



SPEECH SYNTHESIS BY ACOUSTIC CONTROL

Georg HEIKE, Reinhold GREISBACH, Stefan HILGER, Bernd KRÖGER

Institut für Phonetik
Universität zu Köln
Greinstraße 2, D-5000 Köln 41

ABSTRACT

For performing text-to-speech conversion or synthesis by rule with an articulation based synthesis model data on articulatory dynamics have to be collected. Speech synthesis by acoustic control can be used as an aid for acquiring these articulatory data.

For speech synthesis with the aid of a computer various methods have been suggested (1). In trying to increase naturalness a synthesis method which simulates the real physical situation should be favoured, i.e. articulation based synthesis. Articulation based synthesis consists of an articulatory model simulating the movements of the articulatory organs (including glottal movements) and an acoustic model which simulates wave generation and propagation through the vocal tract. Effective models (computationally not too expensive and physically adequate) consist of a two dimensional sagittal model of the articulatory organs, a model of the movements of the vocal cords, and a one dimensional model of wave propagation, coupled by a module which transforms sagittal distance to cross-sectional area.

The main problem in simulating the generation of natural speech by an articulatory model is the acquisition of rules to control the articulatory model, especially in the case of 'synthesis by rule', i.e. speech synthesis that is controlled only by a string of symbols (as input) and a built-in system of rules which transform this discrete string into a continuous speech signal. To formulate these rules data, either of the articulatory or of the acoustic type, have to be collected. Of course articulatory data are most effective in controlling an articulatory model, but the acquisition of data on articulatory dynamics is very expensive (e.g. X-ray films) and available to everybody. Furthermore the techniques used in articulatory data acquisition can affect normal articulation habits and will probably not yield optimal synthetic speech because of (probably) ineffective articulatory to acoustic transformation. Acoustic data, on the other hand, are very easy to obtain, but no direct transform into articulatory data exists. There are two ways of solving the problem: 1. By inspection of the acoustic data and use of his acoustic-phonetic

knowledge an expert may hypothesize a rule which he then may test by using the synthesis model to generate the speech signal. Auditory feedback will help him to accept or revise the rule. 2. Acoustic data are transformed into articulatory data by a resynthesis procedure using the synthesis model. The articulatory data thus generated are used to find the appropriate rule.

A resynthesis system has been created (on an Atari Microcomputer) which can be used to derive the set of rules for the speech synthesis by rule of German. This system consists of an acoustic analysis module, a transformational procedure, an articulatory and an acoustic synthesis module (Fig. 1).

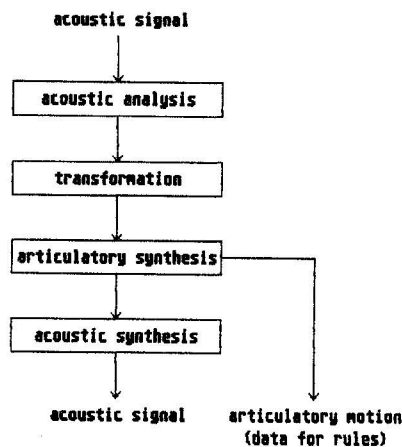


Fig. 1: Structure of the resynthesis system

After sampling various acoustic-phonetic parameters such as pitch, intensity, formant frequency are extracted from the acoustic signal. The procedures used in this acoustic analysis module depend mainly on LPC-analysis methods. For a more detailed description of the algorithm see (2). The module also includes procedures for segmentation and classification of the segments by (phonetic) symbols. Afterwards the analysed parameters as well as the segment boundaries and the symbols can be inspected and corrected manually.

The synthesis model consists of an articulatory model and of an acoustic model. The articulatory model is controlled by 9

10.21437/Eurospeech.1989-173

articulatory parameters which describe geometrically the position of the articulatory organs in a two-dimensional sagittal view (Fig. 2).

