ESTIMATION OF A PHYSICAL MODEL OF THE VOCAL FOLDS VIA DYNAMIC PROGRAMMING TECHNIQUES

E. Marchetto[1], F. Avanzini[1], and C. Drioli[2]
[1] Dept. of Information Engineering, University of Padova, Italy
[2] Dept. of Computer Science, University of Verona, Italy

Abstract: **This work presents a procedure for the estimation of a two-mass vocal fold model starting from a time-varying target flow signal. The model is specified by a large number of physical parameters, computed as functions of four articulatory parameters (three laryngeal muscle activations and subglottal pressure). Flow waveforms synthesized by the model are characterized by means of a set of typical voice source quantification acoustic parameters. Given a sequences of target acoustic parameters, dynamic programming techniques and interpolation based on Radial Basis Function Networks are used to derive sequences of articulatory parameters that lead to resynthesis of the target signal.**

Keywords: **Voice source, Low-dimensional models, Estimation, Synthesis**

I. Introduction

One open problem in research on low-dimensional vocal fold physical models is the relationship between parameters of the models and acoustic parameters related to voice quality. A recent work [1] studied the sensitivity of acoustic flow parameters to variation of physical parameters in a two-mass model, and provided indications of the "actions" that the model employs to target different voice qualities. However low-level parameters (masses, spring stiffnesses, etc.) are not independently controlled by a speaker: more physiologically motivated control spaces are needed. A related issue is the "inverse problem", i.e. the problem of estimating the time-varying control parameters to be used as input to the physical model in order to resynthesize a target acoustic signal. This involves inversion of a non-linear dynamical system with a large number of parameters. Moreover the solution is in principle non-unique. A possible solution to the non-uniqueness problem is working on temporal sequences of acoustic frames and estimating articulatory parameters through minimization of some cost function that includes an "articulatory effort" component. This approach has been applied in [2] to the solution of the inverse problem for an articulatory vocal tract model.

This paper presents a procedure for the estimation of a two-mass vocal fold model [3] starting from time-varying acoustic parameters of a target flow signal. The model is specified by a large number of low-level physical parameters. An additional modeling layer computes these physical parameters as functions of four articulatory parameters (three activation levels of laryngeal muscles and subglottal pressure) [4]. Glottal flow waveforms synthesized by the model are characterized by means of a set of acoustic parameters: fundamental frequency F_0, open quotient *OQ*, speed quotient *SQ*, return quotient *RQ*, normalized amplitude quotient *NAQ* [5], etc., that are used in the literature as typical voice source quantification parameters [6].

Therefore there are three related but distinct spaces of parameters: articulatory, physical, and acoustic parameters. This work deals with the problem of mapping acoustic into articulatory parameters. We tackle the problem by characterizing temporal frames of glottal flow signals via sequences of acoustic parameters, and by developing a methodology to derive the corresponding sequences of articulatory parameters using dynamic programming techniques. The procedure is further improved by using Radial Basis Function Networks (RBFN) to interpolate points in the articulatory space. Results show that the physical model controlled via the estimated parameters is able to resynthesize target flow signal with good accuracy.

Section II describes the physical model used in this work while Sec. III details the techniques used to estimate the model starting from a target time-varying flow signal. Results, as well as and current limitations and shortcomings of the proposed approach, are discussed in Sec. IV

II. The physical model

The analysis developed in the next sections is based on a two-mass model presented in [3] and depicted in Fig. 1. The model assumes in particular one-dimensional, quasi-stationary, frictionless and incompressible flow from the subglottal region up to a time-varying *separation point* z_s along the glottis, where flow separation and free jet formation occurs. No pressure recovery is assumed at the glottal exit. The separation point z_s is predicted in [3] to occur when the glottal area $a(z)$ exceeds the minimum area by a given amount ($10-20$%). By introducing a *separation constant* s (in the range $1.1-1.2$), separation occurs when the glottal area takes the value $a_s = \min(sa_1,\ a_2)$.

