Dept. for Speech, Music and Hearing
Quarterly Progress and Status Report

A new algorithm for speech synthesis based on vocal tract modeling

Lin, Q. and Fant, G.

journal: STL-QPSR
volume: 31
number: 2-3
year: 1990
pages: 045-052

KTH Computer Science and Communication

http://www.speech.kth.se/qpsr

A NEW ALGORITHM FOR SPEECH SYNTHESIS BASED ON VOCAL TRACT MODELING*

Qiguang Lin and Gunnar Fant

Abstract

A new algorithm for articulatory speech synthesis is described in this paper. The algorithm constitutes two main parts: a detailed frequency domain modeling of the vocal tract and a data transform from the frequency domain to the time domain. A computer model was developed to simulate the vocal tract acoustics. The model incorporates all known important components of the vocal system and computes the transfer function between the lip/nostril output and the acoustic source. By decomposing the obtained transfer function into its numerator and denominator, frequencies and bandwidths of resonances (and of anti-resonances if any) can be determined. The transfer function can next be written as a partial fraction expansion series in terms of calculated residues at the poles and can be approximated by retaining the first few terms in the series. Usually, each of these terms is a second-order module and corresponds to an elementary formant resonance. Formants are thus connected in parallel. The time-domain output is obtained by the inverse Laplace transform. Compared with other synthesis methods, this vocal tract oriented synthesis strategy has a number of advantages. For instance, the frequency dependency of loss elements of the vocal tract is preserved and accurate frequency responses can be reproduced. It is also computationally efficient relative to a direct convolution method. Examples of spectrum matching are presented to discuss the properties of the proposed algorithm. The algorithm has been incorporated in an articulatory-based speech synthesis system currently under development at KTH.

1. ACOUSTIC MODELING OF THE SPEECH TRANSMISSION SYSTEM

The speech transmission system consists of the tracheal tubes, larynx, pharyngeal cavities, vocal chambers, and nasal passages. Conventionally, the system of varying cross-sectional dimensions can be approximated by consecutive cylinders, and a planar wave propagation inside the system is assumed. Under this assumption, the transmission properties of each cylinder can be represented in analog to homogeneous transmission line, usually in terms of a T-network. The vocal/nasal tracts are terminated by a radiation impedance at their front end. See Fig. 1. This figure also shows a few additional shunt arms. They have been incorporated to simulate nasal sinuses to provide a good picture of the acoustic characteristics of the nasal system and to simulate the effects of the yielding wall.

A computer model for dealing with the vocal system shown in Fig. 1 has been developed (Lin, 1990), based on the design of Badin & Fant (1984). The system is simulated in the frequency domain, taking advantage of the more convenient and accurate modeling of losses and radiation. Some major features of the model can be summarized as follows [refer to Lin (1990) for more details]:

a) The numerator and the denominator of the calculated transfer function are decomposed so that data of poles (and zeros if any) can be correctly determined. For an ideal all-pole model, the numerator is a constant. Otherwise it becomes a function of frequency, for instance, when there are extra elements in the system. For a shunting ele-

* Paper presented at the 120th Meeting of the Acoustical Society of America, San Diego, November, 1990.

ment, the zero of the transmission occurs at frequency where the shunt has a zero, and for a serial element, the zero of the transmission occurs at frequency where the element has a pole. The location of the transmission zeros will be changed if one sums up the simultaneous outputs from different ports.