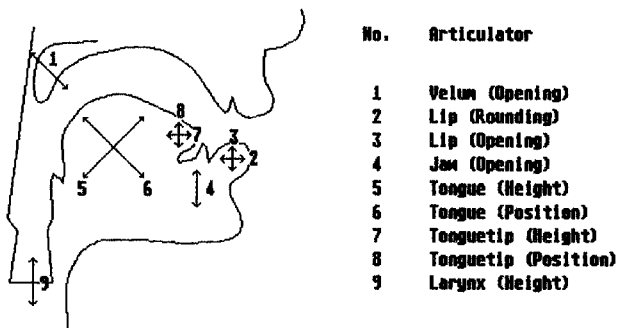


Fig. 2: Articulatory model with parameter numbers

For an exact description of the parameters and the principles of constructing an articulatory picture see (2), (3). Two further articulatory or physiological parameters are needed to describe the motion of the self-oscillating one-mass model of the vocal cords: subglottal pressure (P_s) and tension of the vocal cords (Q).

The one-dimensional acoustic model simulates the wave propagation through the vocal tract delimited by the surfaces of the articulatory organs. For this purpose the articulatory data taken from the two-dimensional sagittal view are transformed into a cross-sectional area function, or, to be more precise, into a series of cylindrical tubes of equal length but varying diameters. These tubes are used to simulate the propagation of the volume velocity (in a Kelly-Lochbaum type model) generated by the oscillating vocal cords. A nasal tract consisting also of such cylindrical tubes

can be coupled. Secondary sources of sound (friction) are at present modelled by inserting random noise into the tube of minimal diameter irrespective of the existing volume velocity. For a detailed description of the current acoustic model see (5).

The transformation module provides acoustic control of the speech synthesis. This acoustic control is performed (to accelerate computation) by simulating the signal generation by reference to tables.

At the moment two tables are used. The first one relates the 9 articulatory parameters of the model to the acoustic data, i.e. the frequency of the first 3 formants. The table is a 9-dimensional grid (representing the articulatory parameters) with 3 values (representing the formant frequencies previously calculated by the synthesis module) in every grid position. The distance of the grid was chosen with respect to availability of computer memory and acceptability of interpolation error. For every given articulatory vector the first three formants can be calculated by interpolating the grid. The second table relates the 2 (physiological) glottal parameters (P_s , Q) to the 2 (acoustical) glottal parameters (F_0 , intensity) for a given articulatory position.

As the vocal tract affects the oscillation of the vocal cords several tables are necessary which again may be ordered in a grid according to the articulatory parameters.

As the whole system should serve as a data acquisition scheme for synthesis by rule it is necessary to assume that the symbols of the (phonetic) input string are represented by specific positions of the articulatory organs (well known from X-ray pictures). Thus it is possible to predict the synthetic speech parameters for all

