and the length of the vocal tract of oral vowels. The method is based on an iterative technique consisting in deforming an initial area function so that the output acoustic vector matches a specified target. The chosen acoustic vector is the formant frequency pattern. In order to regularize the ill-problem, several constraints are added to the algorithm. First, the lip termination area is estimated via a facial capture software. Then, the area function is constrained in such a way that it does not get too far from a neutral position, and it does not change too quickly from a temporal frame to the next, when dealing with dynamic inversion. The method proves to be efficient to approximate the area function and the length of the vocal tract for oral french vowels, both in static and dynamic configurations.

***Index Terms*—** Audiovisual inversion, Vocal tract length, Regularization, Dynamic inversion

1. INTRODUCTION

The problem of the acoustic-to-articulatory inversion, *i.e.* the recovery of the vocal tract shape from the recording of the speech signal has been widely studied since the 60's. Proposed methods are classified into two main categories: the data driven, or machine learning methods [1–3], and the analysis-by-synthesis methods [4–9], which use an acoustic model of the vocal tract in order to adjust the input parameters of the model so that the acoustic output matches the features of the recorded speech signal. The presented method belongs to the latter.

In this paper, the chosen aforementioned input parameter of the direct mapping from the articulatory to the acoustic domain is the area function of the vocal tract, which is modeled as a concatenation of N cylindrical tubes. The input parameter is given by a vector containing the N cross-sectional areas and the lengths of the tubes. The output acoustic vector is the frequency pattern of the first 3 or 4 formants, as used in many studies [4, 5, 8, 10, 11], namely those lying below the frequency limit of the planar wave assumption, at 4 kHz.

The method uses the sensitivity functions derived from Fant [12], and later used for inversion problems by Story [8, 13] and Carré [5]. Basically, the sensitivity function, related to the derivative of the normalized formant frequency with respect to the normalized area function, is computed at each iteration to generate a new area function which reduces the distance between the observed formant frequencies and those generated by the model at the new iteration. Iterations are repeated until this distance is smaller than a certain arbitrary threshold. This method requires an initial function to be generated, that is the first area function from which the algorithm starts. Even though this initial function is arbitrary, it is recommended to include *a priori* knowledge of basic human vocal tract configurations. Indeed, it is a relevant way to regularize the problem, which is very ill-posed. This paper proposes a multimodal inversion, by adding the estimation of the lip aperture, by means of a facial capture software. This method enables the value of three articulatory parameters to be estimated: the lip opening, the lip protrusion, and the jaw opening.

Such multimodal inversion has been previously proposed by Bunton *et al.* [8]. The authors fixed the total length of the vocal tract to an arbitrary value. This can be an important issue since the total length of the vocal tract of a speaker is likely to vary during speech production. This paper proposes, first, to improve the method by taking into account the length of the vocal tract. The length of the vocal tract is computed the same way as the area function, namely in the same iterative loop. Secondly, the paper deals with dynamic inversion, *i.e.* the recovery of the trajectory of the vocal tract area function along a speech segment, where constraints on the articulatory trajectory may be added to better regularize the problem [7, 11].

The paper is organized to describe the main aspects of our approach. Section 2 details the vocal tract model used in this paper as well as the computation of the sensitivity matrices used to approach the target area and length functions. The sensitivity matrices are derived from the theory of area and length perturbations along the vocal tract by Fant [12]. Section 3 presents the entire algorithm to estimate the area and the length functions of the vocal tract, as well as the different constraints used to regularize the inverse problem. Finally Section 4 presents experimental results for static and dynamic configurations.

2. VOCAL TRACT AREA AND LENGTH PERTURBATIONS: THEORETICAL BACKGROUND

This section presents the underlying theory of the presented algorithm.

2.1. Description of the vocal tract

The human vocal tract forms the boundaries of an air volume. The complexity of dealing with acoustic wave propagation along a volume can be overcome by considering the acoustic waves inside the vocal tract as planar or mono-dimensional waves at frequencies below 4 kHz. This assumption is widely considered in the existing literature. Thus, only the cross-sectional area of the vocal tract is needed to describe its acoustic characteristics. The vocal tract is then characterized by the area function $a(x,t)$, corresponding to the cross-sectional area of the vocal tract at the distance x from a certain chosen origin (usually the glottis), at a given instant t. This is valid if the cross-sectional shape is assumed to be uniformly circular.