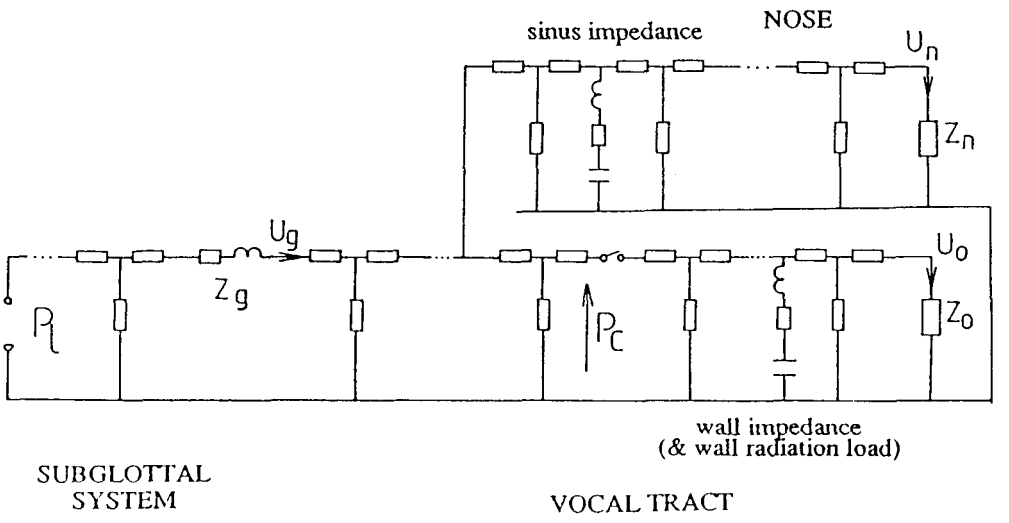

Fig. 1. Network representation of the human speech-producing system.

b) An on-line graphic display facility is provided, see Fig. 2. After each running, it is able to alter some of the simulation conditions and an updated transfer function is computed. This is a convenient tool to study the relations between articulation and acoustics. For example, it is clearly seen in Fig. 2 that a small perturbation of area function in the vicinity of the velum affects only the location of higher formants.

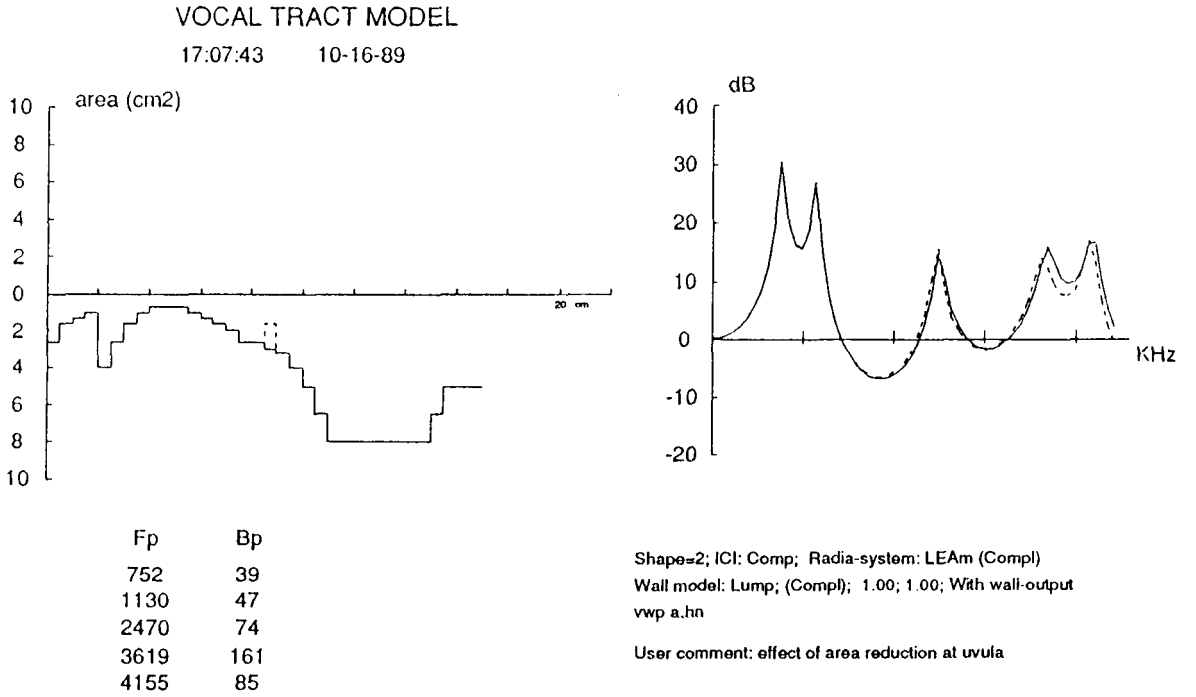

Fig. 2. An example of on-line graphic display when running program TRACT. Note that F_4 and F_5 are shifted upwards when the area in the vicinity of the velum is enlarged.

c) Different models for the radiation impedance and for the wall impedance have been implemented. For a simultaneous radiation occurring at various ports (the lips, the nostrils, and the vibrating walls) the volume velocities are superposed linearly, disregarding the spatial phase difference.

d) The residue data at the poles of the transmission are determined. They will be used later to expand the transfer function into a partial fraction series (see further Section 2).

