GENERATION OF NASALIZED SPEECH SOUNDS BASED ON BRANCHED TUBE MODELS OBTAINED FROM SEPARATE MOUTH AND NOSE OUTPUTS

K. Schnell, A. Lacroix

Institute of Applied Physics, Goethe-University Frankfurt/Main
Robert Mayer-Str. 2-4, D-60325 Frankfurt am Main, Germany
{Schnell, Lacroix}@iap.uni-frankfurt.de

ABSTRACT

A new approach is discussed for analysis and generation of nasalized speech sounds considering separate mouth and nose outputs. For that purpose mouth and nose signals are separated by an insulating panel encircling the head of the speaker. The separated signals are analyzed by a branched tube model each. The model parameters are estimated from mouth and nose signal by iterative inverse filtering. The resulting areas are used for artificial nasalization of speech signals of nonnasalized vowels. Therefor the nonnasalized vowel is analyzed by an unbranched tube model which is extended by the results from the previous analysis of mouth and nose signal. For the artificial nasalization the enlarged model consists of a superposition of two branched tube models representing mouth and nose signal which are excited by residual signals. Depending on the degree of mixture of mouth and nose signals the effect of nasalization is well perceivable.

1. INTRODUCTION

The nasalization can be investigated by simulations of branched tube models which obtained from morphological data of the nasal tract [1, 2]. In [3, 4] a branched tube model is used for articulatory speech synthesis while in [3] the parameters of the nasal tract are estimated from the nasal /N/ (in SAMPA-notation). For an adequate generation of nasalized vowels by a branched tube the mouth and nose signal should be modeled by two system outputs respectively. A parameter estimation from speech signals of nasalized vowels is difficult since two model outputs are estimated from one speech signal representing a mixture of mouth and nose signal. However, in the case of nasalization the analysis of the separated signals from lips and nostrils is advantageous. In our investigation the two signals are separated acoustically so that mouth and nose signals can be analyzed. If mouth and nose signals are described by a single branched tube the parameter estimation is complicated since two outputs of the system have to be modeled simultaneously [5]. This is caused by the fact that a change of one model parameter can improve the approximation of one output whereas the approximation of the other output is degraded. To obtain a better spectral fit of mouth and nose signal the two signals are described by a tube model in each case, utilized in this contribution for analysis and synthesis.

2. SEPARATION OF MOUTH AND NOSE SIGNALS

Mouth and nose signals are recorded separately for the analysis of nasalized vowels. To avoid that mouth and nose signal are mixed the head of the speaker is integrated in an insulating panel shown in fig. 1. The insulating panel divides a room into two parts. Therefore nose and mouth signal radiate into different rooms and can be recorded by two microphones. Especially the velum position in the neighborhood of nasals is interesting which can be seen in the following example [6]. In fig. 2 at the top mouth and nose signals are depicted for the German utterance “Majonaise“ respectively [majonE:s@] in SAMPA-notation. Additionally the power P_N of the nose signal and the

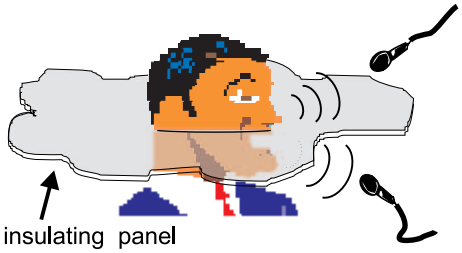

Figure 1: Separation of mouth and nose signal by an insulating panel.

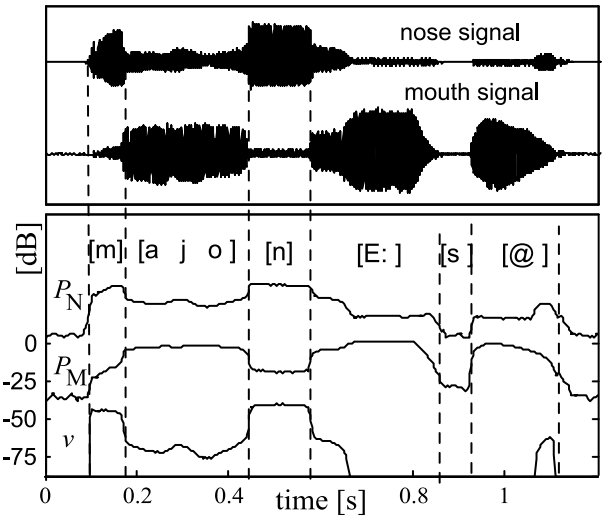

Figure 2: Analysis of separated mouth signal and nose signal of the utterance [majonE:s@]: Time signals (top), power of nose and mouth signal, and measure v of the nasalization (bottom).

