POST-SYNCHRONIZATION VIA FORMANT-TO-AREA MAPPING OF ASYNCHRONOUSLY RECORDED SPEECH SIGNALS AND AREA FUNCTIONS

J. Schoentgen* and S. Ciocea†

Laboratory of Experimental Phonetics, Institute of Modern Languages and Phonetics, CP110, Université Libre de Bruxelles
Av. F. D. Roosevelt, 50, B-1050 Brussels, Belgium.
Tel. +32 2 650 2010, Fax: 32 2 650 2007, E-mail: jschoent@ulb.ac.be

ABSTRACT

The article presents a method of post-synchronization which is the match, by means of formant-to-area mapping, of an area function model to a measured area function. The objective of post-synchronization is to compute a model which is as near as possible to a measured area function and whose eigenfrequencies are identical to the corresponding measured formant frequencies. Different types of acoustic models and constraints are examined. Results show that the best map is obtained in the case of a lossless acoustic model, corrected for lip radiation and wall vibration losses, and the minimization of the Euclidean distance between geometrically fitted and formant-to-area mapped area function models. The differences between measured and mapped area functions are gauged by means of the dynamic length warping distance.

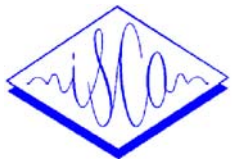
1. INTRODUCTION

Before taking measured vocal tract shapes and formant frequencies for granted, it is wise to confirm that the eigenfrequencies of the measured vocal tract shapes match the measured formant frequencies. Often, the calculated eigenfrequencies of measured vocal tract shapes and the corresponding measured formant frequencies do not agree for the following reasons. 1) With present technology it is difficult to obtain accurate measurements of 3-dimensional vocal tract shapes. In the case of magnetic resonance imaging (MRI), the problems are imaging artifacts related to the air-water boundary and the low concentration of hydrogen in the teeth and bone [7]. 2) It is difficult to record a speech signal via a microphone at the same time as a tract shape via MRI. The acoustic signal and the corresponding vocal tract area function or mid-sagittal profile must be recorded asynchronously. This is a consequence of the intense magnetic fields and noise associated with MRI. 3) The physical behavior of the vocal tract boundaries (i.e. wall vibration and sound radiation at the glottis and lips) is not known exactly, and the acoustic model which describes the sound propagation inside the vocal tract is therefore an approximation only. Furthermore, the use of a 1-dimensional wave propagation model (the Webster equation) may be another source of imprecision, especially for eigenfrequencies higher than 4 kHz. 4) Even if accurate measurements of 3-dimensional vocal tract shapes and an exact description of the sound propagation inside the vocal tract were obtained, area function models with a low number of parameters would still bring about modeling inaccuracies which may give rise to formant frequency values that differ from those measured.

A post-synchronization method has been developed to minimize the difference between measured and calculated formant frequencies. It uses formant-to-area mapping to arrive at an area function model which is as near as possible to a measured area function and whose first three eigenfrequencies are equal to the corresponding measured formant frequencies. The agreement between measured frequencies and the eigenfrequencies of the inferred (via acoustic-to-area mapping) area function model is better than 0.01 Hz. The difference between measured and inferred area functions is estimated by means of dynamic length warping. Results show that the smallest distances are achieved when acoustic losses at the tract boundaries are take into account and the Euclidean distance

*National Fund for Scientific Research, Belgium

†Action de Recherche Concertée, Communauté Française de Belgique



EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

between acoustically mapped and geometrically fitted area function models is kept as small as possible.

One must distinguish between measured area functions (via MRI, for instance), area function models geometrically fitted to measured ones (via a least-squares criterion, for instance) and area function models inferred via formant-to-area mapping. Here, the differences between the latter two are gauged by means of an Euclidean distance, and the differences between the first two by means of dynamic length warping.

2. FORMANT-TO-AREA MAPPING

Formant-to-area mapping is the inference of the shape of a vocal tract model via observed formant frequencies. Here, it consists of the direct calculation of the time derivatives of the cross-sections and the length of a vocal tract model so that the time derivatives of the observed formant frequencies and the model's eigenfrequencies match [5, 4, 6]. The vocal tract model is a concatenation of cylindrical tubelets. Time derivatives of the tubelet cross-sections are obtained by solving a linear algebraic system of equations. The derivatives are then numerically integrated to arrive at the cross-section movements. Since more than one area function is compatible with the observed formant frequencies, constraints are applied to the area function movement to select a unique solution.