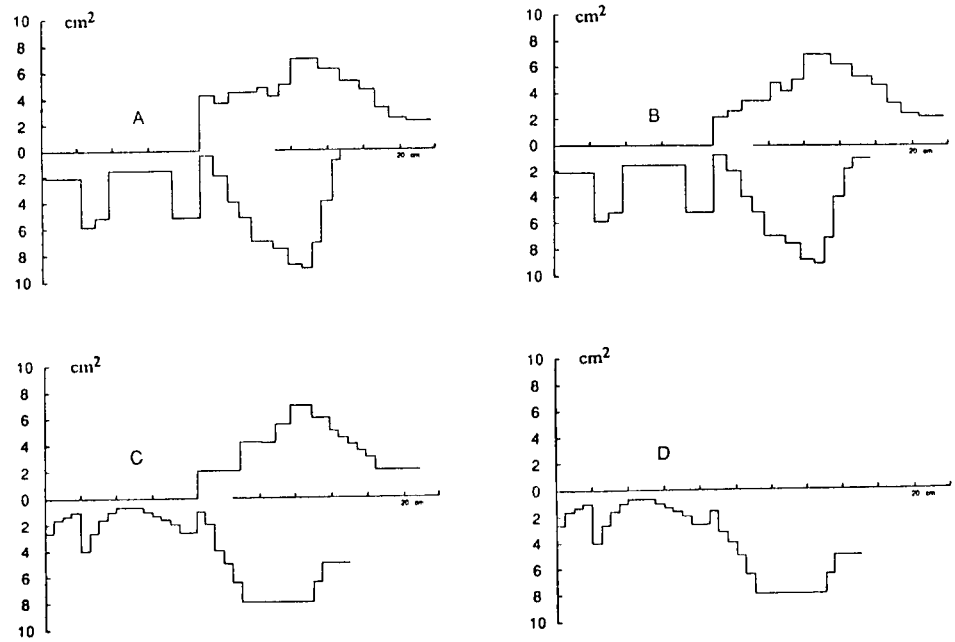

Fig. 3. Plausible area functions of the syllable [ma], sampled at four instants in time. The abscissa, 2 cm per division, has its origin at the glottal end. A: An ideal nasal consonant [m]; B and C: intermediate phases of transition; and D: an ideal oral vowel [a].

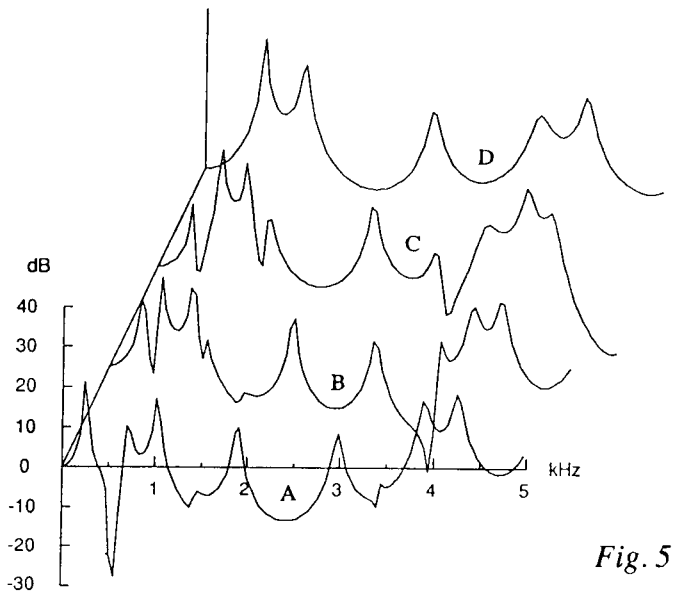

Fig. 4. Transfer functions of the syllable [ma], refer to Fig. 3.

