IMPROVED VOCAL TRACT MODEL FOR SPEECH SYNTHESIS

Minsheng Liu, Arild Lacroix
Institut für Angewandte Physik; University of Frankfurt
Robert-Mayer-Str.2-4; D-60325 Frankfurt am Main,Germany
e-mail:Liu@iap.uni-frankfurt.de,
Lacroix@iap.uni-frankfurt.de

ABSTRACT

Speech synthesis of nasal and non-nasal speech sounds are studied on the basis of an improved model where a nasal tract is included in the vocal tract. The transfer function of the model is analysed. Because of the closure of the oral tract, the three-port adaptor at the velum is reduced to a two-port adaptor, so that the model parameters can be estimated by inverse filtering from the speech signal. Moreover this method is applied to investigate nasalization of vowels.

1 Introduction

The speech signal can be reproduced in a number of ways. The acoustic model of the vocal tract, which incorporates the physiological and physical constraints of the human speech production mechanism and is able to improve the naturalness of synthetic speech, is an attractive way. Since the early attempts[1, 2], a good speech quality of the nasal sounds, which requires a pole-zero model has not been achieved by acoustic models using the oral tract alone.

In this paper the multi-tube model is improved by coupling a nasal tract, in which the velum is represented by a time varying three-port-adaptor. The transfer function is derived, then its properties are studied, so that this model can be used for the synthesis of nasal and non-nasal sounds. The inverse filtering method and an optimization algorithm are used to obtain an estimation of the parameters of nasal and vocal tract and of the velum.

2 Analysis of the Extended Model

The model [2] is extended by coupling the nasal tract at the velum. This model is adapted to produce nasals and non-nasal speech sounds.

2.1 Improved Model for Nasal Sounds

In the production of the nasal consonants [m], [n] and [η], the velum is lowered to couple the nasal tract to the vocal tract, while the oral tract is closed and its length depends on the nasal consonant. In this case the configuration of the vocal tract is characterized as shown in Fig. 1, where the oral tract is represented by two tubes; the number of sections of the first tube is 1, while for the second tube is n.

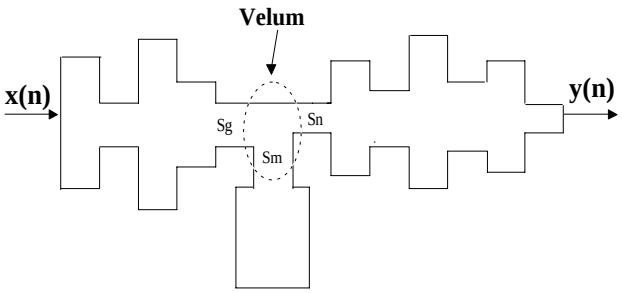

Fig. 1 Tube model for production of nasals,

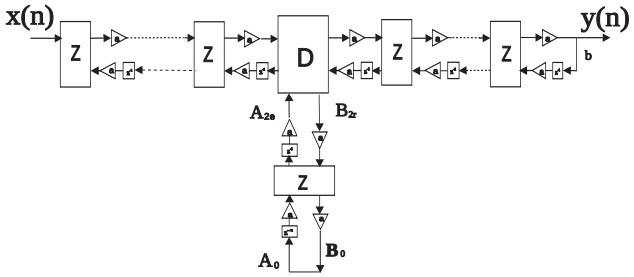

Fig. 2 corresponding discrete-time system

The corresponding discrete system of this model is shown in Fig.2. The transfer function of the vocal tract will be characterized by a set of resonances and anti-resonances. Fig. 2 shows that the discrete time tubes are connected by two- and three-port adaptors, where the α are the attenuation coefficients. The two-port adaptor can be described by the transfer and by the scattering matrix [3]. The three-port-adaptor at the velume can be described by a 3×3 scattering matrix

from the continuity equations for pressure and flow. The scattering matrix for the flow is

$$\begin{pmatrix} B_{1r} \\ B_{2r} \\ B_{3r} \end{pmatrix} = \mathbf{S} \begin{pmatrix} A_{1e} \\ A_{2e} \\ A_{3e} \end{pmatrix} \tag{1}$$

with

$$\mathbf{S} = \frac{\mathbf{1}}{\sum_{\mathbf{i=1}}^{\mathbf{3}} \mathbf{s_i}} \begin{pmatrix} S_g - S_m - S_n & 2S_g & 2S_g \\ 2S_m & S_m - S_g - S_n & 2S_m \\ 2S_n & 2S_n & S_n - S_g - S_m \end{pmatrix} \tag{2}$$

where the S_g, S_m and S_n are the three cross-sectional areas of the pharynx, the mouth cavity and the nose cavity at the velum.