power P_M of the mouth signal are shown. A measure v for the nasalization can be defined by

$$v = 10 \cdot \log\left(\frac{P_N}{P_M}\right)$$

which is shown in fig. 2 (bottom). It can be seen that the sounds between the nasals /m/ and /n/ are nasalized permanently whereas the vowel /E:/ behind the nasal /n/ is nasalized partly in time only. During the vowel /E:/ the velum is closing which is marked by an arrow in fig. 2. This example demonstrates the significance of investigations of nasalization in speech.

3. TUBE MODEL AND ANALYSIS

For the modeling of mouth and nose signals wave digital filters are used which describe the propagation of plane waves through nasal and vocal tract. Branched tube models are chosen since the nasal tract is coupled to the vocal tract in the case of nasalization. The branched tube is depicted in figure 3. The branching of the tube is realized by a parallel three port adaptor (center fig. 3) with the parameters ρ_1 and ρ_2. These parameters are functions of the three adjacent areas A_i at the tube branch by $\rho_j = 2A_j/(A_1 + A_2 + A_3)$. The sidebranch is coupled by the three port and can be described by the transfer

$$\tilde{H}(z) = \frac{X_S^b}{X_S^f} = \frac{Q(z)}{P(z)} = \frac{\sum_i q_i \cdot z^{-1}}{\sum_i p_i \cdot z^{-1}}$$

function relating the signals x_S^f and x_S^b which are located between the three port adaptor and the sidebranch shown in fig 3. $\tilde{H}(z)$ can be described by the numerator Q and the denominator P determined by the reflection coefficients of the side branch. The transfer function can be calculated by 2×2 scattering transfer matrices $\boldsymbol{T}$ which give a relation between the wave quantities at left port and right port. $\boldsymbol{T}_i$ describes the

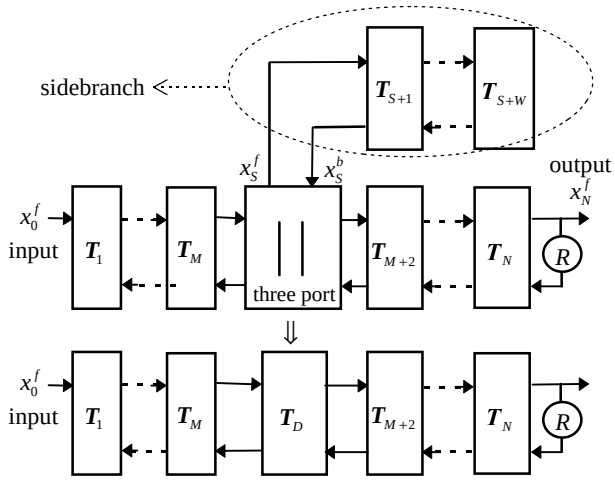

Figure 3: Branched Tube Model.

discontinuity of the cross-sectional area and the following uniform tube element realized by a delay element:

$$\boldsymbol{T}_i = \begin{pmatrix} 1 & r_i z^{-1} \\ r_i & z^{-1} \end{pmatrix} \quad ; \quad \begin{pmatrix} x_i^f \\ x_i^b \end{pmatrix} = \boldsymbol{T}_i \begin{pmatrix} x_{i+1}^f \\ x_{i+1}^b \end{pmatrix}; \qquad (1)$$

x_i^f and x_i^b are the wave quantities propagating forward and backward in the tube. r_i is the reflection coefficient in the ith tube. The three port adaptor with a tube element and the coupled side branch can be transformed into the 2×2 scattering transfer matrix

