Speech Synthesis by Articulatory Models

Helmuth Ploner-Bernard

***Abstract*— This paper is supposed to deliver insights into the various aspects associated with the field of articulatory speech synthesis. After a short overview of human speech production mechanisms and wave propagation in the vocal tract, the acoustic tube model is derived. Several kinds of articulatory models are presented. The "inverse problem" of model parameter estimation is addressed in some detail.**

I. Introduction

RESEARCHERS have been interested in articulatory synthesis for more than two decades. This technique is conjectured to lead up to the most natural sounding synthetic speech. Moreover, articulatory models can be employed in low bit-rate coding ([12]) and, to some extent, in speech recognition.

In order to develop such models, profound knowledge in acoustics, mechanics, physiology, linguistics, phonetics, computer vision and signal processing in general is needed.

Articulatory models attempt to describe the actual speech production mechanisms by a set of slowly time-varying physiological parameters, as lung pressure, glottal widths, shape of the tongue, lip opening and protrusion as well as the amount of coupling of the nasal cavity.

From such a model, area functions describing the geometry of the vocal tract can be inferred by "informed guesses" with the use of interpolation schemes. This function might be sampled and incorporated in a so-called Wave Digital Filter. Together with an appropriate excitation, an output in the form of a speech signal can be generated, as depicted in figure 1.

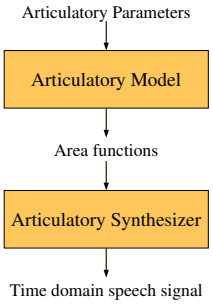

Fig. 1
Speech Synthesis by Articulatory Models

Unlike with simple formant synthesizers, source-tract interactions can be accounted for quite easily ([13]).

In the literature, stress seems to be laid primarily on the synthesis of vowels. For fricatives, noise is generated at the glottis and at the point of the narrowest constriction ([13]).

Graz University of Technology

II. Human Speech Production

In the source-filter model, human speech production is approximated as a source signal being filtered by the vocal tract ([15]). The various positions and movements of the speech-organs, also called *articulators*, are responsible for the acoustic differences between sounds. The articulators comprise the lips, the teeth, the tongue, the jaw and the velum, as shown in figure 2.

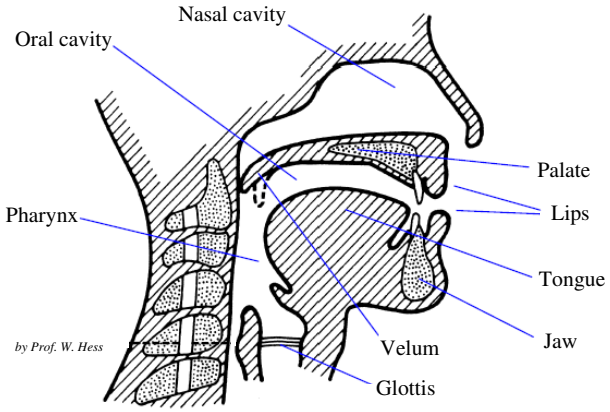

Fig. 2
The human articulators, from [1]

For the articulation of each phoneme, there are *critical* and *non-critical* articulators. The former are essential for a correct production of a phoneme, while the latter are virtually free. Based on this phenomenon, priorities can be assigned to the parameters of articulatory models, determining how critical they are for a given phoneme ([5], [12]).

In fluent speech, the target positions of the articulators are strongly affected by the context in which they appear. This effect is known as *co-articulation*. It can be handled quite naturally by an articulatory model, if it incorporates realistic physiological and dynamic constraints ([14]), as in the functional models outlined in section V.

III. Wave Propagation in the Vocal Tract

The *acoustic theory of speech production* by Fant and Ungeheuer models the vocal tract as an acoustic tube, whose walls are viewed to have an infinitely high sound impedance (cf. [15, chapter 2]). The sound field of such a system is governed by Webster's Horn equation for lossless planar wave propagation:

$$\frac{\partial^2 v(x,t)}{\partial x^2} + \frac{1}{A(x)}\frac{dA(x)}{dx}\frac{\partial v(x,t)}{\partial x} = \frac{1}{c^2}\frac{\partial^2 v(x,t)}{\partial t^2}. \qquad (1)$$

Here $v(x,t)$ is the sound particle velocity, c is the velocity of sound propagation and $A(x,t)$ the so-called *area function*, i. e. the cross-sectional areas as a function of the position between the glottis and the lips. Its shape depends on the specific positions of the articulators.