The vocal tract is modeled as an inertive load. In the limit of fundamental frequencies much lower than the first formant frequency the air column acts approximately as a mass that is accelerated as a unit, and the vocal tract input pressure can be written as $p_v(t) = Ru(t) + I\dot{u}(t)$, where

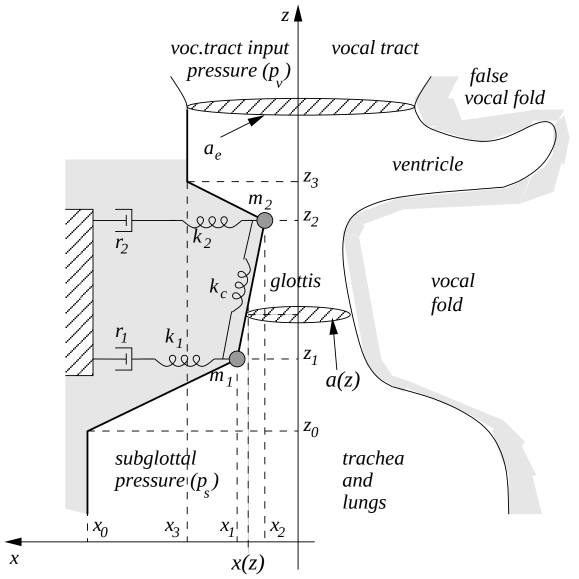


Fig. 1. Right: schematic diagram of the vocal fold, trachea, and supraglottal vocal tract; left: two-mass vocal fold model.

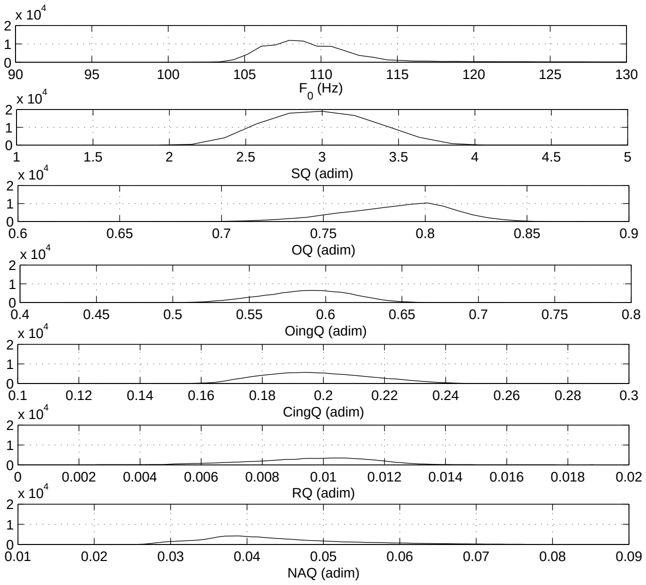


Fig. 2. Distribution of acoustic parameters in the direct codebook.

R, I are the input resistance and inertance, respectively. Values for R, I are chosen from [7]. Being a first-order system, this model does not account for resonances of the vocal tract, however it describes with sufficient accuracy its most relevant effects on vocal fold oscillation, in particular the lowering of the oscillation threshold pressure [7].

Low-level physical parameters (masses, spring stiffnesses, etc.) are not independently controlled by a speaker: more physiologically motivated control spaces are needed, which requires to establish a mapping between physiology (muscle activations) and physics (parameters of the two-mass model). A set of empirical rules, derived from [8], was used in [4] for controlling a two-mass physical model. The rules link vocal fold geometry to activation levels of three muscles: cricothyroid (a_{CT}), thyroarytenoid (a_{TA}) and lateral cricoarytenoid (a_{LC}). These levels are assumed to be normalized in the $[0, 1]$ range. In addition, in this paper we also consider the subglottal pressure p_s. In conclusion, the physical model is completely controlled by the set of four *articulatory parameters* $a_{CT}, a_{TA}, a_{LC}, p_s$.

III. Model estimation

A. An articulatory codebook

The first step of the estimation procedure is to define and populate a *direct codebook*, in which every vector of articulatory parameters $a_{CT}, a_{TA}, a_{LC}, p_s$ is a "key" and is associated with one and only one vector of acoustic parameters. To this aim, a large number of numerical simulations of the two-mass model is run on a dense grid of vectors of acoustic parameters. For each simulation, relevant acoustic parameters are extracted from the synthesized glottal flow signal using the APARAT toolkit [9].

The direct codebook used in this work has been derived on a grid where a_{CT} and a_{TA} vary in the range $0 \div 1$ with a fixed step of 0.05, while the range for a_{LC} is $0.25 \div 0.5$ with a fixed step of 0.025 (because sustained phonation only occurs within this region), and p_s varies in the range $500 \div 1500$ Pa with a fixed step of 50 Pa. The resulting codebook contains 86125 vector pairs. Fig. 2 shows the distribution of the 7 computed acoustic parameters in the direct codebook.