$$\boldsymbol{T}_D = \frac{1}{\rho_2(Q+P)} \cdot \begin{pmatrix} \langle \rho_1 + \rho_2 - 1, 1 \rangle & \langle \rho_2 - 1, 1 - \rho_1 \rangle z^{-1} \\ \langle 1 - \rho_1, \rho_2 - 1 \rangle & \langle 1, \rho_1 + \rho_2 - 1 \rangle z^{-1} \end{pmatrix} \qquad (2)$$

shown in fig. 3 with the abbreviation $\langle x, y \rangle := x \cdot Q + y \cdot P$. $\boldsymbol{T}_D$ can be split into a matrix $\boldsymbol{T}_D'$ with numerators as elements and a common denominator B:

$$\boldsymbol{T}_D = \frac{1}{B(z)} \cdot \boldsymbol{T}_D', \qquad B(z) = \rho_2(Q+P).$$

The termination at the tube model output x_N^f is realized by a real coefficient $R = -0.95$ describing an open tube end with additional losses. The transfer function of the entire tube system H is given by

$$H(z) = \frac{X_N^f}{X_0^f} = \frac{B(z)}{A(z)} = \frac{\rho_2(Q+P)}{A(z)}.$$

The denominator A depends on all parameters of the tube model. The numerator B of $H(z)$ is the common denominator of $\boldsymbol{T}_D$ and depends on the parameters of the sidebranch.

3.1. Parameter Estimation

The parameter estimation is performed by iterative inverse filtering and is divided into two steps. At first the coefficients of the side branch representing the zeros of $H(z)$ are estimated and then the remaining parameters. The estimated zeros obtained by a general ARMA algorithm are converted into the reflection coefficients of the side branch representing B whereas the estimated poles are ignored further on. The ARMA estimation is explained in [7]. By the estimated zeros the recursive part B of the inverse filter can be carried out first by $x' = \mathrm{IDFT}(X \cdot B^{-1})$; x represents the analyzed signal. The remaining parameters are estimated by minimizing the power of the output x_0^f of the nonrecursive part of the inverse filter depicted in fig. 4. The criterion for the minimum is:

$$\mathrm{E}\left[\left(x_0^f\right)^2\right] \to \min \quad \Rightarrow \quad \frac{\partial \mathrm{E}\left[\left(x_0^f\right)^2\right]}{\partial r_i} = 0\,, \quad \frac{\partial \mathrm{E}\left[\left(x_0^f\right)^2\right]}{\partial \rho_i} = 0\,. \quad (3)$$

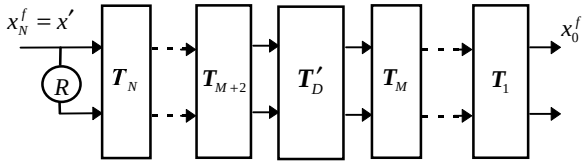

Figure 4: Flow graph of nonrecursive part of inverse filtering.

Formulas can be derived from the equations in (3) which yield one optimal coefficient fulfilling the criterion on condition that the other coefficients are known. Therefore the coefficients are estimated iteratively by these formulas yielding good results. One iteration consists of a calculation of all parameters. In [8] the algorithm is explained in detail.

4. ANALYSIS OF NASALIZED VOWELS

The sampling rate for the nasalized vowels is 16 kHz. To separate the effects of excitation and radiation from the signals mouth and nose signals are prefiltered by an adaptive preemphasis. Mouth and nose signals of a male speaker are analyzed each by the inverse filtering process of a branched tube model. The number of tube sections corresponds to the actual tract lengths. The length of the nasal tract is 12 tubes and the length of the pharynx is chosen as 10 tubes. The number of the tube sections for the mouth cavity is 8. In the case of the analyzed mouth signal the side branch represents the nasal tract with 12 tubes whereas in the case of the nose signal the side branch represents the mouth cavity with a length of 8 tubes since mouth cavity and the nasal tract are exchanged. The results of the analysis of separated mouth and nose signal of the nasalized vowel / ã / are shown in fig. 5. The estimated magnitude responses of the branched tubes and the contribution of the estimated zeros are depicted. The zeros are caused by the side branch. The formants of the vowel can be recognized in the spectra of the mouth signal and the nasal formant can be observed in the spectra of the nose signals. The estimated zeros