Because the oral tract is closed in front, the pressure at the closure is maximum and the acoustic wave is completely reflected, namely $A_0 = B_0$, and we obtain

$$B_{2r} = \frac{z^{-1}T_{11} + z^{-1}T_{12}}{T_{11} + T_{22}} A_{2e}, \tag{3}$$

where T_{11}, T_{12} and T_{22} are elements of the scattering transfer matrix of the closed oral tract. Therefore the 3x3 scattering matrix at the velum can be reduced to a 2x2 matrix

$$\begin{pmatrix} B_{1r} \\ B_{2r} \end{pmatrix} = \tilde{\mathbf{T}}_\mathbf{D} \begin{pmatrix} A_{1e} \\ A_{2e} \end{pmatrix} \tag{4}$$

with

$$\tilde{\mathbf{T}}_\mathbf{D} = \frac{\tilde{\mathbf{T}}_\mathbf{DD}}{-S_{22}z^{-(n+1)} + r_0 z^{-n} - r_0 S_{22} z^{-1} + 1}. \tag{5}$$

$\tilde{\mathbf{T}}_\mathbf{DD}$ is a 2x2 matrix; the elements of it are polynomials. S_{22} is the element of the 3x3 scattering matrix at the velum and r_0 is the reflection coefficient of the oral tract at the velum junction. So the implementation of a three-port-adaptor can be directly expressed not by a 3x3 matrix but by a 2x2 matrix. Therefore we make use of the inverse filter of the model shown in Fig. 3. The transfer function can be derived as follows

$$H(z) = \frac{-S_{22}z^{-(n+1)} + r_0 z^{-n} - r_0 S_{22} z^{-1} + 1}{\begin{pmatrix} a & 1 \end{pmatrix} \tilde{\mathbf{T}}_\mathbf{12} \ldots \tilde{\mathbf{T}}_\mathbf{8} \tilde{\mathbf{T}}_\mathbf{0} \tilde{\mathbf{T}}_\mathbf{DD} \tilde{\mathbf{T}}_\mathbf{6} \ldots \tilde{\mathbf{T}}_\mathbf{1} \begin{pmatrix} b \\ 1 \end{pmatrix}} \tag{6}$$

where $\tilde{\mathbf{T}}_\mathbf{i}$ $(i : 1 \ldots 12)$ are the scattering transfer matrices of pharynx and nasal cavity, while the velum and the simplified oral tract are included in $\tilde{\mathbf{T}}_\mathbf{DD}$.

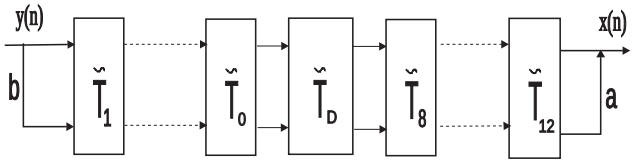

Fig. 3 the inverse filter of the model

It is clear that the vocal tract model is changed into a pole-zero model, if the nasal tract is coupled to the vocal tract. After the parameters of the model are estimated, we can calculate the zero locations from the polynomial, which is the numerator of the transfer function(6):

$$z^{12-(n+1)}(z^{n+1} - r_0 S_{22} z^n + r_0 z - S_{22}) = 0. \tag{7}$$

This eqation shows, that zero locations are determined by the position of the velum and the configuration of the oral tract.