For the neutral vowel /ə/, the vocal tract is approximated by a cylindrical acoustic tube of constant cross-section ($A(x,t) \equiv$ const $\forall x,t$), where planar wave propagation is assumed, i. e. the variables describing the sound field only depend on one coordinate and on time. The resonance frequencies f_k of such a tube of length l are

$$f_k = \frac{(2k-1)c}{4l}, \quad k = 1, 2, \ldots . \tag{2}$$

Experiments have shown that bent pipes, which would model the vocal tract more accurately, have comparable eigenfrequencies ([13]).

Unfortunately, it is not possible to solve eq. 1 analytically for an arbitrary $A(x,t)$. It can be shown, however, that the formant frequencies in eq. 2 vary depending on the amount and on the position of the articulation.

The model of plane wave propagation in the vocal tract is valid for frequencies up to $f_p = 3.5\,\text{kHz}$, corresponding to the largest segments ([13]). Above this frequency, the first cross-modes emerge. Fortunately, most of the energy of speech signals is concentrated in the region below f_p.

The nasal cavity enters the model as a separate tube of fixed length parallel to the vocal tract ([5]).

IV. Acoustic tube model, Kelly-Lochbaum Structure and Wave Digital Filters

As a starting point, consider an acoustic tube of *constant* cross-sectional area. In this case, since $A \equiv$ const and hence $\frac{dA(x)}{dx} = 0$, eq. 1 can be simplified to yield

$$\frac{\partial^2 v(x,t)}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 v(x,t)}{\partial t^2}. \tag{3}$$

It is convenient to introduce the volume velocity $u(x,t)$, defined as

$$u := v\,A. \tag{4}$$

A general solution of eq. 3 can be seen as a combination of two volume velocity waves traveling forward and backward, respectively:

$$u(x,t) = u_\text{f}\Big(t - \frac{x}{c}\Big) - u_\text{b}\Big(t + \frac{x}{c}\Big) \tag{5}$$

As depicted in figure 3, the (continuous) area function in section III can be approximated by a concatenation of homogeneous acoustic tubes. The sudden change in cross-sectional areas at the junctions between the segment k and the segment $(k-1)$ (cf. figure 4) is equivalent to changes in the acoustic impedances, so that part of the traveling wave is reflected according to the reflection coefficient

$$r_k = \frac{A_{k-1} - A_k}{A_{k-1} + A_k}. \tag{6}$$

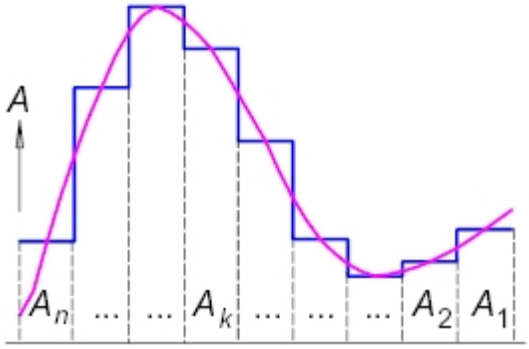

Fig. 3
Approximation of a continuous area function by distinct sections, from [1]

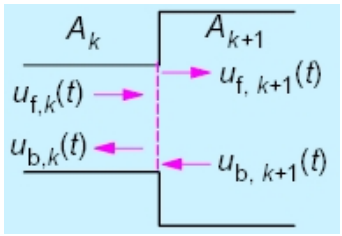

Fig. 4
Junction between segment k and segment $k-1$, from [1]

Fant's model of the vocal tract consists of just a few sections of *variable* length.