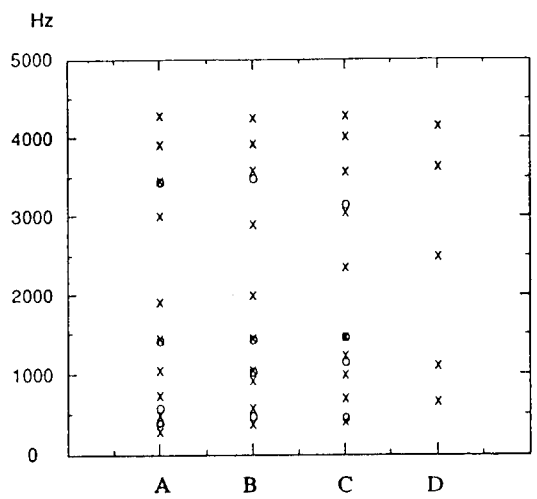

Fig. 5. Plot of zeros (o) and poles (x), based on Fig. 4.

Based the X-ray tracing data from Fant (1960), the model has been used to simulate various sounds, such as vowels, fricatives, liquids, nasal murmurs and nasalized sounds. Representative results have been achieved. An example is shown in Figs. 3, 4, and 5, where the transition from /m/ to /a/ is simulated. Fig. 3 shows the underlying area functions, while Figs. 4 and 5 are plots of the spectra and the zero/pole distribution, respectively.

The particular spectral envelope around 500 Hz for Curve A in Fig. 4 is the result of a zero-pole-zero combination. These two zeros are located closely, so a small frequency scanning increment should be used so as not to miss any of them. In Fig. 5, the positions of the first and third zeros are rather stable from A to B to C. These zeros are related to the zeros of the nasal sinuses. The other two zeros are caused by the entire branched system. Their trajectories vary as the configuration of the system changes.

When the velum is raised and thereby discouples the nasal tract from the system, all of the zeros will then be cancelled by their bound poles, see D in Fig. 5. It can also be seen that the coupled nasal- and vocal tracts have more poles than the vocal tract alone does. When excluding the pole-zero pairs, there are 6 poles for A, B, and C below 5000 Hz and 5 poles for D. This result is expected since the effective length is different.

2. THE NEW ALGORITHM FOR SYNTHESIS IN THE TIME DOMAIN

A new algorithm to interface the above frequency-domain analysis with the time-domain synthesis has been developed. Once the data of poles and the associated residues are known, the transfer function can be written as a partial fraction expansion series. Theoretically, there are infinitely many terms in the series. However, within a limited frequency band of interest, the transfer function can be approximated by explicitly retaining the first few terms. If the order of the numerator of the transfer function is lower than that of the denominator and if there are no real poles, then each of the terms corresponds to an elementary formant resonance. Formants are thus connected in parallel. By performing inverse Laplace-transform, the time-domain output is obtained.

Mathematically, the calculation of residues at poles can, in the S-plane, be interpreted as vector products as depicted in Fig. 6. The desired residue equals the product of vectors from all zeros to the pole under investigation divided by the product of vectors from all other poles to that pole. If there is no zero, the dividend is of course equal to 1.

However, the transfer function is calculated along the $j\omega$-axis. All relevant vectors are pointing at $j\omega_n$ instead at $s_n=\sigma_n+j\omega_n$, which brings about error in the residue calculation. The error is marginal only if the neighbouring poles/zeros are located quite apart (relative to the distance between s_n and $j\omega_n$) and if the real part of the current pole is relatively small. Otherwise a proper correction is needed to move back the shifted vectors. It is found that a correction from the nearest pole/zero is usually sufficient. Note that if no such corrections are necessary, it is then unnecessary to determine the zeros, Lin (1990).