B. Codebook inversion and dynamic codebook access

In order to solve the inverse problem, the direct codebook has to be inverted to obtain the *inverse codebook*. This however suffers from a non-uniqueness problem, i.e. an acoustic vector can be the key to one or *more* articulatory vectors. We tackle the problem by working on temporal *sequences* of acoustic vectors, rather than on a single vector. These may be obtained e.g. by analyzing a time-varying glottal flow signal on a frame-by-frame basis. Given a sequence of acoustic vectors x_k we want to obtain an "optimal" sequence of articulatory vectors v_k^j in the inverse codebook: as already explained, x_k is in principle associated with many candidate vectors v_k^j because of the non-uniqueness problem. In particular we perform a search in the acoustic space of the inverse codebook to find the nearest vectors (according to the euclidean distance) to the given x_k; the v_k^j are therefore the articulatory vectors associated to these nearest vectors in the codebook.

The optimal sequence of articulatory parameters is obtained by minimizing a *cost function* with three terms. An *acoustic* term accounts for the euclidean distance between x_k and its discretized versions in the acoustic space of the codebook (the vectors found by the search). An *articulatory* term minimizes the euclidean distance between v_k^j and v_{k-1}^j, i.e. between every two articulatory vectors *consecutive in time*. This is the key term in the procedure, in order to obtain smooth parameter variations: it minimizes the "articulatory effort", in accordance with the physiological muscle behavior. An *accumulation* term extends the cost function domain to the entire input sequence, so that the obtained articulatory sequence is optimal in a global way. The (simplified) cost function is:

$$f(v_k^j) = \min_{\gamma,\delta}[\tau_1||x_k - c_k^\delta||^2 + \tau_2||v_k^j - v_{k-1}^\gamma||^2 + f(v_{k-1}^\gamma)]$$

where $\tau_{1,2}$ are weights for the acoustic and articulatory terms, respectively; c_k^δ are the discretized acoustic vectors close to x_k. Dynamic programming techniques are the ideal tool for the minimization of the cost function: in particular the accumulation term would lead to exponential complexity, if not computed with this approach.

C. Codebook clustering and interpolation with RBFNs

One problem in the proposed procedure is that a target vector x_k is typically not present in the inverse codebook, which is discrete; therefore every found v_k^j is not associated with x_k, but only with a vector near to x_k. The limitations of the discrete codebook can be overcome by interpolating the articulatory space; this allows to compute articulatory vectors associated exactly to the given x_k.

The interpolation uses RBFNs (Radial Basis Function Networks) [10]. Since RBFNs only interpolate functions and cannot handle multimaps, the inverse codebook has to be manipulated and the non-uniqueness problem avoided. We have developed a novel algorithm that subdivides the codebook in acoustic clusters and articulatory subclusters. Every cluster is associated to one or more subclusters. The algorithm guarantees that for every acoustic vector in a given cluster there will be only one (or none) articulatory vector in each associated subcluster. As a result in every subcluster the subdivided codebook provides a unique mapping, which is needed for RBFNs to work properly.

The algorithm first subdivides the acoustic space in clusters C_i using a standard technique. Random vectors, as many as the desired clusters, are generated and subsequently moved with an iterative procedure [11] to become centroids. Centroids are iteratively displaced in such a way that the sum of the distances between every centroid and the associated vectors is minimized. Clusters C_i are built by associating every acoustic vector with the nearest centroid. In order to obtain a uniform distribution of vectors in every cluster, the iterative procedure is applied in a two-stage fashion. Moreover, in order to ensure a certain degree of overlapping, the vectors which are closest to boundaries between two clusters are replicated in both.