Another approach takes n equidistant samples of the area function $A(x)$ in intervals of $k\Delta x, k = 1, 2, \ldots, n$, cf. figure 5. The delay through each segment of length Δx thus is

$$\tau = \frac{\Delta x}{c}. \tag{7}$$

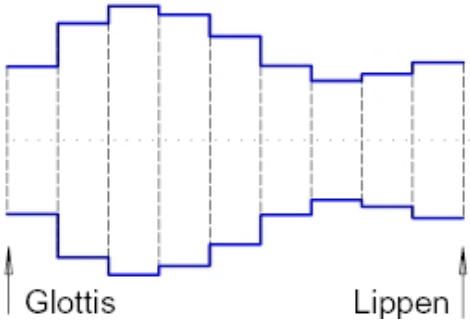

Fig. 5
Example of the tube model, from [1]

From that, the Kelly-Lochbaum implementation of the junction as in figure 6 can easily be derived. The complete structure consists of a multitude of such sections.

This idealized form is assumed to be lossless. In reality, there are of course losses in the vocal tract, mainly due to

- resonances of yielding walls (in opposition to rigid walls),

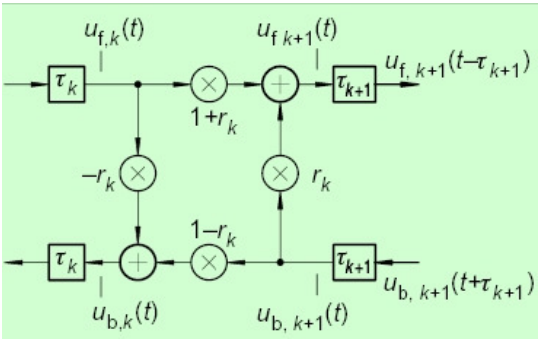

Fig. 6

JUNCTION BETWEEN TWO ACOUSTIC TUBES OF CONSTANT CROSS-SECTIONAL AREAS AS MODELED BY KELLY-LOCHBAUM, FROM [1]

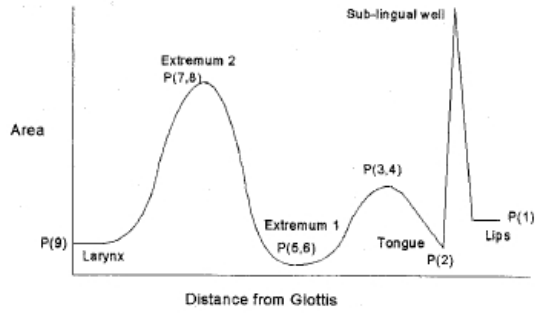

Fig. 7

A STATIC ARTICULATORY MODEL, FROM [8]

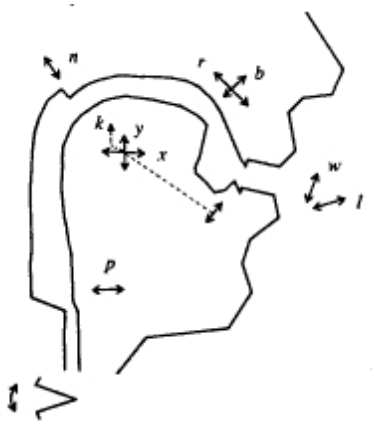

Fig. 8

COKER'S DYNAMIC ARTICULATORY MODEL, FROM [11]

- viscous and thermal losses along the path of propagation, and perhaps due to the most important
- radiation at the lips.

Constant damping factors can be introduced in the model by inserting additional multipliers immediately at the entrance and at the exit of each junction ([7], [8]). As for the radiation at the lips, an additional segment can be inserted in front of the lips.

From the Kelly-Lochbaum structure, a realization in form of a *wave digital filter* can be derived by freezing τ to any given sampling interval.

V. ARTICULATORY MODELS

There are two kinds of articulatory models: Static and Dynamic models ([13]).

- Static or descriptive models describe the vocal tract in terms of area functions. Articulator motion is interpreted as a succession of stationary shapes. As an example, consider a parametrization of an area function by nine parameters defining the areas at certain positions in the vocal tract, like the area at the lips and at the glottis, as well as the coordinates of the highest points of the tongue and the area at these points (cf. figure 7 from [6], [8]).
- Dynamic or functional models set up the equations of motion for each articulator. In such models, the articulators can be considered to be elastic, and from their inertia there might arise constraints regarding their positions, velocities and accelerations. A dynamic model is shown in figure 8.

In general, the parameter space of descriptive models has a higher dimensionality than that of functional models. Note that for vowels, a three-dimensional parameter space spanned by their three first formant frequencies might already suffice.

