AN ARTICULATORY SPEECH SYNTHESIZER TALKING GERMAN

Gernot KUBIN, Vytautas PIKTURNA +)

Institut für Nachrichten- und Hochfrequenztechnik, TU Wien
Gusshausstrasse 25/389, A-1040 Vienna, Austria

ABSTRACT

Modern digital signal processing technology opens the way to real-time implementation of articulatory speech synthesizers as the phonetic-acoustic conversion module in text-to-speech systems. An outline of a workstation for the development of such a prototype synthesizer for the German language is given. This workstation is equipped with fast interactive graphics and acoustics processing capabilities and is used as a tool for both the study of articulatory phenomena as such and the development of simplified algorithms needed for the prospective target realization of the articulatory synthesizer.

1. INTRODUCTION

Until now most of the development in articulatory speech synthesis [1-6] has originated within a phonetic research environment. Only recently [7] attempts have been made to tailor this methodology into a form susceptible for real-time implementation with modern digital signal processing technology.

As articulatory synthesis is expected to yield high synthetic speech quality we have chosen this line for the development of the phonetic-acoustic conversion component within a text-to-speech system for German [8]. Our goal is not to refine the knowledge of human articulation in numerous experiments but rather to use the available knowledge for an operational speech prooduction model.

Therefore we confined the development environment to be small from the beginning: a specifically designed workstation that provides close-to-real-time operation of both the computer animated articulatory model and the acoustic signal synthesizer. The workstation facilitates changes in the detailed model and synthesizer structures while preserving the hard- and software characteristics of the envisaged target system.

+) On leave from Kaunas Polytechnical Institute, Jurostr. 65-302, 233028 Kaunas, USSR

2. SYSTEM OVERVIEW

In the text-to-speech system GRAPHON the articulatory synthesizer bridges the gap between the string of phonetic symbols derived from morphological word parsing [9] of the input text on one side and the synthesized acoustic speech signal on the other side. To this end, the articulatory synthesizer must provide the following four steps:

(1) Interpretation of phonetic symbols in the articulatory domain by means of look-up tables containing geometry and timing parameters. Only ***essential*** or non redundant parameters are used for the definition of a phone, leaving the final determination of the time-varying vocal-tract contours to the following step.

(2) Synthesis of articulatory kinematics by interpolation in the articulatory domain. Thereby ***non-essential*** or redundant parameters (e.g. lip rounding in the articulation of a German [t]) are generated. Secondly, intermediate positions of articulary movements can be generated at an arbitrary rate.

(3) Graphical display and evaluation of sequences of mid-sagittal views. Speech organ contours are generated mathematically from complete sets of geometric parameters defined for a certain time step. Vocal-tract area functions are estimated from linear distances measured between speech organ contours.

(4) Acoustic synthesis with a wave digital filter implementation of a vocal tract model controlled by the time-varying area functions, cf. [7].

As the basic principle of operation has already been discussed in [8,10] only a few points of special interest will be discussed in the sequel.

3. ARTICULATORY PHONETICS AND COMPUTER ANIMATION

It appears obvious that a full account of human articulation is impossible: neither the neuromuscular control of the speech organs nor the dynamics of their movements is fully understood. Even a phenomenological description of their kinematics seems quite untractable as the motion of three-dimensional non-rigid bodies is involved. Especially the continuous change of the tongue shape and position is hard to measure and to model adequately. What is left are a few basic facts describing certain stable articulatory mechanisms either in the steady state [11] or in transitions [12]. The rest is hypothesis.

How can this incomplete knowledge be exploited for speech synthesis? The answer is that even a rudimentary articulatory model introduces an additional level for the representation of spech phenomena such as coarticulation, reduction, assimilation, homorganic articulation, and other contextdependent allophonic variation. This additional level appears more suited to human intuition in the manipulation of hypotheses than the lower levels such as an exclusively acoustic signal description. Furthermore, it opens certain degrees of freedom hidden to the human experimenter at the acoustic level: some simple articulatory movements may induce very complex acoustic mechanisms that would not be recognized as basic to the speech production process at the acoustic level as they appear buried in the mess of signal variability.