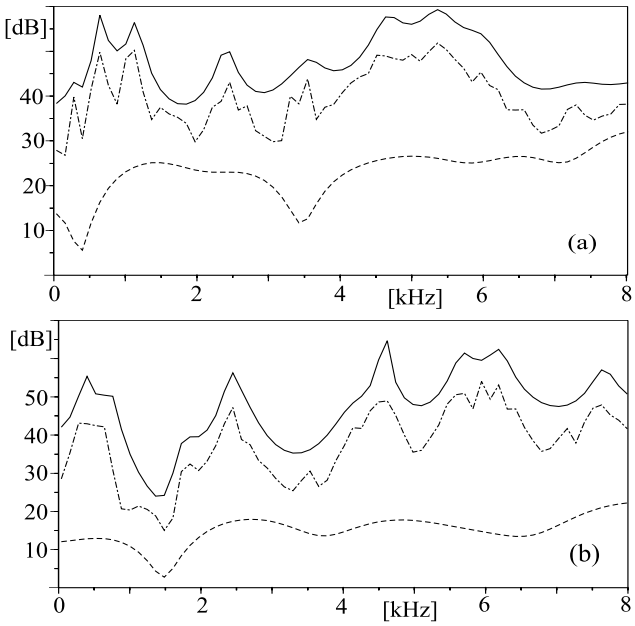

Figure 5: Analysis of mouth signals (a) and nose signals (b) of the nasalized vowel / ã / with preemphasis: Estimated magnitude response after 50 iterations (top solid line), DFT of analyzed speech signal with preemphasis (center dashed-dotted line), estimated zeros of the model (bottom dashed line).

of mouth signals vary from vowel to vowel. In fig. 6 (left) the estimated areas of the nasal tract are depicted obtained from the nose signal of / ã / which represent the areas between the three port and the tube model output. The areas of the side branch estimated from the mouth signal of / ã / are shown in fig. 6 (right) which represent the nasal tract, too. The estimated nasal tract areas in fig. 6 are comparable which cannot be generalized for all vowels. It should be noted that the actual nasal tract is a complicated bipartite tube system with coupled sinus cavities which are connected to the nasal passage by thin channels.

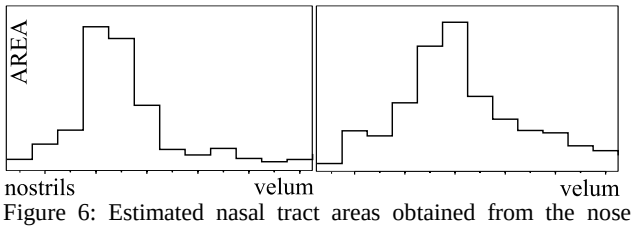

Figure 6: Estimated nasal tract areas obtained from the nose signal (left) and from the mouth signal (right) of the nasalized vowel / ã /.

5. GENERATION OF NASALIZED VOWELS

The estimated areas of the model are utilized to nasalize a speech signal of a nonnasalized vowel. In the following example the vowel /a/ is treated. At first the speech signal of the nonnasalized vowel is analyzed by an unbranched tube model with the tube termination $R = -0.95$. The estimated vocal tract areas A' of the nonnasalized vowel /a/ are shown in fig. 7 and represent a model for the mouth signal without nasal branch. Now this unbranched tube model is extended by the estimated areas of the

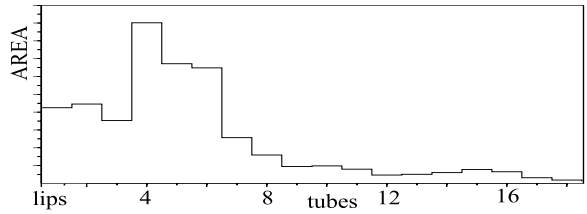

Figure 7: Estimated vocal tract areas A' obtained from a nonnasalized vowel /a/ for an unbranched tube model.

