Quasi-Articulatory Real-Time Speech Synthesis

Peter Meyer, Reiner Wilhelms and Hans Werner Strube

Drittes Physikalisches Institut, Universität Göttingen,
Bürgerstrasse 42-44, D-3400 Göttingen, Fed. Rep. of Germany

ABSTRACT

To produce natural sounding transitions with a speech synthesizer by simple interpolation of its control parameters, these parameters should have articulatory meanings. In this case the synthesizer must have the form of a vocal tract. We embedded such a model into a simple dynamic articulatory system and applied Kalman filtering methods to estimate the articulatory parameters. From these parameters we extract simple rules for speech synthesis. The synthesizer is based on a signal processor system and runs in real time.

THE ARTICULATORY MODEL

The articulatory model is controlled by seven parameters ($a_1,\ldots,a_7$) which determine a discretized 10 tube model of a vocal tract and a 7 tube model of a nasal tract. Parameters a_1 and a_2 describe the tongue body and the shape of the pharynx in a simpified manner by linearly superposing two basic vocal tract shapes and a constant neutral shape. The different places of articulation in the palatal and alveolar region can roughly be described by them. The front palatal and dental articulation is described by parameters a_3 and a_4. a_4 represents the place of the tip of the tongue and a_3 is treated as a parameter of the strength of articulation. a_5 and a_6 determine the radiation from the vocal tract, which is simulated by discretized horns terminating the vocal and nasal tract. a_7 determines the coupling of the nasal tract to the vocal tract.

MODEL FITTING

In order to get transitions of parameters suited for speech synthesis, the model must be fitted to natural speech, that is, we have to find a mapping from an acoustic parameter space to the space of articulatory parameters. It is known from theoretical and practical considerations that this mapping cannot be unique. Thus, we have to restrict ourselves to searching for trajectories that do not contain jumps and that give a representation of measured short-time spectra in a least squares sense. Our method to find this mapping is based on Kalman filtering and Kalman smoothing. We extended the 7-vector of articulatory parameters to the 21-state vector of a dynamic model which is a critically damped 2nd order system with unknown white noise input and unknown control input. Formally:

$$x = (x_1; x_2; u)' \quad , \quad x_1 = (a_1,\ldots,a_7)'$$

x_1: vector of articulatory parameters,
x_2: delayed articulatory parameters,
u: unknown control input.

$$x_{n+1} = \Phi x_n + w_n,$$

w_n: vector of white noise with
$\langle w_n \rangle = 0$, $\langle w_n w_n' \rangle = Q$
The transition matrix is

$$\Phi = \begin{pmatrix} 2A & -A & (I-A)^2 \\ A & 0 & 0 \\ 0 & 0 & I \end{pmatrix}$$

A is a diagonal 7x7 matrix of $\exp(-T/\tau)$;
T: frame length, τ: time constant.

The trajectories of the dynamic system are to be estimated in accordance to natural short time spectra. Thus, based on utterances of one speaker that are digitized with 10 kHz after preemphasis, we estimate ARMA coefficients every 2.5 ms using a Hamming window of 25 ms, and we take the smoothed logarithmic ARMA spectra as a reference. The resulting sequence of short time spectra is called the real measurement process z(t).
The analysis procedure computes the acoustic velocity transfer function of the model's vocal and nasal tract, given the actual estimate of the state $\hat{x}$. The logarithmic spectrum of the transfer function, which is called $h(\hat{x})$, is the model measurement. Then we formally assume that the measurement process z(t) is produced by the model and disturbed with noise:

$$z(t)=h(x(t))+r(t),$$

and thus is related to the 'true' state x. r is a random vector with zero mean and covariance R, it is assumed to contain measurement noise and all model inadequacies as well.
The computation of h(x) requires some computational expense. For this purpose the vocal tract is described by four-terminal networks