Let $A_n=\alpha_n+j\beta_n$ be the (complex) residue at the nth pole, we then have:

$$H(s)=\sum_{n=1}^{N}[H_n(s)]=\sum_{n=1}^{N}\frac{2\cdot\alpha_n\cdot(s-\sigma_n)-2\cdot\beta_n\cdot\omega_n}{s^2-2\cdot\sigma_n\cdot s+\omega_{on}{}^2}\,, \qquad (1)$$

with

$$H_n(s)=\frac{2\cdot\alpha_n\cdot(s-\sigma_n)-2\cdot\beta_n\cdot\omega_n}{s^2-2\cdot\sigma_n\cdot s+\omega_{on}{}^2}\,, \qquad (2)$$

where $\omega_{on}{}^2=\sigma_n{}^2+\omega_n{}^2$, and N is the number of the formants retained.

The correction factor is given by:

$$C_{i,n} = \frac{s_n - s_i}{j\omega_n - s_i} = \frac{(\sigma_n - \sigma_i) + j(\omega_n - \omega_i)}{-\sigma_i + j(\omega_n - \omega_i)}, \qquad (3)$$

where $s_i = \sigma_i + j\omega_i$ may be a neighbouring zero or pole to the current pole. The denominator of Eq. (3) is constructed to eliminate the shifted vector, and the numerator specifies the substituting vector, or the original one.

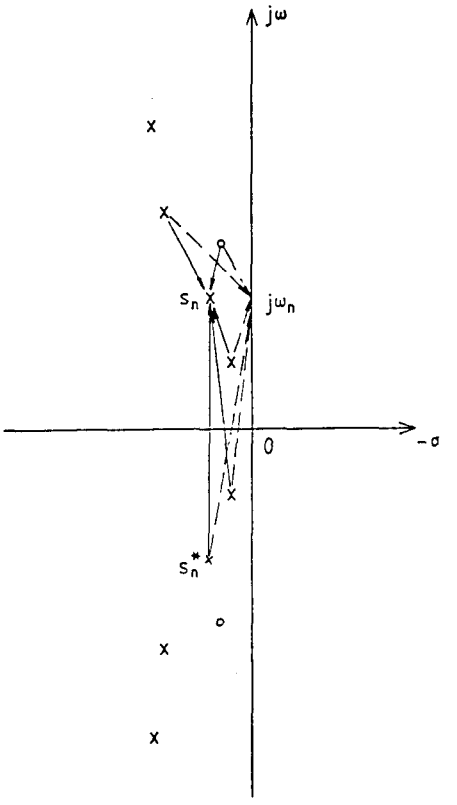

Fig. 6. Vector representation of the residue calculation. Vectors in solid lines are the actual ones, and vectors in dashed lines are those obtainable from a direct vocal tract computation.

The correction factor $C_{i,n}$ can also be used to estimate the transfer function for a lossy system based on the residue data calculated from the corresponding lossless system (Lin, 1990). This can save computation time, provided the lost information on the bandwidths can be estimated by some empirical formulations.

It is found, at least for the transfer function of vowels, that the complex residue $A_n = \alpha_n + j\beta_n$ can be approximated by its dominating part $j\beta_n$. When α_n is ignored, Eq. (2) reduces to:

$$H_n(s) = -\frac{2 \cdot \beta_n \cdot \omega_n}{s^2 - 2 \cdot \sigma_n \cdot s + \omega_{on}^2}. \qquad (4)$$

or:

$$H_n(s) = (-1)^n \cdot \frac{-2 \cdot |H(\omega_n)| \cdot \sigma_n \cdot \omega_n}{s^2 - 2 \cdot \sigma_n \cdot s + \omega_{on}^2}, \qquad (5)$$

where $|H(\omega_n)|$ denotes the spectrum level at ω_n (the nth formant level), which is intimately related to the associated residue. The numerator of $H_n(s)$ is now independent of the frequency. Such independence was actually the basic assumption of the Holmes parallel synthesis system, see Holmes (1983) for the detail of his design. Eq. (5) indicates that the simplified parallel system is specified exclusively by the formant frequency, bandwidth, and its amplitude. They can be estimated from a section display of a narrow-band spectrogram, by means of an analysis-by-synthesis routine. Note that the polarity alternates with pole number n in Eq. (5). When there enter zeros in the transfer function, the polarity alternates with pole plus zero number.

Figs. 7 and 8 give two examples of spectrum matching result, one for a vocalic sound and the other for a nasalized vowel. It is shown in Fig. 8 that a better fit is obtained when the correction is applied. More examples can be found in Lin (1990).