The continuous cross-sectional area can be sampled, so that the vocal tract is modeled as a set of connected tubelets. Note that $a(x,t)$ is not necessarily evenly sampled along x. For the sake of clarity in the notations, the time dependency t is omitted in most

equations of the paper. In that case, the defined quantity are those at a given instant t. The sampled vocal tract is then described by two vectors:

$$\begin{cases} \mathbf{a} = [a_1, a_2, \ldots, a_n, \cdots, a_N]^T \\ \mathbf{l} = [l_1, l_2, \ldots, l_n, \cdots, l_N]^T \end{cases}, \quad (1)$$

In this paper, a_1 corresponds to the cross-sectional area of the vocal tract at the glottis end, and a_N is the cross-sectional area at the lips, *i.e.* the lip aperture.

2.2. Sensitivity matrices

This section deals with the computation of the partial derivative matrices of the formant frequencies with respect to the area and length functions of the vocal tract.

2.2.1. Derivative of the formant frequencies with respect to the area and length functions

The perturbation theory by Fant [12] establishes the relation between a small change of cross-sectional areas and tube lengths along the vocal tract, and the change of formant frequencies. Eq. (2) and Eq. (3) give the relative variation of the m^{th} formant frequency for relative variations of the elements of $\mathbf{a}$ and $\mathbf{l}$ from Eq. (1), respectively. These relationships are as follows:

$$\left[\frac{\Delta\omega_m}{\omega_m}\right]_{\mathbf{a}} = \frac{\sum_{n=1}^{N} \left[\mathcal{T}_n(\omega_m) - \mathcal{V}_n(\omega_m)\right] \frac{\Delta a_n}{a_n}}{\sum_{n=1}^{N} \left[\mathcal{T}_n(\omega_m) + \mathcal{V}_n(\omega_m)\right]}, \quad (2)$$

and

$$\left[\frac{\Delta\omega_m}{\omega_m}\right]_{\mathbf{l}} = \frac{\sum_{n=1}^{N} \Delta\lambda_n \left[\mathcal{T}_n(\omega_m) + \mathcal{V}_n(\omega_m)\right]}{\sum_{n=1}^{N} \left[\mathcal{T}_n(\omega_m) + \mathcal{V}_n(\omega_m)\right]}, \quad (3)$$

where

$$\Delta\lambda_n = -\frac{\Delta l_n}{l_n + \Delta l_n}, \quad (4)$$

and where $\mathcal{T}_n(\omega_m)$ and $\mathcal{V}_n(\omega_m)$ are the kinetic and the potential energy inside the n^{th} tube, at the resonance angular frequency ω_m. The potential and kinetic energies for pressure P and volume velocity U are given by

$$\mathcal{T}_n(\omega_m) = \frac{1}{2}\frac{\rho l_n}{a_n} \left|U_n(\omega_m)\right|^2 \quad (5)$$

$$\mathcal{V}_n(\omega_m) = \frac{1}{2}\frac{\rho c_s^2}{a_n l_n} \left|P_n(\omega_m)\right|^2. \quad (6)$$

One can then define two sensitivity functions, $S_n^a(\omega_m)$ and $S_n^l(\omega_m)$ for the cross-sectional area perturbations and the length perturbation respectively

$$S_n^a(\omega_m) = \frac{\mathcal{T}_n(\omega_m) - \mathcal{V}_n(\omega_m)}{\mathcal{H}(\omega_m)}, \quad (7)$$

and

$$S_n^l(\omega_m) = \frac{\mathcal{T}_n(\omega_m) + \mathcal{V}_n(\omega_m)}{\mathcal{H}(\omega_m)}, \quad (8)$$

where the index m denotes m^{th} formant frequency, and $\mathcal{H}(\omega_m)$ is the total energy in the vocal tract at the m^{th} formant frequency, hence

$$\mathcal{H}(\omega_m) = \sum_{n=1}^{N} \left[l_i^2 \left|U_n(\omega_m)\right|^2 + c_s^2 \left|P_n(\omega_m)\right|^2\right]. \quad (9)$$

The functions $S_n^a(f_m)$ and $S_n^l(f_m)$ describe the relative variation of the m^{th} formant frequency with respect to the relative variation of area and length function of the vocal tract, respectively.

