Joint Analysis of Speech Frames for Synthesis Based on Lossy Tube Models

Karl Schnell and Arild Lacroix

Institute of Applied Physics, Goethe-University Frankfurt
Max-von-Laue-Str. 1, D-60438 Frankfurt am Main, Germany
schnell@uni-frankfurt.de

Abstract

This paper discusses a model-based synthesis approach focused on the estimation of model parameters. For the treated approach, tube models are used for analysis and synthesis of speech units. In comparison to the standard lossless tube model, an extended tube model is used which includes the frequency dependent vocal tract losses. The parameters of the tube models are estimated by minimizing the spectral error between the tube model and a speech segment. For the analysis of speech units, the time evolution of the parameters is taken into account. For that purpose, the speech segments are analyzed jointly which ensures smooth parameter trajectories. The investigations show that, especially for extended tube models, the joint analysis of frames improves the quality of the synthesized speech signals. Additionally, the differences of the results obtained by the standard and the extended tube model are discussed.

1. Introduction

Speech generation is nowadays often performed by concatenation of speech units. The speech units to be concatenated can be represented by the speech signals themselves [1] or by model-based descriptions. The model-based description has the advantage of flexibility and possible data reduction with the disadvantage of decreasing more or less the speech quality in comparison to natural speech signals. For model-based synthesis or re-synthesis, the speech signals are usually generated by a model describing the vocal tract and/or nasal tract ranging from the standard LPC-model to articulatory models, e.g. [2-4]. For synthesis, the common task of these models is to shape the spectral envelope of the synthesized speech. It is known that, for linear prediction, the harmonic structure of the spectrum of voiced speech influences the estimation of formant frequencies and bandwidths, especially for high pitch. Underestimating of bandwidth by linear prediction decreases the synthesis quality, which can be compensated by a subsequent bandwidth expansion or by specialized analysis methods [5-6]. In contrast, in [7] an analysis is proposed considering multiple measurements. Besides spectral approximation, an important feature of the models is the type of model parameters. The interpretation of the parameters varies from parameters describing the spectral envelope such as MFCCs, LSF, or formants to parameters describing the geometry of the vocal tract. Tube models describe the geometry of the vocal tract by tube areas. Articulatory models are mostly based on tube models with articulatory parameters such as the center or tip position of the tongue. Different articulatory vocal tract models exist mapping articulatory parameters to vocal tract areas or to mid-sagittal cross-sections, which restricts the scope to feasible vocal tract configurations. One problem is to control the articulatory parameters for synthesis. For a data-driven approach the parameters are estimated from speech signals, which is not an easy task [8, 9]. The fact that the articulatory vocal tract models are more or less imperfect affects the estimation. A more practical obstacle is that model adaptation to an individual speaker needs measurements of the speaker's anatomy [9], and another more general problem of estimation is the non-uniqueness of acoustic-to-articulatory mappings, which has several reasons. One reason for ambiguity can lie in the type of spectral features. For example, if only the first formants are taken into account, not the whole spectral information is used for the estimation. To tackle the problem of non-uniqueness, look-up tables, obtained from acoustic and articulatory measurements, combined with dynamic constraints can be used [10]. The benefit of articulatory parameters is their meaningful interpretation; however, their drawbacks for data-driven synthesis are the difficulties of the parameter estimation and the restriction of the area function, which can be unfavorable for a precise spectral approximation. In this contribution, a lossy tube model is used whose parameters are estimated from the magnitude spectrum of a speech segment. In comparison to the standard lattice filter, the lossy tube model implies the frequency dependent losses of the vocal tract. The losses which are introduced influence spectral estimation, especially on the formant bandwidths, and, additionally, on the vocal tract areas. The areas of the model are unconstrained enabling detailed spectral modeling. In comparison to the investigations in [11-12], the main focus of this contribution is the discussion of a joint parameter estimation of speech segments implying dynamic constraints and the realization of the acoustic synthesis.