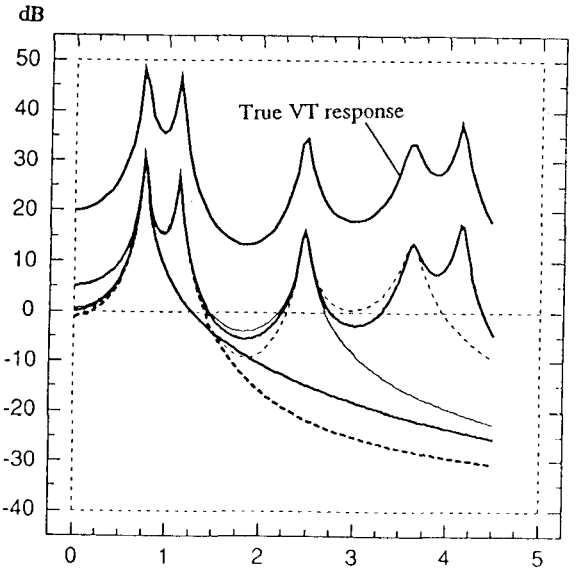

Fig. 7. Magnitude spectrum of vowel [a]. Accumulative output of formants, in comparison with the true vocal tract response.

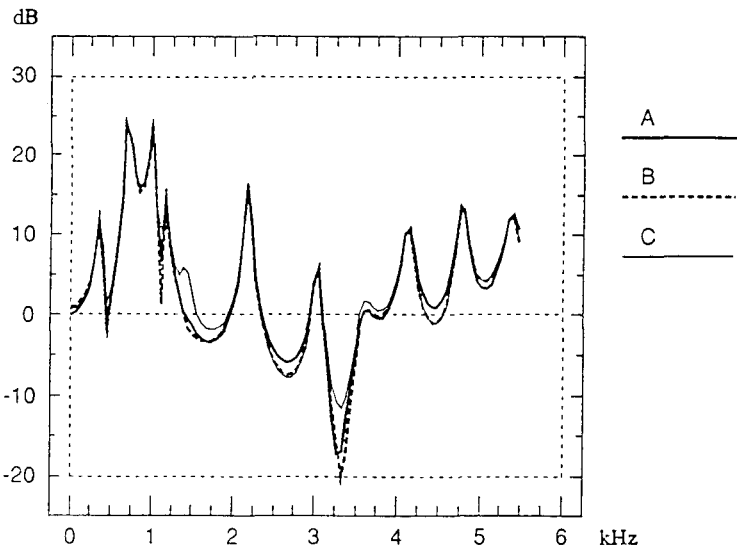

Fig. 8. Transfer functions for a nasalized vowel. The first 11 poles are included. Curve A: Calculated transfer function; Curve B: Resynthesized transfer function, with the residues being corrected; Curve C: Same as Curve B but with no residue correction. Observe that the false hump around 1.4 kHz is removed by applying the correction term $C_{i,n}$, see text.

When formant circuits are connected in cascade, the spectrum of a vowel-like sound is defined solely by the formant frequencies and bandwidths, in addition to the term of higher pole correction, HPC (Fant, 1960). In terms of residue calculation, the cascade connection may be changed to an equivalent parallel connection. Fig. 9 presents such an example. Thus, by this technique it is able, though not necessary, to avoid the cascade configuration in a formant speech synthesizer. The control strategy remains, however, the same as if a cascade structure of formant was used.

3. COMPARISON WITH OTHER TYPES OF SYNTHESIS SYSTEM

In the Holmes design (Holmes, 1983), the spectrum match is based on the spectrum of radiated speech. The present system models, however, the transfer function only. The residues can be correctly determined (assuming that the underlying area function is known). Therefore, no formant shaping filters are required to avoid spectrum distortion. On the other hand, Holmes' design allows a formant amplitude to vary over a certain range thanks to the introduction of shaping filters. The present system does not generally have this feasibility, but by introducing a notch filter one can alter formant levels without disturbing the spectrum (Lin, 1990).

Information of all poles and zeros has been included when the residues are calculated. The higher pole/zero correction is therefore inherently and accurately preserved. This is an advantage over the conventional cascade formant synthesizer, where one has to specify the HPC term, explicitly or implicitly. Alternatively, the algorithm suggests a new method for specifying the HPC term.