Since articulatory parameters vary much more slowly than acoustic parameters, this domain seems to be more suitable for parameter interpolation ([12]).

More often than not, articulatory model assume a fixed vocal tract length.

VI. PARAMETER ESTIMATION

The so-called "inverse problem" of acquiring model parameters for use in articulatory synthesizers directly or indirectly from speech signals still represents a major difficulty. This acoustic-to-articulatory mapping can be shown to be non-unique, i. e. more than one vocal tract shape can produce speech signals with almost identical spectra ([13]). Thus, besides good acoustic matching, a smooth evolution of the area functions and anatomical feasibility of the corresponding articulatory parameters might be required.

Furthermore, many of the procedures mentioned below are unable to determine the length of the vocal tract configurations.

A. Linear Predictive Coding

A simple method of determining the vocal tract shape directly from the speech signal consists in evaluating the reflection coefficients arising from the recursive Levinson-Durbin-algorithm for Linear Predictive Coding analysis.

However, these parameters characterize the idealized acoustic tube model, but result from considering a real world lossy signal. As described in [15], this discrepancy is responsible for the inaccurate results of this method.

B. Magnetic Resonance Imaging

Perhaps the most intuitive way of inferring the vocal tract shape is the measurement with methods such as Magnetic Resonance Imaging (MRI) as in [3] or in [9].

Several scans of the subject have to be made in order to obtain 3D models. It should be noted that in general 3D models of the

vocal tract are crucial for a representation of lateral sounds like /l/, where the exclusive information in the mid-sagittal plane does not suffice.

The present drawbacks of being remarkably costly, time consuming and noisy during image acquisition might be circumvented by better scanning devices.

An example of the equipment used in this technique is shown in figure 9.

Fig. 9
MAGNETIC RESONANCE IMAGING EQUIPMENT, FROM [2]

C. Acoustic Impedance Measurement

As explained in [4], the geometry of the vocal tract can be calculated from the acoustic impedance of the human speech production system as seen from the lips.

In this procedure, a particular acoustic volume velocity impulse is sent toward the lips while a speaker articulates a certain sound, travels through the vocal tract and gets reflected at the closed glottis, as depicted in figure 10. From the sound pressure wave recorded at the lips it is possible to draw conclusions about the geometry of the vocal tract assuming plane wave propagation without losses.

To account for losses, the inferred area function is placed into a lossy transmission line and an impedance is computed. The results are compared to the measured data and optimized in an iterative manner.

This procedure provides a cheap, fast method for the measurement even of a large number of vocal tract shapes, although the cases of nasal sounds is not taken into consideration.

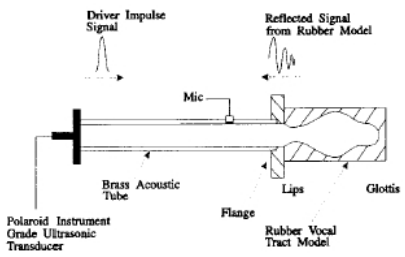

Fig. 10
ACOUSTIC IMPEDANCE MEASUREMENT, FROM [4]

D. Analysis by Synthesis

Analysis-by-synthesis procedures permit automated parameter estimation by successively changing the parameters of the articulatory synthesizer in order to approximate a given natural utterance.

The speech signal is segmented on a phoneme basis or with a fixed frame length. For each such section, a set of descriptive parameters is extracted, e. g. LPC-coefficients, mel frequency cepstral coefficients, or the coefficients of any spectral transformation. These parameters are used to search a so-called codebook for the "best match". A codebook contains a huge amount of possible parameters along with their corresponding vocal tract shapes, represented as (sampled) area functions or articulatory parameters, depending on the kind of model used (cf. section V). For details on how to populate a codebook, see section VI-D.1. The signal is then re-synthesized using the parameters from the codebook, compared to the original signal and optimized iteratively.

Due to the aforementioned non-uniqueness of the acoustic-to-articulatory mapping, more than one configuration might result in the same acoustic representation. Thus, other constraints than good spectral agreement alone are introduced by optimizing a more complicated cost function. The parameters of such a cost function may consist of: (1) acoustic distance from spectral coefficients, (2) smoothness of vocal tract shape, (3) smooth temporal evolution of the vocal tract and (4) of the vocal folds between adjacent frames, (5) energy of natural and synthetic signals ([10]).