2.2 Model for Non-Nasal Speech Sounds

If the cross-sectional area S_m of the oral tract at the velum is decreased to zero, then the oral tract is eliminated. It can be shown that the reduced two-port adaptor $\tilde{\mathbf{T}}_\mathbf{D}$ becomes

$$\tilde{\mathbf{T}}_\mathbf{D} = K_0 \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} \tag{8}$$

with

$$\tilde{\mathbf{T}}_\mathbf{7} = \tilde{\mathbf{T}}_\mathbf{0} \tilde{\mathbf{T}}_\mathbf{D} = K_0 \begin{pmatrix} z^{-1} & r \\ rz^{-1} & 1 \end{pmatrix}. \tag{9}$$

This is a scattering transfer matrix and the transfer function becomes

$$H(z) = \frac{1}{\begin{pmatrix} a & 1 \end{pmatrix} \tilde{\mathbf{T}}_\mathbf{12} \tilde{\mathbf{T}}_\mathbf{11} \ldots \tilde{\mathbf{T}}_\mathbf{7} \tilde{\mathbf{T}}_\mathbf{6} \tilde{\mathbf{T}}_\mathbf{5} \ldots \tilde{\mathbf{T}}_\mathbf{1} \begin{pmatrix} b \\ 1 \end{pmatrix}} \tag{10}$$

having only poles and the remaining model is an all-pole-model.

3 Estimation of the Parameters and Synthesizer

The sampling frequency is chosen to be 8 kHz so that each elementary tube of the model is 21 mm long. The system for speech synthesis is composed of 12 tubes. The pharynx between glottis and velum consists of five tubes and the nasal tract has seven tubes. The length of the tubes for the oral tract is chosen according to the nasal consonant.

For each section two parameters α and r have to be estimated, while there are three parameters for the reduced

three-port adaptor.
The model parameters for speech synthesis are analysed and optimized from natural speech. The method is based on the minimization of the signal power behind each section in the cascade of the inverse filter [4]. The parameters of the reduced three-port adaptor are estimated by an adaptive algorithm [5].
To get a good speech quality it is important to use a sythesizer structure which is strictly inverse to the analyzer.

4 Results and Discussion

We choose 6.3 cm as the length of the oral tract for the nasal [m] that is closed at the lips corresponding to three tube sections. Fig. 4 shows the fourier spectrum and the magnitude response (6) of the nasal [m] . It indicates that the nasal [m] has five resonances. The magnitude response shows two antiresonances located at 812 Hz and 4000 Hz; the antiresonance at 812 Hz can be seen from the group delay curve.
We can calculate zero locations by using the estimated parameters from equation (7). The result indicates that there are three zeros, conjugate complex zeros $z_{1,2} = 0.857e^{\pm j40.1^{\circ}}$ located at 891 Hz and a zero $z_3 = -0.975$ at 4000 Hz. We can see that the spectrum of the speech signal is in good coincidence with the spectrum of a synthesized speech signal shown in Fig. 5.

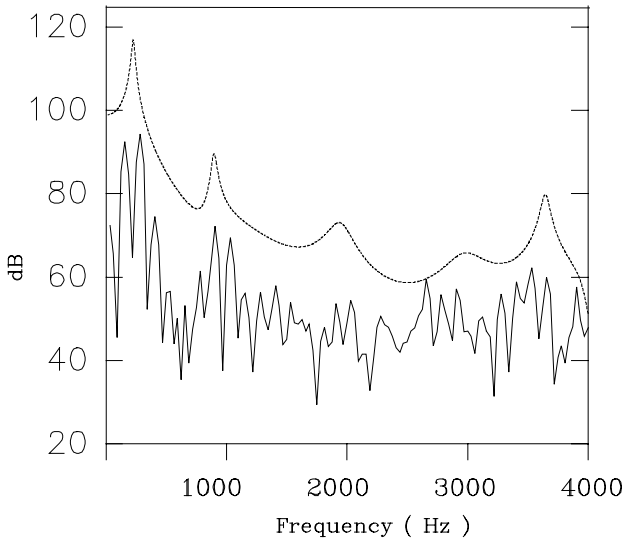

Fig. 4 Magnitude of the transfer function of the nasal [m] (dotted line) and related fourier spectrum.