vowel / ã / for artificial nasalization which is depicted schematically in figure 8. As mentioned in the introduction the estimation of two outputs simultaneously for one model is complicated [5]. Therefore two branched tubes are used for the representation of the mouth and nose signal respectively which is advantageous for better estimation results yielding improved synthesized speech quality. The procedure is illustrated in the following. The vocal tract of /a/ from fig. 7 is coupled by the side branch with the nasal tract areas of fig. 6 (right) yielding a branched tube model for the mouth signal y_M which also models the effect of the lowered velum. The modification of the magnitude response caused by the coupled side branch is depicted in fig. 9. Due to the coupling additional zeros appear in the frequency response; in this case primarily the first resonance

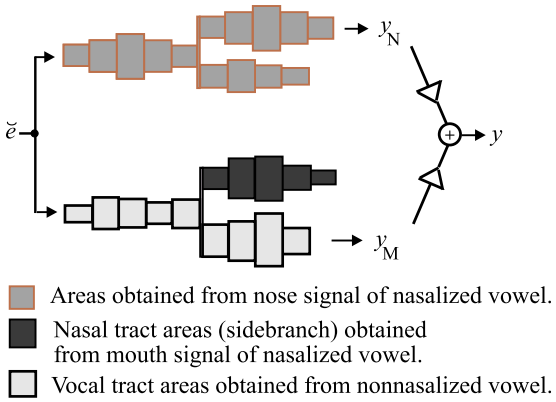

Figure 8: Schematic generation of nasalized vowels.

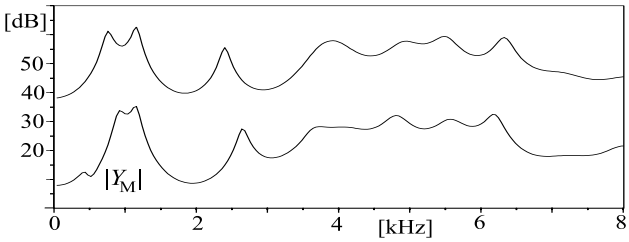

Figure 9: Magnitude response of the unbranched tube with the vocal tract areas A' of the nonnasalized vowel /a/ (top) and with the coupled side branch additionally (bottom).

is affected. Since the nasalized vowel is a superposition of two signals besides y_M an additional nose signal should be generated, too. The nose signal y_N is generated by the branched

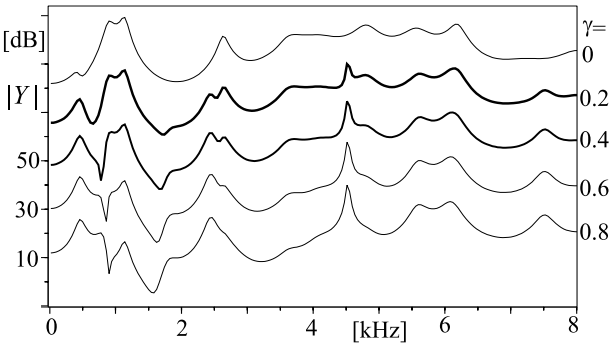

Figure 10: Magnitude responses of two branched tube systems depending of mixture of the outputs of lips and nostrils.

tube system of the analyzed nose signal of / ã / with the magnitude response in figure 5 (b). The system of the artificial nasalized vowel consists of two branched tube models representing mouth and nose signal which are shown in fig. 8. Resulting magnitude responses

