ARTICULATORY STATE ESTIMATION FROM INCOMPLETE SPEECH MEASUREMENTS

Luis Rodrigues[1] and John Kroeker[2]
[1]Dept. of Mechanical and Industrial Engineering, Concordia University, Montreal, Canada, luisrod@encs.concordia.ca
[2]Eliza Corporation, Beverly, MA, USA, jkroeker@elizacorp.com

ABSTRACT

This paper presents a method to find the unknown components of the state of articulators that produce a given speech signal. The proposed method reconstructs the unknown states such that the energy needed to produce the recorded speech is minimized.

1. Introduction

Although increasingly sophisticated models for human speech production have been developed in the past thirty years [1], they have made little impact on the progress of computer synthesized human speech. From a practical point of view, these models are either too complicated to implement, or lack the comprehensiveness in covering all classes of sounds, or both. On the other hand, the current approach used in commercial speech synthesizers cannot provide the natural transitions between phonemes as the human articulatory system does. However, two important computational models in speech synthesis have led to promising results: the Haskins Laboratory computational model [2] and the University of Waterloo nonlinear observation articulatory dynamical model [3], with close connections to the linear observation hidden dynamic model (HDM) from the University of Edinburgh [4].

Traditionally, human speech has been seen as having two structures: one considered physical, the other cognitive. The relation between the two structures is generally not an intrinsic part of either description. Articulatory Phonology was first suggested by [5] at the Haskins Laboratory, and it offers the different assumption that these apparently different domains are, in fact, the low and high level description of a single complex feedback system. Crucial to this approach is the identification of phonological units with dynamically specified units of articulatory action, called *gestures*. Thus, an utterance is described as an act that can be decomposed into a small number of primitive units, with a particular spatio-temporal configuration. Examples of gestures for different words can be found in [1]. This paper uses the dynamical model from [3] (based on the articulatory theory of speech production described in [1]), and proposes a method that reconstructs the unknown articulatory states such that the energy needed to produce a given speech signal is minimized.

2. Problem Formulation

We assume a mathematical model for speech production consisting of a discrete-time linear and time-invariant dynamical system [3]. The state then evolves according to

$$x_{k+1} = \Phi x_k + w_k \tag{1}$$

with $x \in R^n$ being the state vector of positions of the different articulators and $w \in R^m$ being the forcing input vector. We furthermore assume that the state vector is not completely known. In other words, at each instant k of discrete-time, some components of x_k are not known and some components of x_{k+1} are also not known. We will use the following notation: an upper index in a variable corresponds to the row index when the variable is a component of a vector or matrix; a lower index corresponds to the time index when the variable is a component of a vector and to the column index when the variable is an entry of a matrix. Using this notation, suppose that, without loss of generality, for a given instant of discrete-time k, the component l of x_k (x_k^l) and the component m of x_{k+1} (x_{k+1}^m) are unknown. The problem then, for each instant k, is to compute the unknown components of both x_k and x_{k+1} such that the input energy

$$J = w_k^T w_k \tag{2}$$

is minimized. In other words, the problem to be solved is

$$\min_{x_k^l,\, x_{k+1}^m,\, l\in I_l,\, m\in I_m} \quad w_k^T w_k$$
$$s.t \qquad (1)$$

3. Problem Solution

To solve this problem we start by noting that from (1) the forcing input vector w_k is

$$w_k = x_{k+1} - \Phi x_k \tag{3}$$

and the functional (2) can then be rewritten as

$$J = \left(x_{k+1} - \Phi x_k\right)^T \left(x_{k+1} - \Phi x_k\right) \tag{4}$$

Then, the solution that minimizes (4) must verify

$$\frac{\partial J}{\partial x_k^l} = 0, \quad \frac{\partial J}{\partial x_{k+1}^m} = 0. \tag{5}$$