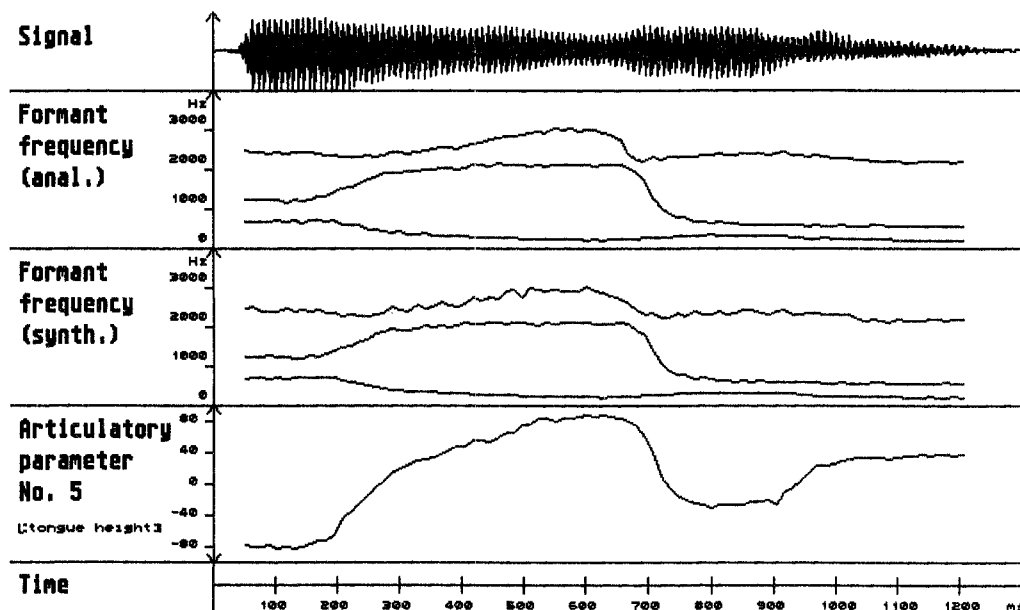


Fig. 3: Resynthesis of the sequence /aeiou/. From top to bottom: a) original speech signal; b) analysed formant frequencies; c) formant frequencies of optimal resynthesized speech signal; d) contour of one of the articulatory parameters corresponding to resynthesized speech signal.

these positions by reference to the first table and interpolation. Two procedures have to be applied for comparison of the two parameter sets:

- Adaption of the synthetic resonator to the human (analytic) resonator. This is done by normalisation of the corresponding acoustic vowel triangles.
- Adaption of the synthetic acoustic parameters to the analysed ones. As it is almost certain that the synthetic parameters will not correspond exactly, an optimization procedure is applied. This procedure tries to minimize the difference between synthesized and analysed values by varying the articulatory parameters and evaluating the resulting acoustic parameters. The optimization process is restricted by the fact that the optimal solution or one of them must lie in the vicinity of the first predicted one.

The optimized positions set up a kind of grid for the articulatory movement defined by the input symbol string as a whole. This grid is filled up by moving the articulatory organs continuously, i.e. interpolating the parameters from one grid position to the following. This interpolation procedure is done by using known facts about the dynamics of articulatory movements (e.g. (6)). It can be assumed that the actual movement for the utterance to be resynthesized lies in the vicinity of the reconstructed one. So the same optimization procedure as the one described above is applied restricted by the fact that no jumps in the trajectory of the articulatory parameters are allowed. The reconstruction of the vocal tract dynamics is followed by the reconstruction of the vocal cord movement. As in a self-oscillating glottis the control parameters are dependent on the coupled resonator it is necessary to know the actual form of the vocal tract. (It is assumed that the glottal effects on vocal tract resonance are negligible.) Again reference to the second table and interpolation are necessary to predict the correct control parameters of the vocal cords. Fig. 3 shows an example.

The result of this reconstruction process is the optimal articulatory parameter trajectory to control the synthesis process. These parameters can now be used for the acquisition of rules or be fed into the synthesis system resulting a) in a sequence of mid-sagittal pictures of the articulatory organs which can be displayed as a trickfilm on the computer monitor, and b) in an acoustic signal which can be made audible via D/A-converter and loudspeaker.

(1) D H Klatt, "Review of text-to-speech conversion for English", JASA 82, pp 737-793: 1987

(2) R Greisbach, Grundlagen der Automatisierbarkeit phonetischer Transkription. Hamburg 1988

(3) G Heike, Artikulatorische Synthese. IPKöln-Berichte 10, pp 8-31: 1980

(4) G Heike / J Philipp, "LISA - Ein Verfahren zur artikulatorischen Sprachsynthese", in: B Müller, Sprachsynthese, pp 39-53: Hildesheim 1985

(5) S Hilger, Entwicklung eines Systems zur interaktiven Analyse und Synthese von Sprache unter besonderer Berücksichtigung der Glottissimulation, Diss. Köln 1988

(6) J S Perkell, Physiology of Speech Production, Cambridge 1969