3. CORPUS AND METHODS

The corpus was the first three formant frequencies and the corresponding area functions of twelve American English resonants (vowels and laterals) asynchronously measured via MRI by Story et al. [7].

The area function model consisted of a concatenation of 8 cylindrical tubelets that had lengths equal to $L/10$, $L/15$, $2L/15$, $L/5$, $L/5$, $2L/15$, $L/15$, $L/10$ respectively, where L was the distance from the glottis to the lips [3]. This choice was the outcome of a previous study according to which this model fitted the measured area functions geometrically better than the corresponding model consisting of 8 tubelets of equal lengths. The cross-section area of the first tubelet (adjacent to the glottis) was fixed at 2 cm^2 [5, 4]. The areas of the other seven cross-sections were variable and the total length, L, depended via formula (1) on the cross-section area of lip tubelet A_8. The symbol cm means that length L was measured in centimeters and cross-section A_8 in square centimeters.

$$\frac{L}{cm} = 20 - 0.5\frac{A_8}{cm^2}. \qquad (1)$$

Formula (1) was obtained by means of regression analyses of published data [7].

The total length of the vocal tract model is a parameter which simultaneously affects all the eigenfrequencies, whereas individual cross-sections mainly influence one or two. However, an unconstrained tract length would have taken on unrealistic values since from a mathematical point of view, length changes were the most efficient way of effecting formant changes. Therefore, any changes in total vocal tract length were made to depend on the cross-section of lip tubelet A_8 [6]. Consequently, the area function model had 7 independent control parameters.

Radiation at the lips and wall vibration losses were taken into account within the wave propagation model. This choice had been the outcome of a preliminary study during which three models of acoustic losses were compared [1]. The lossless acoustic model, corrected for wall vibration and lip radiation, was preferred because it was the simplest, and formant frequencies were mainly affected by these types of losses.

Post-synchronization consisted of determining, for each phonetic segment of the corpus, the cross-sections of the area function model by means of formant-to-area mapping.

The initial length (before formant-to-area mapping) of the model was equal to the measured tract length. The initial cross-sections were those obtained via the geometric fit of the model to the measured area function.

The mapping was constrained via one of four conditions. (i) The Euclidean distance between acoustically mapped and geometrically fitted area function models was kept as small as possible. (ii) The time derivatives of the cross-section movements were minimized. (iii) & (iv) Linear combination Ω^2(i)+(ii) was reduced to a minimum. Weight Ω was equal to 20 or 30 respectively.

A total of eight matching experiments was carried out by combining the four constraints (i)-(iv) with the lossless acoustic model or lossless model corrected for lip radiation and wall vibration. The agreement between the measured and model-generated formant frequencies was better than 0.01 Hz for all the phonetic segments and types of constraints.

To compare the effects of different acoustic models and constraints, the distances were computed between acoustically inferred and MRI-measured area functions. A possible problem was that these were of different lengths and their numbers of concatenated tubelets were different.

A method to deal with the alignment mismatch was dynamic length warping. Dynamic length warping is a mathematical technique which applies an optimum non-linear length scale distortion to achieve a best match at all points. The two area functions to be compared were inputted. The output was the length of the best path aligning the two.

The procedure was the following. When two area

	Lossless model				Lossy model			
	C(1)	C(2)	C(3)	C(4)	C(1)	C(2)	C(3)	C(4)
[i]	3.61	3.29	3.31	3.25	2.95	3.38	3.06	3.15
[ɪ]	2.26	2.59	2.36	2.43	2.18	2.55	2.28	2.35
[ɛ]	1.97	2.43	2.12	2.06	1.90	2.54	1.96	2.09
[æ]	2.02	3.11	2.70	2.94	2.01	3.09	2.22	2.82
[ʌ]	3.24	3.30	3.34	3.31	3.07	3.16	3.10	3.17
[ɑ]	3.04	3.60	3.30	3.44	2.95	3.84	3.29	3.43
[ɔ]	3.86	5.51	5.04	5.29	3.80	5.18	4.71	4.93
[o]	4.85	6.20	4.88	5.58	4.71	5.84	5.44	5.89
[ʊ]	3.41	3.49	3.46	3.47	3.39	3.42	3.41	3.41
[u]	6.55	6.97	6.78	6.82	6.08	6.28	6.09	6.14
[ɝ]	3.22	3.52	2.99	3.01	3.11	3.97	3.33	3.54
[l]	3.45	6.72	4.05	5.44	3.42	6.71	4.58	5.33