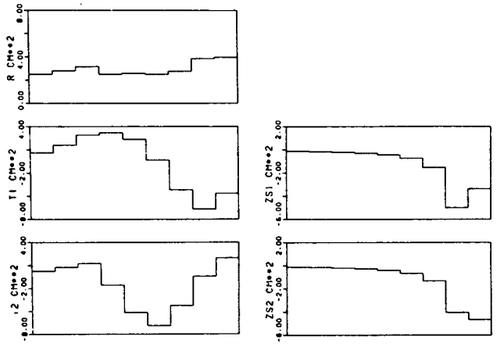

Fig. 1: The articulatory model.
Left: Constant neutral and two basic shapes.
Right: Tongue-tip component in its two extremal points of articulation.

in the view of electrical circuit analogue. The transfer functions from the vocal source and from the fricative source (only one assumed) to the mouth and to the nostrils can be computed, and h(x) is obtained by adding the partial transfer functions and computation of the power spectrum.
As h(x) is a nonlinear function, we can not apply the linear Kalman filter algorithms, an iterated Kalman filter is more convenient. It requires that at each step of iteration the matrix H(x) containing the partial derivatives of h(x) must be evaluated numerically at the actual estimate of x. As h(x) is of high dimension (we take a 64-point FFT), we make use of an inverse covariance Kalman filter. The algorithm works as follows:
Let $\hat{x}^-$ be the estimate of the state at time t with covariance P^-, before the actual measurement z is incorporated. Starting with $x_0=\hat{x}^-$ it iterates:

$$x_{k+1} = x_k + (H'R^{-1}H + (P^-)^{-1})^{-1}$$
$$\cdot(H'R^{-1}(z-h(x_k)) - (P^-)^{-1}(x_k-\hat{x}^-))$$

for k=0,..., K. $\hat{x}^+ := x_K$.

$$P^+ = (H(\hat{x}^+)'R^{-1}H(\hat{x}^+) + (P^-)^{-1})^{-1} .$$

The inverse covariance R^{-1} of the measurement noise r is simply defined as a time varying diagonal matrix. It plays the role of a weighting function for the particular measurements. The algorithm presented above can be extended to a (computationally rather tedious) smoothing algorithm by requiring that the estimate of state $\hat{x}$ at time n is not only determined by the measurement history up to time n but also by future measurements up to time n+m. For one update of the smoothing algorithm a Kalman filter, starting with the present state of the smoother, first runs forward up to time n+m, then, using the adjoined backward dynamic model, backwards to time n. At each measurement it makes an update of its state. The state of the smoother is then updated by assembling its actual state and the state of the backwards filter.

INCORPORATION OF ARTICULATORY CONSTRAINTS

As could be expected, the described procedure works sufficiently for most of the pure vocalic transitions. For voiced-unvoiced transitions and for nasalized transitions some constraints have to be incorporated into the model. We do this in a straightforward way: If we know, e.g. that the velum must be open at an interval and closed at another, we 'tell' the Kalman filter that we measure the behavior of the velum parameter, that is, we include the pseudo-measurements into the general measurement history z(t) and give them more or less influence by defining the corresponding entries in the diagonal of R^{-1}. In a similar manner all parameters that are functionally related to the state of the model can be prescribed, such as place of articulation and strength of fricative excitation.