Summarizing, the actual structure of an articulatory model is only partly determined by human articulation itself whereas an even larger part is due to the ***means of representation*** used in the desired application. As our application requires interaction with a human experimenter, a graphical display of speech organ movements is indispensible. Thus the principles of ***computer animation*** govern largely the design of our articulatory model.

(1) Animation of axionometric displays of three-dimensional shapes would be too clumsy on the envisaged "small" hardware environment.

(2) Two-dimensional shapes can be adequately displayed by their ***contours.*** Representing speech organs by their contours only, deliberately dismisses all knowledge concerning their morphology and internal dynamics. The prevailing information about articulator contours is conveyed by mid-sagittal (cine-)radiographies [11,13,14]. These are taken as the starting point for modeling the vocal-tract geometry.

(3) The methods for the synthesis of articulatory movements may be classified according to [15] as follows:

- ***Image-based key-frame animation*** generates intermediate frames from fully specified ***key-frames*** by interpolation of shapes without taking into account any structural information about this shape. This principle is similar to the diphone synthesis concept in exclusively acoustic speech synthesizers. As the velar movement shows a single degree of freedom it can be adequately modeled by this technique.
- ***Parametric key-frame animation*** has previously been used in articulatory synthesis [3] for a ***synthesis-by-script*** mode of operation. Still the human experimenter provides fully specified key-frames but these are interpreted in a parameter domain so that interpolated frames preserve certain structural characterictics of the parameterized shape. This principle is similar to the (allophone) synthesis by rule concept in exclusively acoustic speech synthesizers.
- ***Kinematic algorithmic animation*** is our approach for the modelization of highly mobile variable-shape articulators, in particular the tongue, lips and epiglottis. There exists no similar concept in exclusively acoustic speech synthesizers. The synthesized frame sequence is no more specified from key-frames but from algorithmic parameter control laws. Because there is a direct *open-loop relationship* between the control laws and the controlled geometry and timing parameters this technique is a *kinematic* one. In our system, typical laws specify the durations of on-glide, stationary, and off-glide phases in the movement of a particular articulator within a given phone [16]. These durations may assume negative values e.g. to emphasize anticipatory coarticulation or reduction.
- ***Dynamic algorithmic animation*** requires the replacement of the above kinematic laws by models of the internal speech organ *dynamics.* This approach [6, p. 279] goes beyond our previous option for simple contour line geometry. It introduces an additional level of representation, i.e. complexity, which we consider only worthwhile to be studied in the context of text-to-speech synthesis after completing the study of pure kinematic models.

(4) Sampling of articulatory movements is sufficient at a rate of approx. 20 frames/sec for the human eye. However, this rate does not fully captûre the true motion of speech organs. For this purpose, a rate of at least 50 frames/sec should be used. Yet, it is important to separate the two rate requirements when implementing the computational model: every second, 50 frames must be calculated and evaluated by the graphics processing system while only 20 midsagittal contour plots must be output via the video display.

4. ACOUSTIC PHONETICS AND SIGNAL PROCESSING

Acoustic phonetics is seemingly more tractable than articulatory phonetics as there exist highly refined models of the vocal-tract acoustics such as [17]. More often than not, these models are delineated as an electrical circuit analog which can in turn be transformed into a digital circuit. The most elegant strategy consists in the ***wave digital filter*** (WDF) concept [18] which provides a direct translation of the analog voltage and current relations into the digital domain. A reasonably simplified WDF version of [17] has been implemented for the development of a quasi-articulatory speech synthesizer in [7].