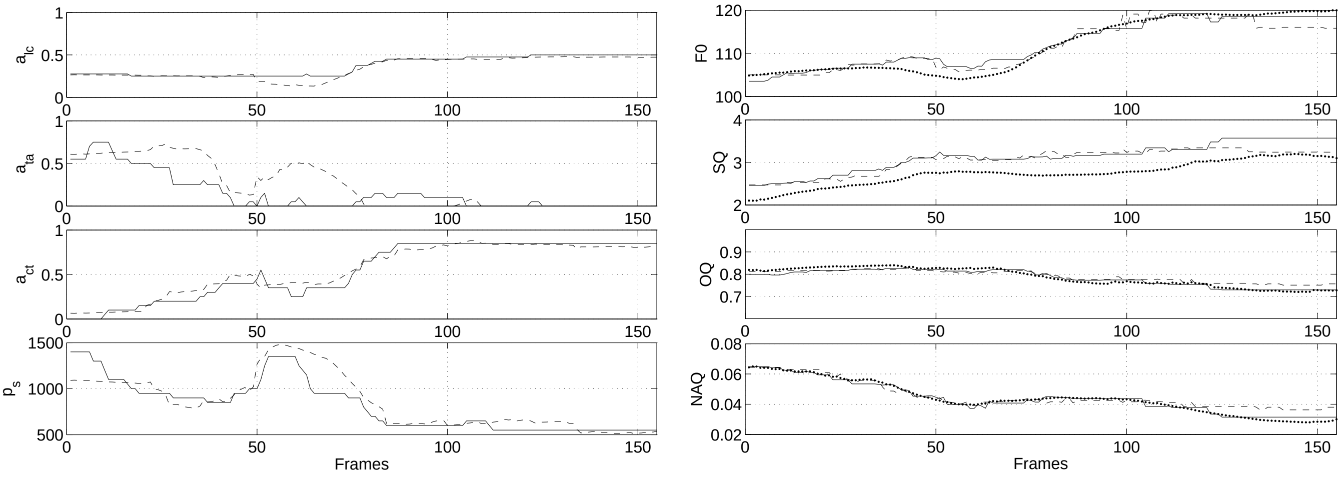
Once the acoustic clusters C_i are built, the algorithm determines the s articulatory subclusters S_j^i ($j = 1 \dots s$) associated to each C_i. Here s equals the maximum number of articulatory vectors associated to the same acoustic vector x^* in C_i. Every articulatory vector associated with x^* is assigned to a distinct subcluster and used as a "seed". The remaining articulatory vectors are allocated as follows. When many articulatory vectors v_k^j are associated to the same acoustic vector x_k, every v_k^j is assigned to a different subcluster, chosen as the one with the nearest *articulatory* centroid. The location of the subcluster centroid is updated after every new vector is added.

Having determined the clusters C_i, each associated with one or more subclusters S_j^i, within every S_j^i we construct four different RBFNs to interpolate each dimension of the articulatory space. Every acoustic vector associated to the subcluster is used as center for one RBF (gaussian functions in our application). Values for the parameters of the functions (standard deviation, etc.) are found after an extensive set of experiments on the codebook. After the determination of all the RBFNs, the articulatory space can be interpolated. The following procedure is used to feed the dynamic programming with interpolated vectors. Given an acoustic vector we find the k nearest acoustic clusters and all the associated subclusters. The acoustic vector is used as input for the set of RBFNs in each subcluster. Finally, all the computed interpolated articulatory vector (as many as the subclusters) are passed to the dynamic programming procedures, which proceeds with the optimization.

IV. Results and discussion

The proposed algorithms were initially tested and tuned using artificial target sequences of acoustic vectors. These were used as input to the system to obtain the corresponding articulatory parameters. Results from these preliminary tests provided two main indications. First, the synthetic signals obtained by driving the physical model with the derived articulatory parameters follow closely the target acoustic vectors. Second, the derived muscular activations and subglottal pressure have physiologically plausible evolutions, i.e. they have smooth variations in time. These initial results confirm the validity of the employed cost function, and of the RBFN interpolation.

In order to test the proposed algorithms on real signals, we have realized a complete *synthesis-by-analysis* procedure. Starting from a recorded utterance (a sustained vowel with varying pitch and voice quality) the signal is inverse filtered with APARAT. The estimated glottal flow is analyzed frame-by-frame and a sequence of acoustic vectors is obtained. The corresponding articulatory vectors (derived using the techniques described in Se. III) are used to drive the physical model, and the resynthesized glottal flow is convolved with the time-varying formant filter of the vocal tract. The final result is a resynthesis of the utterance, in which the evolution of pitch and voice quality are close to those of the original signal.



(a)

(b)

Fig. 3. Example of the analysis-by-synthesis procedure. (a) Time sequences of articulatory parameters retrieved by the optimization procedure (solid line: no RBFNs; dashed line: RBFNs). (b) Time sequences of glottal flow acoustic parameters (dotted line: target sequences extracted from a recorded utterance; solid line: resynthesis without RBFNs; dashed line: resynthesis with RBFNs).