ANALYSIS OF FITTED ARTICULATORY TRAJECTORIES

For analysing and testing the fitted articulatory trajectories, the vocal tract model was implemented on a signal processor system. This system consists mainly of a fast signal processor TMS-32010 from Texas Instruments, a fast parallel interface and a 16-bit D/A-converter. The signal processor is fast enough to calculate the vocal tract model in real time. The articulatory model as well as the "articulator-to-filter" transformations are done on a laboratory computer (Gould 32/9705). The filter parameters are transferred to the signal processor system at a frame rate of 200/s.
It was possible to resynthesize intervals of an adapted trajectory as well as fixed parameter sets. On this way it was possible to extract subjectively constant parts of vowels or consonants.
Fig. 4 shows extracted vowels in the plane of the first two articulatory parameters (a_1, a_2). If we interprete the first parameter as a front versus back, the second as a high versus low parameter, the positions of the vowels are close tongue hump positions. In this plane some vowels like /e:/ and /i:/ or /a/ and /ə/ are very close. They differ from each other in the third (tongue tip) parameter. For example this parameter is higher for /i:/ than for /e:/ which can be interpreted as an articulation more in the front of the tract.
For the most cases the transitions between phonemes are regular. Attempts to find a fixed dynamics for the articulation failed. Every transition seemed to have a different dynamics. The best and most easy description of the transitions was linear interpolation or a critically damped second order system with varying parameters.

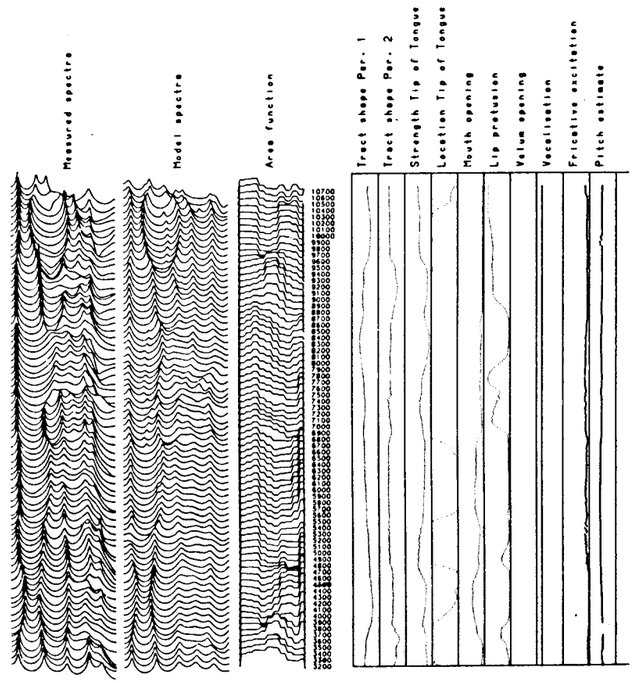

Fig. 2.
Fitted utterance /la:le:li:lo:/.
Time runs from bottom to top.
Velum opening was fixed to zero.

For both figures on this page:
The vocalization parameter is chosen 'by hand'. It determines the energy of the input signal of the four terminal network which represents the transfer from the glottis to the radiation. This parameter also determines the glottal shunt, which increases for low vocalization. The fricative excitation parameter is computed by the fitting procedure based on simple assumptions about the relation between closest constriction and turbulent noise strength.