This yields the following sensitivity matrices

$$\mathbf{J_a} = \begin{bmatrix} S_1^a(f_1) & S_2^a(f_1) & \cdots & S_N^a(f_1) \\ S_1^a(f_2) & S_2^a(f_2) & \cdots & S_N^a(f_2) \\ \vdots & \ddots & \ddots & \vdots \\ S_1^a(f_M) & S_2^a(f_M) & \cdots & S_N^a(f_M) \end{bmatrix}, \quad (10)$$

and

$$\mathbf{J_l} = \begin{bmatrix} S_1^l(f_1) & S_2^l(f_1) & \cdots & S_N^l(f_1) \\ S_1^l(f_2) & S_2^l(f_2) & \cdots & S_N^l(f_2) \\ \vdots & \ddots & \ddots & \vdots \\ S_1^l(f_M) & S_2^l(f_M) & \cdots & S_N^l(f_M) \end{bmatrix}, \quad (11)$$

which quantify relative variations of formant frequencies with respect to relative variations of areas and lengths along the vocal tract.

In this study, the sensitivity matrices are computed using the chain-matrix paradigm of Sondhi and Schroeter [14].

3. RECURSIVE ALGORITHM FOR ESTIMATING THE VOCAL TRACT AREA AND LENGTH FUNCTIONS

This section describes the algorithm used to estimate the area function and the length of the vocal tract, from the formant frequencies, by means of the sensitivity matrix technique.

3.1. General idea of the algorithm

The technique uses an iterative computation of the area function, aiming to match the measured formant frequencies. The algorithm requires an initial area function. A common solution for such inverse problems uses the Jacobian inverse technique [15]. Given an initial vector $\mathbf{a}_0$ or $\mathbf{l}_0$, denoted by $\mathbf{x}_0$, producing a initial formant frequency pattern $\mathbf{f}_0$, the problem can be written

$$\Delta\mathbf{f} = \mathbf{J}_x \Delta\mathbf{x}, \quad (12)$$

where $\Delta\mathbf{x} = \mathbf{x} - \mathbf{x}_0$ and $\Delta\mathbf{f}$ is the difference between the formant frequency target $\mathbf{f}$ and $\mathbf{f}_0$. $\Delta\mathbf{x}$ is then the vector to which $\mathbf{x}_0$ should be added to reach the target. It is then expressed

$$\Delta\mathbf{x} = \psi\hat{\mathbf{J}}\Delta\mathbf{f}, \quad (13)$$

where $\hat{\mathbf{J}}$ is either the Moore-Penrose pseudo inverse of $\mathbf{J}$, or the transpose, and ψ is a scalar coefficient used to scale the difference vector to a sufficiently small value. This operation is iteratively applied until $\Delta\mathbf{x}$ vanishes. In our case, the algorithm uses the transpose technique since it has been found empirically to be the most efficient technique.

In the paper, the iterations for the area correction and the length correction are simultaneously applied. It is equivalent to a vector $\Delta\mathbf{x}$ made of the vertical concatenation of $\Delta\mathbf{a}$ and $\Delta\mathbf{l}$, and $\mathbf{J}_x$ is the horizontal concatenation of $\mathbf{J}_a$ and $\mathbf{J}_l$. Consequently, the area correction and the length correction are not concurrent and there is no problem running them simultaneously.

3.2. Estimation of vocal tract areas

The next area function is computed according to the following relationship

$$\mathbf{a}_{k+1} = \mathbf{a}_k + \psi \mathbf{A}_k \mathbf{J}_\mathbf{a}^T \mathbf{f}_k, \tag{14}$$

where $\mathbf{A}_k \in \mathbb{R}_+^{*N\times N}$ is a diagonal matrix where the diagonal elements are the elements of the area function vector $\mathbf{a}_k$, namely $\mathbf{A}_k = \text{diag}\,(a_1, a_2, \ldots, a_N)$. $\mathbf{J}_\mathbf{a} \in \mathbb{R}^{M\times N}$ is the matrix defined in Eq. (10) and $\mathbf{f}_k \in \mathbb{R}^M$ is a scale function taking into account the difference between the formant frequency obtained at the k^{th} iteration, and the measured formant frequency:

$$\mathbf{f}_k = [f_{1_k}, f_{2_k}, \ldots, f_{m_k}, \ldots, f_{M_k}]^T, \tag{15}$$

where

$$f_{m_k} = \left[\frac{F_m - F'_{m_k}}{F'_{m_k}}\right]. \tag{16}$$

In Eq. (16), F_m is the m^{th} measured formant frequency, and F'_{m_k} is the m^{th} formant frequency obtained by the area function of the k^{th} iteration. The symbol ψ in Eq. (14) denotes a speed factor, greater than 1, enabling to enhance the iterative process. A value in the order of magnitude of 10 is usually good enough [8]. In this study, a value of $\psi = 15$ is used along the paper.