Table 1: The dynamic length warping distances [in cm] between the square roots of measured and inferred area functions, Story et al.'s data. C(i), i=1,2,3,4 designates the type of constraint.

functions were compared, that were composed of n_m and n_c cylindrical tubelets respectively, local distance $d_{ij} = |\sqrt{A_i} - \sqrt{a_j}|$ between tubelet cross-sections A_i and a_j of the measured and computed vocal tract shapes was calculated, $i = 1 \ldots n_m$ and $j = 1 \ldots n_c$. Cumulative distance D_{ij} along the optimum path from the beginning at the glottis to tubelets i and j was as follows:

$$D_{ij} = \sum_{x,y=1,1 \ along \ the \ best \ path}^{i,j} d_{xy}, \qquad (2)$$

which is assumed to be equivalent to:

$$D_{ij} = d_{ij} + \min\left(D_{i-1j}, D_{i-1j-1}, D_{ij-1}\right). \qquad (3)$$

The procedure starts with $D_{11} = d_{11}$, where all the possible paths begin. The value obtained for $D_{n_c n_m}$ was the score of the best path aligning the two area functions [2].

4. RESULTS

The dynamic warping distances between the mapped and measured area functions were calculated for each of the 12 resonants of Story et al.'s corpus and for each of the eight experiments. The results are presented in Tables 1 and 3. They show that the smallest distances were achieved when losses at the boundaries were taken into account and constraint (i) was used. The minimization of cross-section speeds either alone (ii), or in combination with constraint (i), was not strong enough to guarantee the smallest possible distances between the modeled and measured area functions.

Table 2 presents the vocal tract shapes arrived at via the post-synchronization of vowel [ɑ] of Story et al.'s corpus for the two acoustic models (lossless and corrected lossless) and the four constraints.

5. REFERENCES

[1] S. Ciocea. *Semi-analytic formant-to-area mapping.* PhD thesis, Université Libre de Bruxelles, Brussels, Belgium, 1997.

[2] J.N. Holmes. *Speech Synthesis and Recognition.* Van Nostrand Reinhold, UK, 1988.

[3] M. Mrayati, R. Carré, and B. Guérin. Distinctive regions and modes: A new theory of speech production. *Speech Comm.*, 7:257–286, 1988.

[4] J. Schoentgen and S. Ciocea. Direct calculation of the vocal tract area function from measured formant frequencies. In *Eurospeech*, volume 1, pages 745–748, 1995.

[5] J. Schoentgen and S. Ciocea. Kinematic acoustic-to-geometric mapping. In *ICPhS*, volume 2, pages 194–197, 1995.

[6] J. Schoentgen and S. Ciocea. Kinematic formant-to-area mapping. *Speech Comm.*, 4, 1997.

[7] B.H. Story, I.R. Titze, and E.A. Hoffman. Vocal tract area functions from magnetic resonance imaging. *J. Acoust. Soc. Amer.*, 100:537–554, 1996.

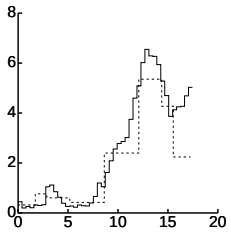
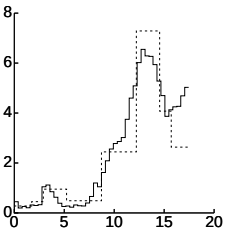
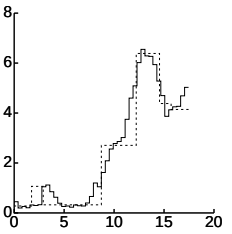
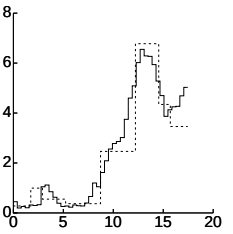
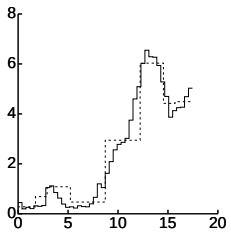
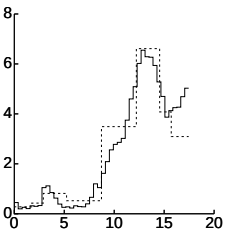
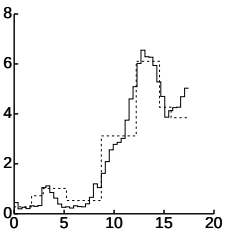
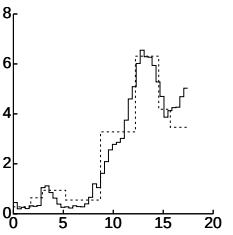
	Constraint (1)	Constraint (2)	Constraint (3)	Constraint (4)
Lossless model				
Lossy model				