From (4), the first condition leads to

$$\frac{\partial J}{\partial x_k^l} = 0 \Leftrightarrow 2\left(x_{k+1} - \Phi x_k\right)^T \frac{\partial}{\partial x_k^l}\left(x_{k+1} - \Phi x_k\right) = 0, \tag{6}$$
$$\Leftrightarrow \left(x_{k+1} - \Phi x_k\right)^T \left(-\Phi_l\right) = 0$$

where Φ_l is the column l of matrix, i.e,

$$\Phi_l = \left[\Phi_l^1 \quad \cdots \quad \Phi_l^n\right]^T.$$

Using (4), the second condition from (5) leads to

$$\frac{\partial J}{\partial x_{k+1}^m} = 0 \Leftrightarrow 2\left(x_{k+1} - \Phi x_k\right)^T \frac{\partial}{\partial x_{k+1}^m}\left(x_{k+1} - \Phi x_k\right) = 0 \tag{7}$$
$$\Leftrightarrow \left(x_{k+1} - \Phi x_k\right)^T e_m = 0$$

where e_m is the mth unitary coordinate vector, i.e,

$$e_m = \underbrace{\left[0 \quad \cdots \quad 1 \quad \cdots \quad 0\right]^T}_{m}.$$

Note that equations (6) and (7) express the fact that the approximation error between x_{k+1} and the linear model Φx_k must be orthogonal to all vectors Φ_l (for l in the set of indices of unknown components of x_k) and to all coordinate vectors e_m (for m in the set of unknown components of x_{k+1}). Rearranging equations (6)-(7) we finally get the following set of linear equations to be solved

$$\begin{cases} \Phi_l^T x_{k+1} - \Phi_l^T \Phi x_k = 0, & l \in I_l \\ e_m^T x_{k+1} - e_m^T \Phi x_k = 0, & m \in I_m \end{cases} \tag{8}$$

Note that (8) is a system of linear equations, which has the advantage that it can be solved numerically very efficiently. If we use the subscript *knownk* for all column indices corresponding to known values of x_k, *knownk1* for all column indices corresponding to known values of x_{k+1}, *unknownk* for all column indices corresponding to unknown values of x_k and *unknownk1* for all column indices corresponding to unknown values of x_{k+1}, then the system of equations (8) can be rewritten as

$$Ax = b \tag{9}$$

where

$$A = \begin{bmatrix} \Phi_{unknownk}^T \\ e_{unknownk1}^T \end{bmatrix} \begin{bmatrix} I_{unknownk1} & -\Phi_{unknownk} \end{bmatrix},$$

$$b = -\begin{bmatrix} \Phi_{unknownk}^T \\ e_{unknownk1}^T \end{bmatrix} \begin{bmatrix} I_{knownk1} & -\Phi_{knownk} \end{bmatrix} \begin{bmatrix} x_{k+1} \\ x_k \end{bmatrix}_{known}$$

$$x = \begin{bmatrix} x_{k+1} \\ x_k \end{bmatrix}_{unknown}$$

4. Numerical Example

Consider

$$\Phi = \begin{bmatrix} 1.75 & 0.8 \\ -0.95 & 0 \end{bmatrix}, \quad k = 1, \quad l = 1, \quad m = 2,$$

$$x_1 = \begin{bmatrix} a \\ 1 \end{bmatrix}, \quad x_2 = \begin{bmatrix} 1 \\ b \end{bmatrix}$$

The system of linear equations (8) to be solved is

$$\begin{cases} 3.965a + 0.95b = 0.35 \\ 0.95a + b = 0 \end{cases}$$

and the solution is $a = 0.1143, \quad b = -0.1086$.

REFERENCES

[1] R. D. Kent, S. G. Adams, and G. S. Turner, "Principles of Experimental Phonetics", chapter in *Models of Speech Production*, pp. 2–45, Mosby, 1996, Editied by N.J. Lass.

[2] Haskins Laboratory, "Introduction to Articulatory Phonology and the Gestural Computational Model", http://www.haskins.yale.edu/research/gestural.html

[3] L. J. Lee, P. Fieguth, L. Deng, "A Functional Articulatory Dynamic Model for Speech Production", *Proc. IEEE International Conference on Acoustics Speech and Signal Processing*, Salt Lake City, Utah, May 2001.

[4] S. King and A. Wrench, "Dynamical system modelling of articulator movement", *Proc. International Congress of Phonetic Sciences*, San Francisco, California, pp.2259—2262, 1999.

[5] C. P. Browman and L. Goldstein, "Towards an articulatory phonology", *Phonology Yearbook*, 3:219—252, 1986.

ACKNOWLEDGEMENTS

The first author would like to acknowledge NSERC and NATEQ for partially funding this research.