3.3. Estimation of vocal tract length

The estimation of the vocal tract length is based on the same technique. First, the variation of λ is computed by

$$\delta\lambda = \psi \mathbf{J}_\mathbf{l}^T \mathbf{f}_k, \tag{17}$$

where, after Eq. (4),

$$\delta\lambda = \left[\frac{l_{1_k}}{l_{1_{k+1}}} - 1, \frac{l_{2_k}}{l_{2_{k+1}}} - 1, \ldots, \frac{l_{N_k}}{l_{N_{k+1}}} - 1\right]^T. \tag{18}$$

The new length vector $\mathbf{l}_{k+1}$ is then

$$\mathbf{l}_{k+1} = \mathbf{\Lambda}_k \mathbf{l}_k, \tag{19}$$

where $\mathbf{\Lambda}_k \in \mathbb{R}^{N\times N}$ is a diagonal matrix, where the diagonal elements are such that $\mathbf{\Lambda}_k = \text{diag}\left(\frac{1}{1+\delta\lambda_1}, \frac{1}{1+\delta\lambda_2}, \ldots, \frac{1}{1+\delta\lambda_N}\right)$.

The iteration process continues until the L1-norm of $\mathbf{f}$ is less than a certain arbitrary threshold ϵ, namely until $\|\mathbf{f}_k\|_1 < \epsilon$.

Since the formant frequencies are estimated with a rough precision, it is not necessary to impose a very small value to ϵ. Setting $\epsilon = 1\%$ has been shown to be sufficient. A value less than 1 % increases the time computation and does not improve the estimation.

3.4. Constraints

Since the problem may be ill-posed, namely an infinite number of area functions may give the same formant frequency pattern, it should be regularized by imposing constraints on the solution. Beside the lip aperture constraint, explained in Section 4.1, the solution deals with two other constraints: one related to the potential energy and one related with the kinetic energy. The latter is important only when dealing with trajectory estimation in a dynamic configuration.

3.4.1. Potential energy constraint

This constraint attempts to avoid unrealistic area function configurations by constraining the area function not to deviate too far from the rest configuration, *i.e.* the area function given by the null articulatory vector of the Maeda model [16]. Let $\mathcal{V}_{art}$ be the articulatory potential energy (which should not be confused with the acoustical potential energy introduced in Eq. (6)), defined by

$$\mathcal{V}_{art} = \|\mathbf{a} - \mathbf{a}_0\|_2^2, \tag{20}$$

where $\mathbf{a}_0$ is the area function of the rest position, given by the null articulatory vector of the Maeda model [16]. Note that the articulatory potential energy expression for the length function is obtained by replacing $\mathbf{a}$ by $\mathbf{l}$ in Eq. (20).

The constraint term $\mathcal{C}_\mathcal{V}$ is then

$$\mathcal{C}_\mathcal{V} = \frac{\partial \mathcal{V}_{art}}{\partial \mathbf{a}} \mathcal{V}_{art}, \tag{21}$$

where

$$\frac{\partial \mathcal{V}_{art}}{\partial \mathbf{a}} = 2\left[\mathbf{a} - \mathbf{a}_0\right]. \tag{22}$$

3.4.2. Kinetic energy constraint

This constraint should be used only for inversion of dynamic configurations, namely to estimate an articulatory trajectory. This aims at reducing the difference between the area function at a given instant t and the area function at $t+1$. Let $\mathcal{T}_{art}$ be the articulatory kinetic energy (which should not be confused with the acoustical kinetic energy introduced in Eq. (5)) defined by

$$\mathcal{T}_{art}(t) = \|\Delta\mathbf{a}(t)\|_2^2, \tag{23}$$

where $\Delta\mathbf{a}(t) = \mathbf{a}(t+1) - \mathbf{a}(t)$ is the difference between two successive area functions at a given instant t. Similarly, the articulatory kinetic energy expression for the length function is obtained by replacing $\mathbf{a}$ by $\mathbf{l}$ in Eq. (23).