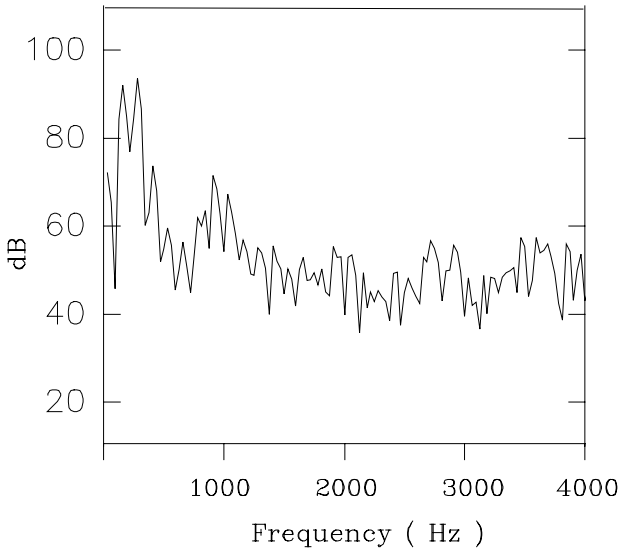

Fig. 5 Spectrum of the synthesized signal for the nasal [m].

The oral tract length for the nasal [n], in which a closure is formed at the middle of the oral tract, is chosen as 4.2 cm corresponding to two tube sections. The analytic results are displayed in Fig. 6. The nasal [n] has also five resonances and a conjugate complex zero at 1058 Hz, which can be detected from the group delay of the frequency response.

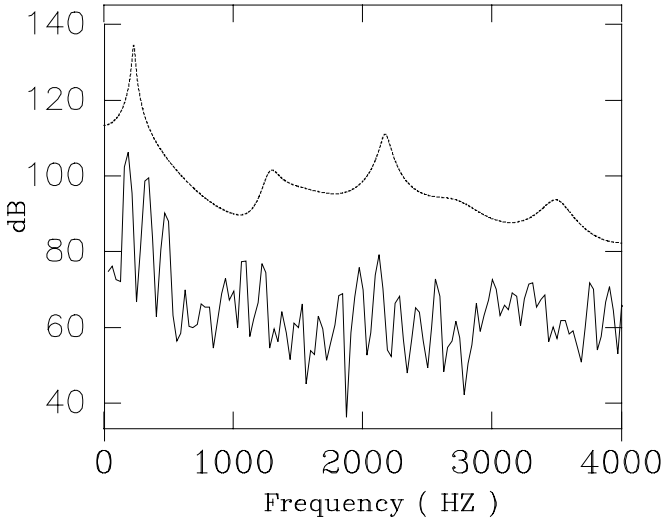

Fig. 6 Magnitude of the frequency response for the nasal [n] (dotted line) and spectrum of synthesized speech signal.

Nasalization of the vowels has been produced by inserting a pole-zero pairs near the first resonance [6]; this implies that the coupling of the nasal tract to the vocal tract results in pole-zero pairs in the transfer function. We have also analysed nasalizations by using the transfer function (6). In this case the reduced three-port