We adopt this procedure while modifying its implementation according to our hardware system that comprises a vector-oriented bit-slice signal processor for the acoustic signal synthesis . This processor is controlled by a MC68000 microprocessor system developed at our department with special attention to the fast high-resolution graphics as needed for the animated articulatory model. The two processor systems are coupled via a parallel interface with a transfer rate of up to 3 Mbyte/sec. This interface transmits the area function values estimated from linear distances between speech organ contours on the basis of piece-wise approximation formulae given in [4]. The vocal-tract synthesis filter is tuned according to the area function in a time-varying manner. The operation of the signal synthesis can be supervised with a waveform editor and linear predictive analysis module integrated in the workstation utilities.

5. PERCEPTUAL PHONETICS AND SYSTEM EVALUATION

There are a lot of open issues that can only be studied in perceptual experiments implementing a feedback loop for system opimization through a *human experimenter*:

(1) How accurate must an acoustic vocal-tract model be, given its control by a fairly coarse articulatory model?

(2) What is the adequate level of representation for various speech phenomena? Adequacy should be defined by the human listener's judgement while the choice among several adequate representations should be made such that implementation complexity is minimized. For instance, it is not at all clear which articulatory transitions really *need* to be represented in the articulatory domain and which could be established by simplified rules operating directly on area functions or acoustic parameters.

(3) Feedback control should be made possible at all system levels. This calls for comparison mechanisms for mid-sagittal views and area functions as well as for spectrographic measurements. To fulfill this requirement, an interactive phonetic editor is built with thumb-wheel control of articulatory geometry and real-time output of the speech organ contours, the area function, and the synthetic speech signal.

(4) Special attention is devoted to rapidly time-varying speech events such as the explosion in stop consonants. For their detailed study both adaptive methods as well as new time-frequency analysis methods [19] are under investigation.

6. CONCLUSION

Several concepts fundamental to the design of a workstation for the development of a real-time articulatory speech synthesizer have been discussed. At the present state of the system, articulatory kinematics can be computed and displayed by our graphics system at a rate of 10 frames/sec approx. Speech signals can be produced with a sampling rate of 10 kHz. For a target system with 50 frames/sec and 20 kHz sampling an increase in computational capacity by a factor of 5 is needed. This is well within reach of off-the-shelf components (e.g. MC68020 with floating-point co-processor and 4 DSP chips such as TMS 32010). These data show an impressive technology step when they are compared to run-time data of articulatory models published a few years ago, e.g. 360 times real time in [2] or 20 to 60 times real time in [3]. Taking up this step is essential for applied articulatory synthesis.

7. REFERENCES

[1] J.L. Kelly, C.C. Lochbaum: Speech Synthesis. 4th Int. Congr. on Acoustics, Copenhagen, August 1962, paper G42.

[2] C.H. Coker: A Model of Articulatory Dynamics and Control. Proc. IEEE 64 (1976), pp. 452-460.

[3] Ph. Rubin, Th. Baer, P. Mermelstein: An articulatory synthesizer for perceptual research. JASA 70 (1981), pp. 321-328.

[4] G. Heike et al.: Berichte des Instituts für Phonetik der Universität zu Köln, IPKöln-Berichte 10(1980), 12(1982), 13(1986).

[5] T. Thomas, F. Fallside: A new articulatory model for speech production. Proc. IEEE Int. Conf. on Acoustics, Speech, and Sig. Proc. ICASSP'85, Tampa (Fla.) March 1985, pp. 1105-1108.

[6] R. Carré, R. Descout, M. Wajskop [eds.]: Articulatory Modeling and Phonetics. G.A.L.F. Groupe de la Communication Parlée - Proc. of the Symposium at Grenoble, July 1977.

[7] P. Meyer, R. Wilhelms, H.W. Strube: An efficient vocal tract model running in real time. In: I.T. Young et al. [eds.]: Signal Processing III: Theories and Applications (Proceedings of EUSIPCO'86), North-Holland 1986, pp. 377-380.