The constraint term $\mathcal{C}_\mathcal{T}$ is then

$$\mathcal{C}_\mathcal{T} = \frac{\partial \mathcal{T}_{art}}{\partial \mathbf{a}} \mathcal{T}_{art}, \tag{24}$$

where

$$\frac{\partial \mathcal{T}_{art}}{\partial \mathbf{a}} = \begin{cases} 2\Delta\mathbf{a}(t), & t = 1 \\ 2\left[\Delta\mathbf{a}(t) - \Delta\mathbf{a}(t-1)\right], & 2 \le t \le t_{max} - 1 \\ 2\Delta\mathbf{a}(t-1), & t = t_{max} \end{cases} \tag{25}$$

3.5. Dynamic configuration

When dealing with dynamic configurations, the algorithm presented in Section 3 is slightly modified. Indeed, temporal frames are simultaneously inverted. Eq. (14) becomes

$$\tilde{\mathbf{a}}_{k+1} = \tilde{\mathbf{a}}_k + \tilde{\mathbf{A}}_k \left[(1 - c_{kin} - c_{pot})\tilde{\mathbf{J}}_\mathbf{a}^T \tilde{\mathbf{f}}_k + c_{kin}\mathcal{C}_\mathcal{T} + c_{pot}\mathcal{C}_\mathcal{V}\right], \tag{26}$$

where

$$\tilde{\mathbf{a}} = [\mathbf{a}(0), \mathbf{a}(1), \ldots, \mathbf{a}(t), \ldots, \mathbf{a}(t_{max})]^T \tag{27}$$

$$\tilde{\mathbf{f}} = [\mathbf{f}(0), \mathbf{f}(1), \ldots, \mathbf{f}(t), \ldots, \mathbf{f}(t_{max})]^T \tag{28}$$

$$\tilde{\mathbf{J}}_\mathbf{a} = \text{diag}\,(\mathbf{J}_\mathbf{a}(0), \mathbf{J}_\mathbf{a}(1), \ldots, \mathbf{J}_\mathbf{a}(t), \ldots, \mathbf{J}_\mathbf{a}(t_{max})), \tag{29}$$

c_{kin} and c_{pot} are weighing coefficients less than 1 for kinetic and potential energy constraints.

Temporal frames are thus simultaneously inverted. In that case, the threshold $\tilde{\epsilon}$ should be proportional to the number of frames T, namely $\tilde{\epsilon} = T\epsilon$.

4. EXPERIMENTS

This section details the experimental protocol for acquiring input data of the algorithm and results.

4.1. Capture of the lip aperture

The lip aperture is estimated thanks to the markerless facial capture software *Faceshift*[1]. A range camera enables the depth image of the face to be dynamically captured, as shown in Figure 1. In this paper, the frame rate is 30 frame per seconds. The knowledge of the lip contour gives the area of the lip opening.

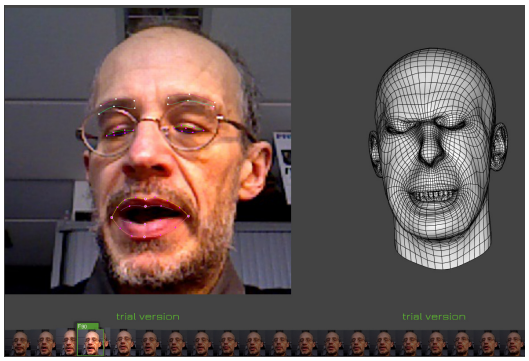


Fig. 1. Screen print image from the markerless facial capture software Faceshift. The depth and the position of the digitally marked points are given for each temporal frame.

4.2. Computing the initial function

The initial function is set to a neutral position, which is an area function corresponding to the articulatory position requiring the least effort from the speaker, with respect to the lip termination area estimated by the facial capture software. For that purpose, we use the articulatory model of Maeda [16], which uses 7 coefficients corresponding to the position of articulatory components (jaw position, tongue dorsum position, tongue dorsum shape, apex position, lip height, lip protrusion, and larynx height). Among these 7 parameters, only three articulatory deformation modes (jaw position, lip height, and lip protrusion) control the lip opening area. The corresponding parameters can be easily estimated by an inverse technique (such as Jacobian inverse technique [15], for instance). The initial function is then computed by setting the other parameters to 0, which is the rest position of the speaker.

4.3. Results

This section shows examples of inversion of french oral vowels produced by a native french speaker. Section 4.3.1 shows results for static configurations, while Section 4.3.2 displays the trajectory of the area function along a speech time segment corresponding to transition between two successive vowels. The chosen constant parameters are summarized in Table 1.