$$Y(\omega) = (1-\gamma)Y_M + \gamma Y_N, \quad 0<\gamma<1$$

of the two added systems are shown in fig. 10 depending on the mixture γ of the outputs y_N and y_M. The top curve represents the mouth signal y_M with $\gamma = 0$. The magnitude responses below represent a nasal tract coupling with values $\gamma > 0$. The second and third curve from top (thick line) are realistic mixtures. Between 500 Hz and 1 kHz a zero appears as consequence of the superposition damping the first formant. This effect can be seen, too, by a theoretical calculation of the transfer function of the nasalized vowel / ã / in [9]. For the synthesis of the artificial nasalized vowel the residual signal e from the nonnasalized vowel is processed which has an almost flat spectral envelope. e is prefiltered by real poles of the adaptive preemphasis resulting in $\breve{e}$. The two branched tube models with the entire magnitude responses of fig. 10 are excited each by the signal $\breve{e}$ so that the mouth and nose signals are generated. Audio examples of the mixed synthesized signals demonstrate that the effect of the nasalization is well perceivable. The degree of the nasalization depends on the mixture of the two signals. The nasalization of other vowels have been processed in the same way. For comparison of the resulting magnitude responses fig. 11 shows the DFT of a nasalized vowel / ã / of the same speaker by conventional speech recording. The main zero about 600 Hz in fig. 11 corresponds to the zero which is obtained by the mixture of the outputs of the two branched tubes in figure 10.

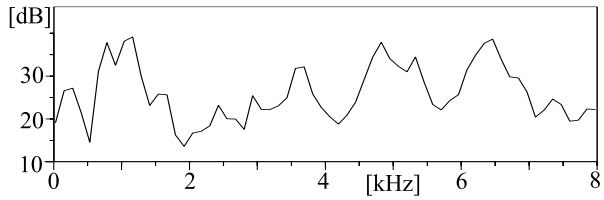

Figure 11: DFT of nasalized vowel / ã / with preemphasis.

6. CONCLUSIONS

The use of two branched tube models is advantageous for the analysis and generation of mouth and nose signals of nasalized vowels yielding good results. For the investigation of the nasalization of vowels mouth and nose signals are recorded separately. Both signals are analyzed individually by iterative inverse filtering of a branched tube model. The resulting model parameters are used for an additional nasalization of a nonnasalized vowel. For the synthesis of the artificial nasalized vowel two branched tube models are excited by the residual signal of the nonnasalized vowel and the outputs of the two systems are mixed representing the mouth and nose signal. The nasalization of the synthesized vowels is well perceivable depending on the degree of mixture of the mouth and nose signals. The magnitude response of the added two tube models representing the output of lips and nostrils shows a zero corresponding to the estimated zero of natural speech signals of the same speaker.

REFERENCES

[1] Maeda, S., "The Role of Sinus Cavities in the Production of Nasal Vowels", *Proc. Int. Conf. ICASSP'82,* Paris France, pp. 911-914, 1982.

[2] Dang, J., Honda K., Suzuki H., "Morphological and Acoustical Analysis of the Nasal and the Paranasal Cavities", *J.Acoust.Soc.Am.*, Vol. 96, No. 4, pp. 2088-2100, 1994.

[3] Meyer P., Wilhelms R., Strube H.W., "A quasiarticulatory speech synthesizer for German language running in realtime" *J.Acoust.Soc.Am.*, Vol. 86, pp. 523-539, 1989.

[4] Sondhi M., Schroeter J., "A hybrid time-frequency domain articulatory speech synthesizer", *IEEE Trans. on Acoustics, Speech and Signal Processing*, ASSP-35, pp. 1070-1075, 1987.

[5] Schnell, K., Lacroix, A., "Parameter Estimation from Speech Signals for Tube Models", *Joint ASA/EAA Meeting - Forum Acusticum*, Berlin, CD-ROM, 1999.

[6] Bettinelli, K., Schnell, K. Lacroix, A., "Separate Messung und Analyse von Mund- und Nasensignalen bei natürlicher Sprache", *Proc. 13th Conf. ESSV-2002*, Dresden Germany, pp. 237-244, 2002.

[7] Schnell, K., Lacroix, A., "Pole Zero Estimation from Speech Signals by an Iterative Procedure", *Proc. Int. Conf. ICASSP-2001*, Salt Lake City USA, pp. 109-112, 2001.

[8] Schnell, K., Lacroix, A., "Parameter Estimation of Branched Tube Models by Iterative Inverse Filtering", *Proc. 14th Int. Conf. on Digital Signal Processing, DSP-2002*, Santorini Greece, Vol. I, pp. 333-336, 2002.

[9] Stevens, K. N., *Acoustic Phonetics*, MIT Press, Cambridge London, 1998.