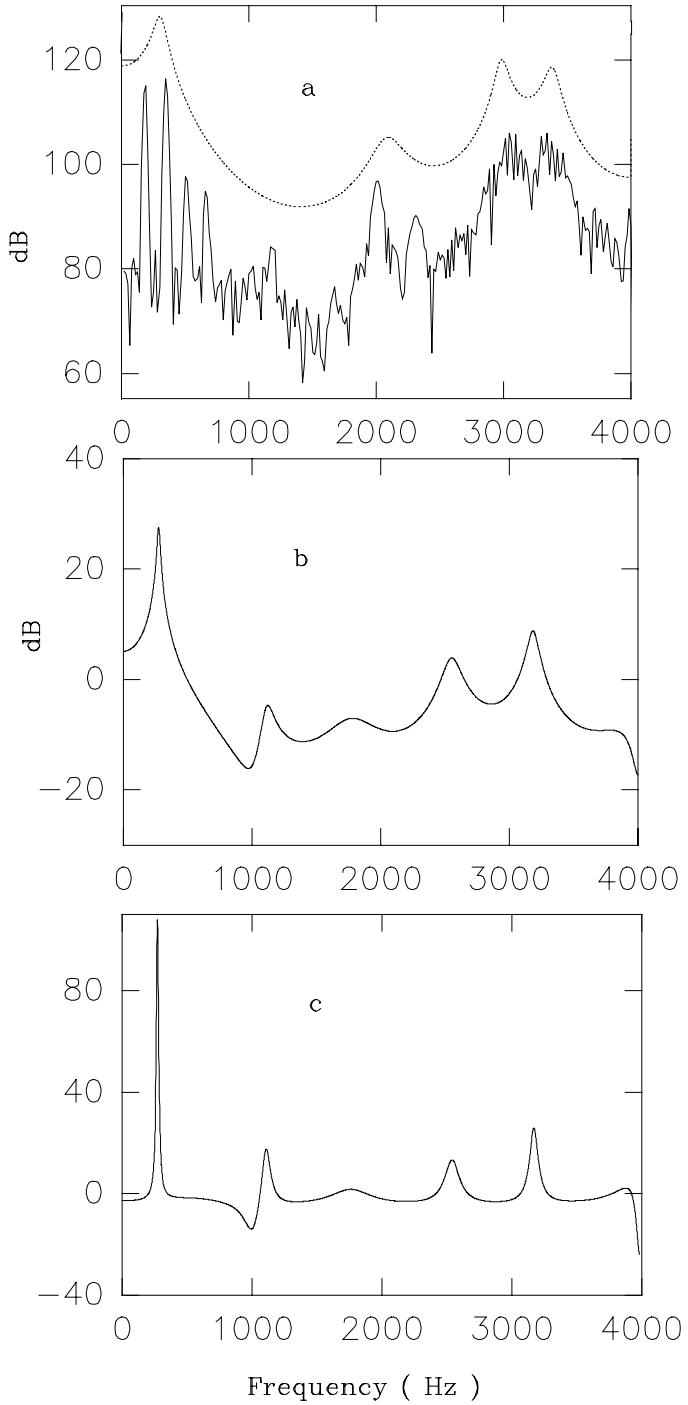

Fig. 7 Results of the nasalization [i]: a) magnitude of the frequency response (dotted line) and fourier spectrum; b) and c) Magnitude and group delay of frequency response of the improved model.

adaptor $\tilde{\mathbf{T}}_{\mathbf{D}}$ describes antiresonances and resonances introducing zeros in the transfer function; if the oral tract has two tubes, namely, $n = 1$, pole-zero pairs are introduced by the reduced three-port adaptor $\tilde{\mathbf{T}}_{\mathbf{D}}$.

Fig. 7 presents as an example for nasalization the vowel [i]. Fig. 7 a) shows the fourier spectrum and the magnitude response of the all-pole model. It indicates that the fourier spectrum has five resonances, but the magnitude response of the all-pole model displays only four (dotted line). The magnitude and group delay of the frequency response of the improved model in Fig. 7 b) and 7 c) show that pole-zero pairs appear at 1029 Hz in comparison to the result of the all-pole model.

The results display that this method can be applied for speech synthesis with good quality of nasal- and non-nasal speech sounds resulting in a considerable improvement as compared to the results of usual all-pole techniques.

ACKNOWLEDGEMENT

M. Liu is grateful to Katholischer Akademischer Ausländer-Dienst (Germany) for a scholarship.

References

[1] KELLY, J. and LOCHBAUM, C.: Speech Synthesis. Proc. 4. Int. Congr. on Acoustics, Copenhagen 1962, Paper G 42, pp. 1-4.

[2] FRANK, W. and LACROIX, A.: Multi Tube Models for Speech Synthesis, Proc. 3. European Signal Processing Conference 1986, pp. 373-376.

[3] LACROIX, A.: Digitale Filter, 4. Auflage, Oldenbourg Verlag, München, 1996, pp. 104-110.

[4] LACROIX, A.: Ein neues Verfahren für die inverse Filterung mit dem Kreuzglied-Kettenfiler. NTZ Archiv 7, pp. 155-158, 1985.

[5] HAYASHI, S., SUGUIMOTO, M. and KISHI, G.: Low bit-rate MPC coders with adaptively controlled synthesis filter parameters. Signal Processing. Vol.35, No.3, pp. 285-293, 1994.

[6] HAWKINS, S.and STEVENS, K. N. : Acoustic and perceptual correlates of the non-nasal-nasal distinction for vowels, J. Acoust. Soc. Am 77, pp. 1560-1575, 1985.