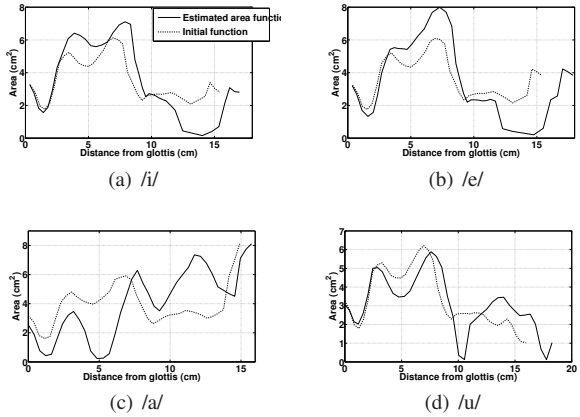
Parameter	Value	Parameter	Value
Number of tubes N	40	ψ	15
Number of formants M	3	ϵ_1 (%)	1
c_s (m.s^{-1})	343	c_{kin}	0.9
ρ (kg.m^{-3})	1.204	c_{pot}	10^{-3}

Table 1. Constant parameters chosen for the study

The formants of the recorded acoustic signal are estimated by a concurrent curve strategy approach using LPC (*Linear Predictive Coding*) [17]. The inversion technique is also valid with other formant technique estimation, such as cepstral coefficients based methods, for instance [9].

4.3.1. Static vowels

Figure 2 displays results obtained for static vowels. The estimated area functions (solid line) are in good agreement with the vocal tract shapes expected for these vowels. /i/ and /e/ present a wide back cavity and a narrow front cavity, while /a/ presents a rather narrow back cavity and a wide front cavity, with a wide mouth opening. Note that /a/ presents the smallest vocal tract length, which is again in agreement with expected shapes, since the lip aperture is very large and there is almost no lip protrusion. Finally, as expected, /u/ presents two cavities, separated by a constriction at the mid-point of the vocal tract. The length of the vocal tract inverted for /u/ is the longest since it uses a strong protrusion and a small lip aperture.



(a) /i/ (b) /e/

(c) /a/ (d) /u/

Fig. 2. Estimated area functions (solid line) of several static french vowels. The initial area function is represented by the dashed line.

The computation time for the different estimations spans a range from 0.95 s for /u/ to 4.73 s /a/. These very low computation times are a good step toward a real-time utilization. This could be achieved with a few improvements in the implementation of the algorithm.

4.3.2. Dynamic transition

Figure 3 shows the trajectory of the area function from /a/ to /u/. As expected, the length of the vocal tract increases during the transition from /a/ (large mouth opening, no protrusion) to /u/ (narrow mouth opening, strong protrusion). The transition between both phonemes

[1] http://www.faceshift.com/

can be clearly seen around the middle of the sentence: there is a sudden increase of the vocal tract length, of the mouth cavity, and a sudden narrowing of the constriction around 9 cm from the glottis.

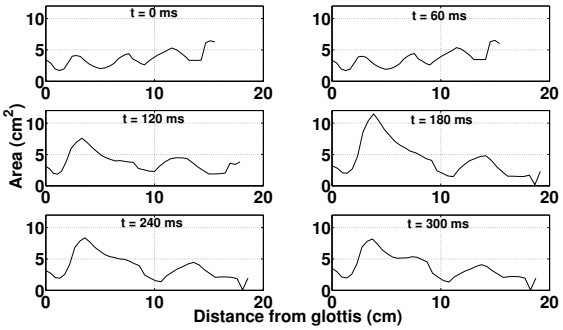


Fig. 3. Evolution of the estimated area function from /a/ to /u/. The time interval between each figure is 60 ms. The upper left figure corresponds to $t = 0$ ms, the bottom right to $t = 300$ ms. The computation time is 126 s.

Note that the computation time is considerably longer than for static vowels. This is due to the kinematic constraint, which tend to slower the algorithm. However, these constraints are necessary for accurate inversions.

5. CONCLUSIONS AND FURTHER WORKS

The method proposed in this paper performs a close estimation of the area and length functions of the vocal tract from the simultaneous acquisition of the speech acoustic signal and the lip opening area. In comparison with preexisting methods, the presented technique has the advantage to run efficiently without the need of neither an *a priori* learning of a database, nor a codebook search, which are time-consuming methods. It also easily adapts to any speaker, and does not require a fixed vocal tract length to be set arbitrarily. The estimation time is very low, less than 5 seconds for a static configuration, and therefore could be eventually implemented in a real-time estimation software. This is an important result for potential applications of the inversion, such as voice re-education, or foreign language learning for instance. However, it works only for oral vowels. For other phonemes, such as nasals or fricatives, a practical solution could be to compute the sensitivity matrices numerically.