Fig. 3 shows the performance of the synthesis-by-analysis procedure on a real utterance (a sustained /e/). The time-varying acoustic vectors obtained in the resynthesis follow with good accuracy the target ones, and informal listening tests confirm that the resynthesis is qualitatively similar to the target signal. In particular the NAQ is usually well followed, as shown in Fig. 3(b). This is a positive result as the NAQ is known to be strongly related to voice quality [5]. The effect of using RBFNs can be noticed in Fig. 3(a): the sequences of articulatory vectors interpolated by RBFNs are smoother than those obtained using bare dynamic programming. A second advantage of using RBFNs is that the amount of vectors that feeds the dynamic programming procedure is significantly reduced and this leads to a corresponding decrease in the computation time.

While the results reported in this work indicated that the proposed approach is effective in estimating control parameters of the physical model, both with synthetic target data and with real utterances, a number of limitations are still hindering the performance of the estimation procedure described in this work. These are mainly related to intrinsic limitations of the two-mass model. Ranges of variation for the acoustic parameters are generally narrow (see Fig. 2), and are sometimes non realistic. RQ and NAQ in particular assumes exceedingly low values, due to poor description of the flow at small glottal apertures, which results in abrupt glottal closure and exceedingly high absolute values of the flow derivative peak. The relationship between physical parameters of the models and acoustic parameters also need to be assessed: as an example, the relation between p_s and F_0 observed in the model is not in accordance with results reported in the literature. Finally, a more systematic approach to the determination of RBFNs parameters is needed in order to fully exploit the benefits of interpolation in the codebook.

References

[1] D. Sciamarella and C. D'Alessandro, "On the acoustic sensitivity of a symmetrical two-mass model of the vocal folds to the variation of control parameters," *Acta Acustica united with Acustica*, vol. 90, no. 4, pp. 746–761, Jul. 2004.

[2] J. Schroeter and M. Sondhi, "Speech coding based on physiological models of speech production," in *Advances in Speech Signal Processing*, S. Furui and M. Sondhi, Eds. New York: Dekker, 1992, pp. 231–263.

[3] N. J. C. Lous, G. C. J. Hofmans, R. N. J. Veldhuis, and A. Hirschberg, "A symmetrical two-mass vocal-fold model coupled to vocal tract and trachea, with application to prosthesis design," *Acta Acustica united with Acustica*, vol. 84, pp. 1135–1150, 1998.

[4] F. Avanzini, S. Maratea, and C. Drioli, "Physiological control of low-dimensional glottal models with applications to voice source parameter matching," *Acta Acustica united with Acustica*, vol. 92, no. Suppl.1, pp. 731–740, Sep. 2006.

[5] P. Alku, T. Bäckström, and E. Vilkman, "Normalized amplitude quotient for parametrization of the glottal flow," *J. Acoust. Soc. Am.*, vol. 112, no. 2, pp. 701–710, Aug. 2002.

[6] P. Alku and E. Vilkman, "A comparison of glottal voice quantification parameters in breathy, normal and pressed phonation of female and male speakers," *Folia Phoniatr. Logop.*, vol. 48, no. 5, pp. 240–254, Sep. 1996.

[7] I. R. Titze and B. H. Story, "Acoustic interactions of the voice source with the lower vocal tract," *J. Acoust. Soc. Am.*, vol. 101, no. 4, pp. 2234–2243, Apr. 1997.

[8] ——, "Rules for controlling low-dimensional vocal fold models with muscle activation," *J. Acoust. Soc. Am.*, vol. 112, no. 3, pp. 1064–1027, Sep. 2002.

[9] M. Airas, H. Pulakka, T. Bäckström, and P. Alku, "A toolkit for voice inverse filtering and parametrisation," in *Proc. 9th European Conf. on Speech Communication and Technology (Interspeech'2005 - Eurospeech)*, Lisbon, Sep. 2005, pp. 2145–2148.

[10] T. Poggio and F. Girosi, "Networks for approximation and learning," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1481–1497, Sep. 1990.

[11] A. Gercho and R. M. Gray, *Vector quantization and signal compression*, ser. The Kluwer international series in engineering and computer science. Kluwer, 1992, boston.