[8] G. Dorffner, M. Kommenda, G. Kubin: GRAPHON-The Vienna Speech Synthesis System for Arbitrary German Text. Proc. IEEE Int. Conf. on Acoustics, Speech, and Sig. Proc. ICASSP'85, Tampa (Fla.) March 1985, pp. 744 - 747.

[9] A. Pounder, M. Kommenda: Morphological Analysis for a German Text-to-Speech System. Proc. 11th Int. Conf. on Computational Linguistics COLING'86, Bonn(FRG) August 1986, pp.263-268.

[10] G. Dorffner, G. Kubin: Artikulatorische Sprachsynthese. Mikroelektronik in Österrreich. Wien: Springer 1985, pp. 456-461.

[11] G. Fant: Acoustic Theory of Speech Production. s'Gravenhage: Mouton 1970(2).

[12] O. Fujimura: Temporal Organization of Articulatory Movements as a Multidimensional Phrasal Structure. Phonetica 38 (1981), pp. 66-83.

[13] Wängler: Atlas deutscher Sprachlaute. Berlin: Akademie-Verlag 1981(7).

[14] G. Lindner: Optische Analysen der Koartikulation durch Röntgenkinematographie. Hochschulfilm T-HF 719. Berlin: Institut für Film, Bild und Ton, 1972.

[15] L. Forest, N. Magnenat-Thalmann, D. Thalmann: Integrating Key-Frame Animation and Algorithmic Animation of Articulated Bodies. In: T.L. Kunii [ed.]: Advanced Computer Graphics. Tokyo: Springer-Verlag 1986, pp. 263 - 274.

[16] G. Dorffner, M. Kommenda: Ein Artikualtionsmodell zur Sprachsynthese. Fortschritte der Akustik - DAGA'85. Bad Honnef: DPG-GmbH 1985, pp. 615-618.

[17] J.L. Flanagan, K. Ishizaka, K.L. Shipley: Synthesis of Speech From a Dynamic Model of the Vocal Cords and Vocal Tract. BSTJ 54 (1975), pp. 485-506.

[18] A. Fettweis: Wave Digital Filters: Theory and Practice. Proc. IEEE 74 (1986), pp. 270-327.

[19] W. Wokurek, G. Kubin, F. Hlawatsch: Wigner Distribution - A New Method for High-Resolution Time-Frequency Analysis of Speech Signals. Proc. 11th Int. Congr. of Phon. Sciences, Tallinn August 1987, this volume.

APPENDIX

As a reference to our articulatory model two figures are presented:

Fig. 1 shows the parameterization of the articulatory geometry by approximation of the speech organ contours with simple mathematical functions (circle, tangent).

Fig. 2 shows a (subsampled) synthetic frame sequence for the German word [matəma:tik].

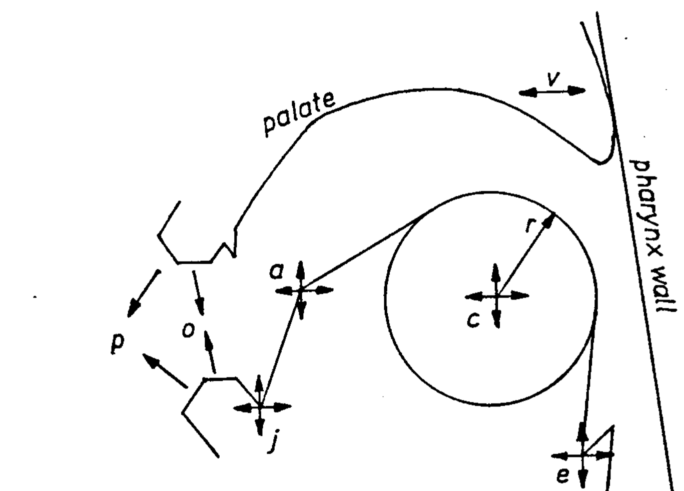

Fig. 1. Articulatory parameters:
- a tongue apex
- c tongue body centre
- e epiglottis
- j jaw (lower incisors)
- o lip opening
- p lip protrusion
- r tongue body radius
- v velum