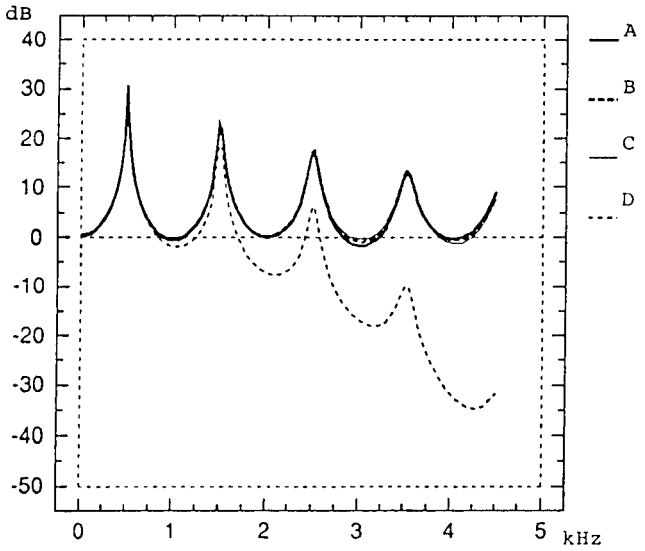

Fig. 9.Transfer functions of the neutral tube, generated by the following methods: A: Synthesized transfer function by the parallel system; the residue data have been directly determined from a vocal tract computation; B: Same as A, but the residue data have been estimated based on a cascade model of formants. C: Synthesized transfer function by a cascade model; The effect of the higher pole contribution is included; D: Same as C, but without the higher pole contribution.

The proposed system can also be compared with time-domain analogs, such as the transmission line analog or the reflection-type line analog, which will directly generate the time-domain functions. The present system may have difficulty in following rapid dynamic aspects of the articulation events. However, it has the advantage of i) preserving the accurate frequency-dependency of loss elements and of boundary conditions; ii) no constraints on the variation of the overall length of the vocal tract; iii) less computation time.

The computational efficiency is also a merit of the present system in comparison with a direct convolution method, see, for instance, Sondhi and Schroeter (1986). In the convolution method, the transfer function is first calculated by a frequency domain analysis and then its impulse response is determined by the inverse Fourier transform. The speech output is obtained by convoluting the resultant sequence with the source of excitation. In the proposed algorithm, the convolution is accomplished by a filtering process.

4. CONCLUDING REMARKS

A new algorithm for speech synthesis based on the vocal tract simulation has been described. Satisfactory results of spectrum match have been achieved. But it remains to be extended so that it can also deal with real poles. Such an extension may be of importance for the simulation of fricatives.

The algorithm can also be utilized to specify the driving-point impedance seen downstream from the glottis in terms of formant frequencies, bandwidths and residues at the poles. This

specification is useful in studying the effects of the interaction between the glottal source flow and the acoustic load provided by the vocal tract during phonation.

The algorithm has been incorporated in an articulatory-based speech synthesis system currently under development at KTH.

ACKNOWLEDGEMENTS

The work was in part supported by the grants from the Swedish Board for Technical Development. The scholarship from the L M Ericsson Telephone Company (Stiftelsen för främjande av elektroteknisk forskning) to Qiguang Lin is gratefully acknowledged.

References

Badin, P. & Fant, G. (1984): "Notes on vocal tract computation," *STL-QPSR* No. 2-3, pp. 53-107.

Fant, G. (1960): *Acoustic Theory of Speech Production*, Mouton, The Hague.

Holmes, J.N. (1983): "Formant synthesizers cascade or parallel?" *Speech Comm.* **2**, pp. 251-273.

Lin, Q. (1990): *Speech Production Theory and Articulatory Speech Synthesis*, Ph.D. thesis, Dept. of Speech Communication and Music Acoustics, Royal Institute of Technology, Stockholm.

Sondhi, M.M. & Schroeter, J. (1986): "A nonlinear articulatory speech synthesizer using both time- and frequency-domain elements", *Proc. ICASSP-Tokyo*, pp. 1999-2002.