The algorithm needs several relevant constraints in order to regularize the problem. For that purpose, we proposed to constrain:

- the lip termination area, by means of a facial capture device.
- the estimated area function to be as close as possible to an initial area function corresponding to the one requiring the least effort for the speaker (the rest position). This initial area function is derived from the Maeda articulatory model [16]. This constraint prevents the recovery of unrealistic area functions, by staying close to realistic physiological configurations. The initial function could be also adapted to the vowel to be inverted, by setting it to a typical vowel area function, for instance.
- the difference between two successive area functions in a dynamic configuration to be as small as possible. This constraints avoids unrealistic articulatory movements.

The main challenge of the technique is to adjust these different constraints. Indeed, their modification may change the estimated area function. Therefore, they should be appropriately chosen. A robust definition of the constraints is an important challenge to tackle in the future.

REFERENCES

[1] A. Soquet, M. Saerens, and P. Jospa, "Acoustic-articulatory inversion based on a neural controller of a vocal tract model," in *The ESCA Workshop on Speech Synthesis*, 1991, pp. 1–5.

[2] S. Hiroya and M. Honda, "Estimation of articulatory movements from speech acoustics using a HMM-based speech production model," *IEEE Trans. Speech Audio Proc.*, vol. 12(2), pp. 175–185, 2004.

[3] T. Toda, A. W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a gaussian mixture model," *Speech Communication*, vol. 50, no. 3, pp. 215–227, 2008.

[4] Z. Yu, "A method to determine the area function of speech based on perturbation theory," *STL-QPSR*, vol. 34(4), pp. 77–96, 1993.

[5] R. Carré, "From an acoustic tube to speech production," *Speech communication*, vol. 42, pp. 227–240, 2004.

[6] S. Ouni and Y. Laprie, "Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion," *J. Acoust. Soc. Am.*, vol. 118(1), pp. 444–460, 2005.

[7] S. Panchapagesan and A. Alwan, "A study of acoustic-to-articulatory inversion of speech by analysis-by-synthesis using chain matrices and the Maeda articulatory model," *J. Acoust. Soc. Am.*, vol. 129(4), pp. 2144–2162, 2011.

[8] K. Bunton, B. H. Story, and I. R. Titze, "Estimation of vocal tract area functions in children based on measurement of lip termination area and inverse acoustic mapping," in *Proceedings of meetings on acoustics*, 2013, vol. 19, pp. 1–8.

[9] J. Busset and Y. Laprie, "Acoustic-to-articulatory inversion by analysis-by-synthesis using cepstral coefficients," in *Proceeding of meetings on Acoustics*, 2013, vol. 19.

[10] P. Mermelstein, "Determination of the vocal-tract shape from measured formant frequencies," *J. Acoust. Soc. Am.*, vol. 41(5), pp. 1283–1294, 1967.

[11] B. Potard and Y. Laprie, "A robust variational method for the acoustic-to-articulatory problem," in *Interspeech, Brighton 2009*, 2009.

[12] G. Fant, "Vocal-tract area and length perturbations," *Roy. Swedish Academy of Music*, vol. 16(4), pp. 1–14, 1975.

[13] B. H. Story, "Technique for "tuning" vocal tract area functions based on acoustic sensitivity functions," *J. Acoust. Soc. Am.*, vol. 119(2), pp. 715–718, 2006.

[14] M. M. Sondhi and J. Schroeter, "A hybrid time-frequency domain articulatory speech synthesizer," *IEEE Trans. Acoust. Speech Sig. Process.*, vol. 35(7), pp. 955–967, 1987.

[15] S. R. Buss, "Introduction to inverse kinematics with jacobian transpose, pseudoinverse and damped least squares methods," *IEEE Journal of Robotics and Automation*, vol. 17, pp. 1–19, 2004.

[16] S. Maeda, "Un modele articulatoire de la langue avec des composantes linéaires," 1979, pp. 152–162.

[17] Y. Laprie, "A concurrent curve strategy for formant tracking," Jegu, Korea, Oct. 2004.