AN ARTIFICIAL SPEECH SYNTHESIZER TALKING GERMAN

Gernot KURIN, Vytautas PIKTURNA

Institut für Nachrichten- und Hochfrequenztechnik, TU Wien

Jurostr. 65-302, A-1040 Vienna, Austria

ABSTRACT

Modern digital signal processing technology opens the way to real-time implementation of artificial 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 interact graphics and acoustic signal 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-acoustic research context only. Recently [7] attempts have been made to tailor this methodology into a form susceptible for real-time application.

As articulatory synthesis is expected to yield high-fidelity speech quality, we provide details to the hardware environment. Since we concentrated the design effort on the workstation for German [8], our goal is not to refine the knowledge of human articulation in numerous experiments but rather to use the available know—

ledge for an operational speech production model. Additionally, we aim to provide a small hardware environment.

2. SYSTEM OVERVIEW

In the test-to-speech system GRAPFR the articulatory synthesizer bridges the gap between the phonetic-acoustic domain of acoustic speech signal on the one side and the synthetic 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 geometric parameters and timing parameters. These 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-essentials such as redundant parameters (e.g. lip rounding in the articulation of a German /u/) are eliminated. Second, intermediate positions of articulatory 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 the contour shape measured between speech organ contours.

4. Acoustic synthesis with a wave digital filter implementation of a vocal tract model controlled according to speech organ contours.

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. ARTIFICIALTENOMICS AND COMPUTER ANIMATION

It appears obvious that full account of human articulation should be provided by a neuromuscular model of the speech organs nor the dynamics of their movements is fully understood. Even a phenomenological description of their kinematics would quite untractable as the motion of three-dimensional non-rigid bodies. Except for the continuous change of the tongue shape and position is hard to measure and model adequately. What is left are a few basic facts describing certain stable articulatory movements in a steady state [11] or in transitions [12]. The rest is hypothesis.

How can this incomplete information be exploited for speech synthesis? The synthesis of a rudimentary articulatory model introduces an additional level for speech representation of phenomena such as articulation, reduction, assimilation, and other context-dependent 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 invoke very complex mechanical mechanisms that would not be recognized as basic to the speech production process. This artefact is only apparent in the case of closed vowels.

Summarizing, the actual structure of an articulatory model is partly determined by human articulation itself whereas an even larger part is due to the means of representation used in the designed application. As our application requires interaction with a human experimenter, a graphical display of speech organ movements is indispensable. 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 is a rather simple, effective technique. Due to the means of representation used in the design. Kinematic algorithmic animation is our approach [6, p. 279] goes beyond our previous approach [11,12,14]. Instead, it is important to separate the two rate requirements when implementing the computational model: one rate for less than 50 frames/sec in order to imitate the changing rate of articulation, another rate of at least 50 frames/sec for the human experimenter. It is conceivable to separate these two rate requirements when using an interactive approach. In the case of computer animation, details such as stressing of articulatory movements may be studied in the context of text-to-speech synthesis after completing the study of pure kinematic models.

3. Sampling of articulatory movements is sufficient at a rate of approx. 20 frames/sec for the human experimenter. This rate does not fully capture the true motion of speech organs. For this purpose, a rate of at least 50 frames/sec should be used. It is important to separate the two rate requirements when implementing the computational model: one rate for less than 50 frames/sec in order to imitate the changing rate of articulation, another rate of at least 50 frames/sec for the human experimenter. It is conceivable to separate these two rate requirements when using an interactive approach. In the case of computer animation, details such as stressing of articulatory movements may be studied in the context of text-to-speech synthesis after completing the study of pure kinematic models.

4. Acoustic synthesis—by-script mode of operation. Still the human experimenter provides fully specified key-frames while the graphics processing system while only 20 mid-sagittal 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 [16], the human experimenter has to deal with a more subtle set of parameters than the human experimenter. How to transform a set of articulatory laws into an acoustic signal synthesis 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 optimization 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.

(3) Feedback control should be made possible at all system levels. This calls for comparison mechanisms on the 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 (circle, tangent).

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 art, articulatory kinematics can be computed and displayed by our graphics system at a rate of 10 frames/sec or more. Speech signals can be produced with a rate of 10 kHz. For a target system with 50 frames/sec and 50 kHz sampling an increase in computational capacity by a factor of 5 is needed. This is well within reach of off-the-shelf computers (e.g. MCB-802 with floating-point co-processor and 4 DSP chips such as TMS 32020). 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