palate and *pharynx wall* are fixed reference positions

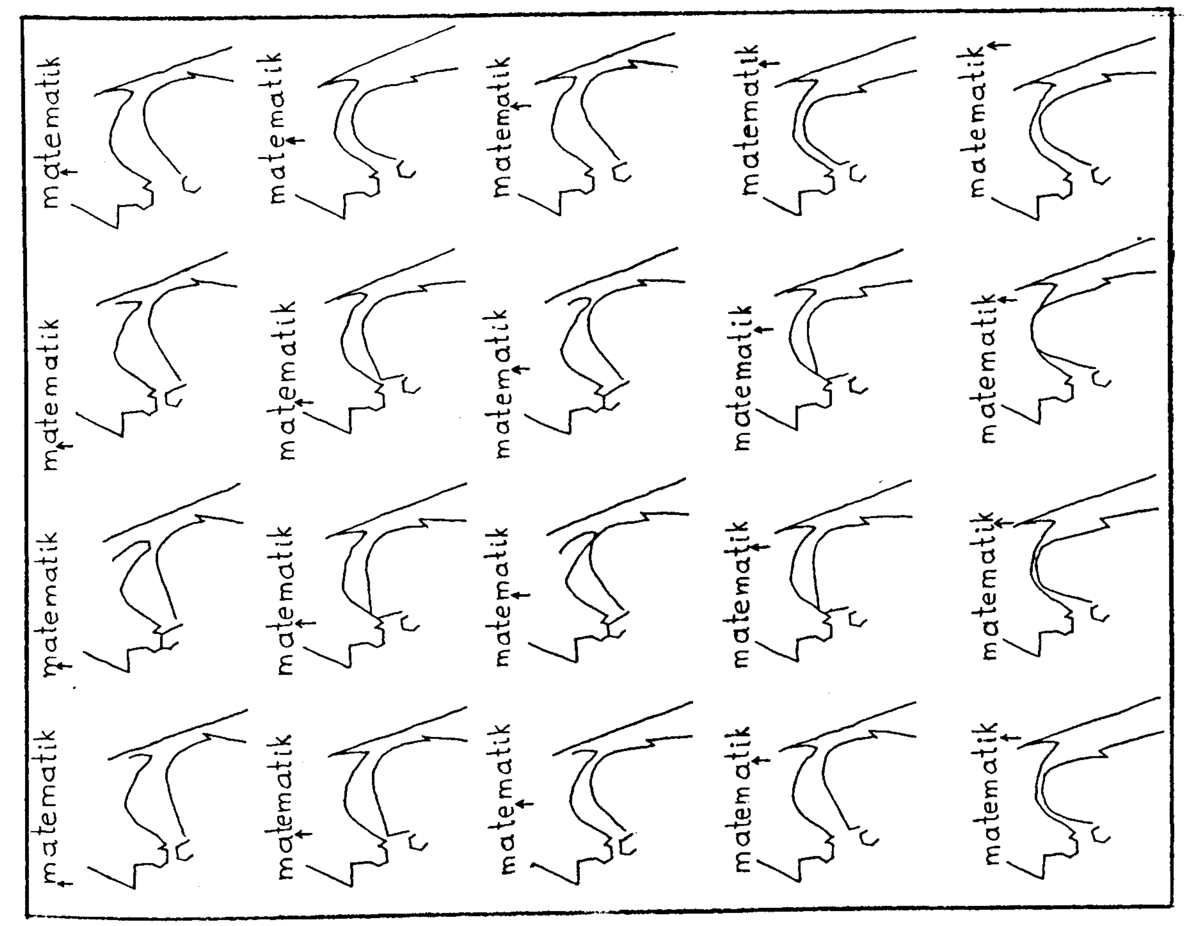

Fig. 2. (Subsampled) synthetic frame sequence for [matəma:tik], the arrow points at the current time position of the frame within the whole word.