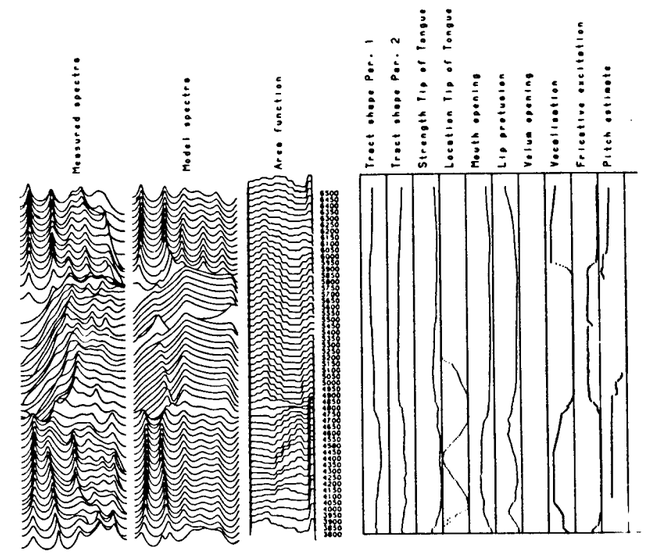

Fig. 3.
Fitted utterance /taʃə/.
Starting at the voice onset of /ta/

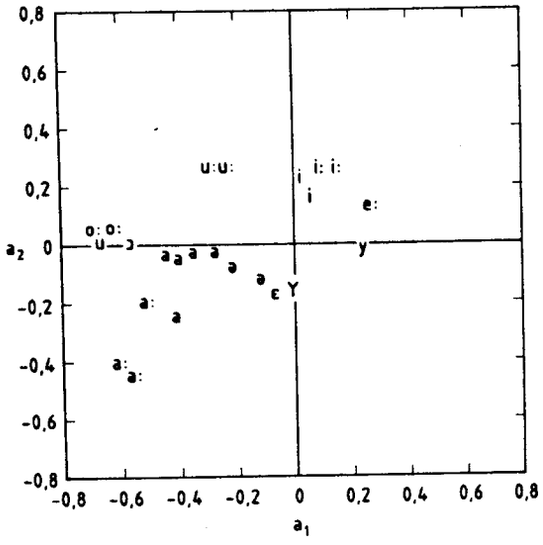

Fig. 4. Points of articulation of some German vowels in the plane of a_1,a_2.

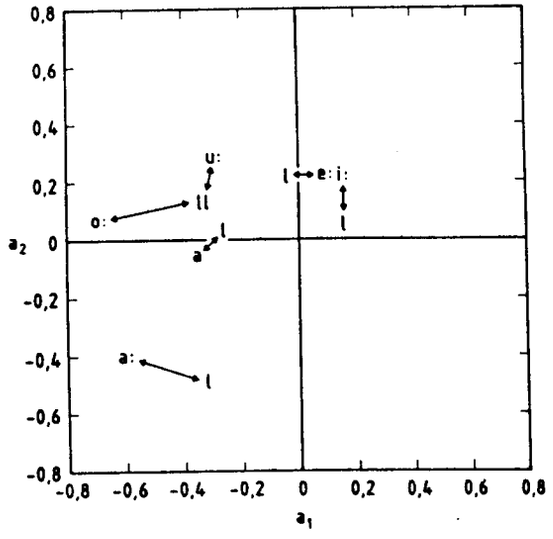

Fig. 5. Points of articulation for liquid /l/ embedded in different vowel context like /a:la:/.

The adapted articulatory parameters show strong effects of coarticulation. We examined different VCV-transitions with the liquid /l/ like /a:la:/,/e:le:/... Fig. x shows the points of articulation of the l's in the plane of the first two parameters. The articulation is close to the position of the surrounding vowel. The main articulation is done by the third, the tongue tip parameter. For all shown articulations this parameter produces a constriction of the 8'th or the 9'th tube segment of about 0.5 cm^2. This effect can be seen if we look at the articulation of the l's in the plane of the first and third parameter where they lay on a straight line. Similar results can be found for other consonants, for example for the nasal /n/.

Strong articulatory effects can be found for the articulation of plosives which is close to the following vowel. This effect is extreme for the plosives /p/,/b/ which are articulated nearly in the same way as the following vowel with a mouth opening of zero.

SYNTHESIS

For speech synthesis purposes we stored 36 parameter vectors of different German phonemes in a table. This table contains also information about voiced or unvoiced exitation, the strength of the fricative exitation, duration, voice onset times etc. Only for few transitions it is possible to make a synthesis by interpolating between these parameter vectors. So we add further vectors which all represent the same consonant to describe coarticulatory effects, such as the articulations of different /l/ in Fig. x. If we synthesize a transition from a vowel to a consonant we interpolate to the representant which is nearest to the vector of the vowel. If we want to synthesise a VCV-transition we interpolate during the 'constant' consonant part to the representative which is nearest to the following vowel vector. These rules are described by addresses to the consonant vectors written in a 36x36 matrix. This matrix also contains information about the duration of the transition.

Parameters like voice onset times for plosives and strength of fricative exitation are chosen subjectively.