Table 2: Measured (solid lines) and mapped (dashed lines) vocal tract shapes corresponding to the American English vowel [ɑ] of the Story et al. corpus. The vertical axes represent the vocal tract area function (in cm^2), and the horizontal axes the position (in cm), with 0 corresponding to the glottis.

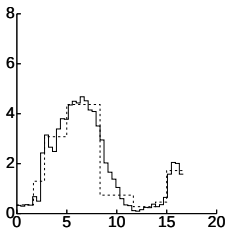
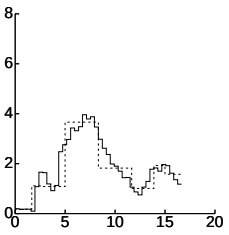
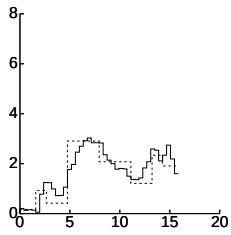
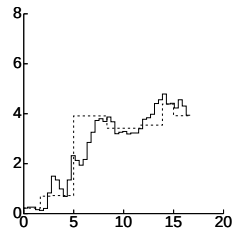
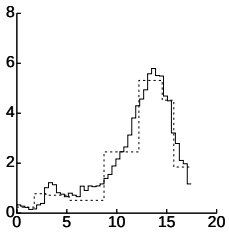
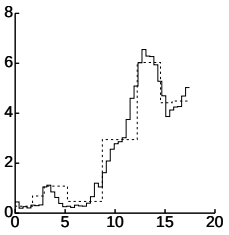
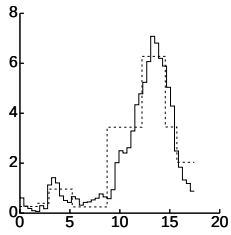
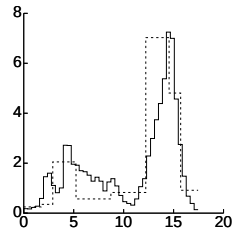
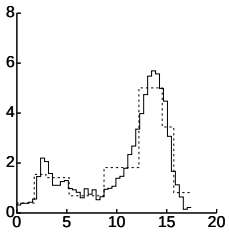
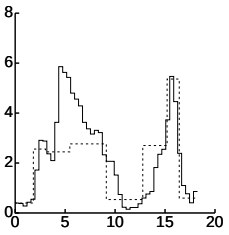
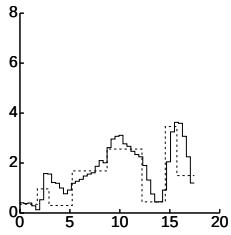
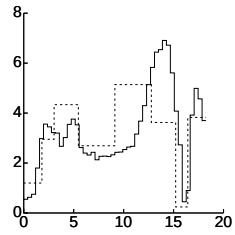
[i] 	[ɪ] 	[ɛ] 	[æ] 
[ʌ] 	[ɑ] 	[ɔ] 	[o] 
[ʊ] 	[u] 	[ɝ] 	[l] 

Table 3: Measured (solid lines) and mapped (dashed lines) vocal tract shapes corresponding to the 12 American English resonants of Story et al.'s corpus. The mapped shapes displayed give rise to the smallest dynamic length warping distances and correspond to the boxed data in Table 1. The vertical axes represent the vocal tract area function (in cm^2), and the horizontal axes the position (in cm), with 0 corresponding to the glottis.