The algorithm can be improved by optimizing multiple frames at a time, making it easier to obtain the smooth variations of vocal tract shapes essential for natural sounding speech.

However, in each case the re-synthesized signal should be time aligned to the original speech before performing any processing. In addition, any analyses should be done pitch-synchronously to minimize the influences of the glottal excitation ([12], [11]).

Dynamic programming provides a technique for drastically reducing search times while accessing the codebook ([13]).

1) Population of the codebook: A simple approach to populate the codebook for the acoustic-to-articulatory mapping consists in randomly iterating through various configurations of articulatory parameters and storing them together with their spectra or other acoustic representations. Unfortunately, such codebooks will contain many unnecessary data, in that some articulatory configurations are not used in a given language or by a particular speaker ([13]).

Alternatively, the "inching"-approach starts out at extreme articulatory parameters and performs some kind of interpolation between them on trajectories in the articulatory space. The risk here is that there might still remain sparsely populated areas ([6], [8]).

REFERENCES

[1] http://www.ikp.uni-bonn.de/dt/lehre/materialien/aap/aap_1f.pdf.
[2] http://www.radiologyinfo.org/.
[3] P. Badin, G. Bailly, M. Raybaudi, and C. Segebarth. A three-dimensional linear articulatory model based on MRI data. *SCA/COCOSDA Workshop on Speech Synthesis, Jenolan Caves, Australia*, 1998.
[4] J. W. Devaney and C. C. Goodyear. A comparison of acoustic and magnetic resonance imaging techniques in the estimation of vocal tract area functions. *International Symposium on Speech, Image Processing and Neural Networks*, pages 575–578, April 1994.

[5] Georg Dorffner, Markus Kommenda, and Gernot Kubin. GRAPHON – the Vienna speech synthesis system for arbitrary german text. *IEEE International conference on Acoustics, speech and signal processing*, pages 744–747, April 1985.

[6] C. C. Goodyear and Dongbing Wei. Articulatory copy synthesis using a nine-parameter vocal tract model. *IEEE international conference on acoustics, speech and signal processing*, pages 385–388, May 1996.

[7] Andrew R. Greenwood. Articulatory speech synthesis using diphone units. *IEEE international conference on acoustics, speech and signal processing*, pages 1635–1638, April 1997.

[8] A. R. Greenwood and C. C. Goodyear. Articulatory speech synthesis using a parametric model and a polynomial mapping technique. *International symposium on speech, image processing and neural networks*, pages 595–598, April 1994.

[9] A. R. Greenwood, C. C. Goodyear, and P. A. Martin. Measurements of vocal tract shapes using magnetic resonance imaging. *Communications, Speech and Vision, IEE Proceedings I*, pages 553–560, December 1992.

[10] S. K. Gupta and J. Schroeter. Low update rate articulatory analysis/synthesis of speech. *International conference on acoustics, speech and signal processing*, pages 481–484, April 1991.

[11] S. Parthasarathy and C. H. Coker. Phoneme-level parametrization of speech using an articulatory model. *International Conference on Acoustics, Speech and Signal Processing*, pages 337–340, April 1990.

[12] S. Parthasarathy, J. Schroeter, C. Coker, and M. M. Sondhi. Articulatory analysis and synthesis of speech. *Fourth IEEE region 10 international conference*, pages 760–764, November 1989.

[13] J. Schroeter and Man Mohan Sondhi. Techniques for estimating vocal-tract shapes from the speech signal. *IEEE transactions on speech and audio processing*, pages 133–150, January 1994.

[14] S. Terepin and F. Fallside. A polynomial vocal tract model for speech synthesis. *Acoustics, Speech and Signal Processing, IEEE International Conference on ICASSP*, pages 919–922, May 1982.

[15] Peter Vary, Ulrich Heute, and Wolfgang Hess. *Digitale Sprachsignalverarbeitung*. B. G. Teubner Stuttgart, 1998.

[16] Zhen-Li Yu and Shang-Cui Zeng. Acoustic-to-articulatory mapping codebook constraint for determining vocal-tract length for inverse speech problem and articulatory synthesis. *5th international conference on signal processing proceedings*, pages 827–830, 2000.