2. Extended Tube Model

Tube models can be described in the time domain or frequency domain. The advantages of the frequency-domain description are the direct realization of frequency dependent vocal tract losses and variable tube lengths; however, one drawback is that for the realization of the acoustic synthesis a conversion from the frequency domain to the time domain has to be performed, which is usually realized via the calculation of the impulse response [13]. In comparison to frequency-domain models, time-domain tube models enable a direct realization of the acoustic synthesis. Here, time-domain tube models are treated for synthesis.

The simplest tube model is the LPC-model in lattice structure, which describes a lossless tube model. The standard lossless tube model consists of tube elements realized by lossless delays z^{-1} and adaptors describing the area discontinuities; the tube termination at the lips is realized by a reflection

coefficient ± 1. In this contribution, a lossy tube model is used, which considers losses within the vocal tract and at the lips by radiation. The frequency dependent loss effects by vibrating walls, viscous friction, and heat conduction within the vocal tract are modeled by lossy tube elements, which includes a lossy delay $\vartheta(z)$ instead of the lossless delay z^{-1}. ϑ is realized by a pole-zero system

$$\vartheta(z) = \frac{0.9875 \cdot z^{-1} - 0.9047 \cdot z^{-2}}{1 - 0.9182 \cdot z^{-1} + 0.0041 \cdot z^{-2}} \quad . \tag{1}$$

The coefficients of ϑ are obtained by an optimization with respect to the mentioned loss effects [11]; here, for a sampling rate of 16 kHz. The lossy delays are placed alternately in the upper and lower path of the signal flow graph of the lossy tube model depicted in Fig. 1. The reflection coefficients $r(i)$ can be transformed into the areas by $a(i+1) = a(i) \cdot (1 - r(i))/(1 + r(i))$. For synthesis, power waves

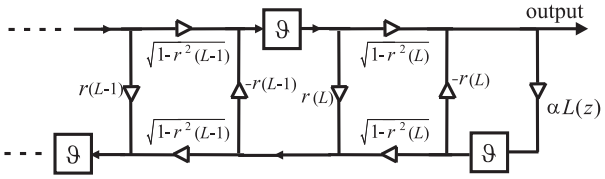

Figure 1: Flow graph of the lossy tube model for synthesis.

are chosen as wave quantities determining the adaptors in Fig. 1. The advantage of power waves is that alterations of the coefficients don't change the wave energy. The termination $\alpha \cdot L(z)$ at the lips is realized frequency dependent by the pole-zero lip-impedance model from Laine [14] with the lip opening area as parameter; α is an additional real damping factor. The termination of the tube model at the other end is reflection free since a fixed termination at the glottis has the disadvantage that the vocal tract length has to be estimated. The tube model consists of $L = 24$ tube elements whose area configuration is described by the vector $\boldsymbol{r} = (r(1), r(2), \ldots r(L))^{\mathrm{T}}$ of reflection coefficients. Since for a sampling rate of 16 kHz the vocal tract length is smaller than 24 tubes, the first reflection coefficients can model the constriction of the glottis.

2.1. Parameter estimation of the lossy tube model

For the analysis of speech units, the model parameters are determined from the corresponding speech signals of the units. For that purpose, the units are segmented into overlapping frames s_k, which are multiplied by a Hanning window. For each frame, the parameters to be estimated are the reflection coefficients of the lossy tube model, whereas the parameters of the lossy delays and the lip termination are constant for the analysis due to ambiguity; the lip opening is chosen to 2.5 cm². To eliminate the influence of excitation and radiation on the spectral envelope, the speech segments s_k are filtered by a repeated adaptive pre-emphasis which is realized by inverse filtering with linear prediction of first order. The resulting pre-emphasis filter *P* consists of two real zeros

$$P(z) = \prod_{i=1}^{2} \left(1 - p(i) \cdot z^{-1}\right), \tag{2}$$

which can balance better the spectral decrease of voiced speech than only one zero. Each segment *k* has its individual estimated pre-emphasis coefficients. After pre-emphasis, the reflection coefficients of the lossy tube model are estimated from the pre-emphasized speech segments s'_k by minimizing the error

$$e_k(S'_k, H_k) = \frac{1}{\pi} \int_0^{\pi} \left| \frac{S'_k(\omega)}{H_k(\mathrm{e}^{\mathrm{j}\omega})} \right|^2 \mathrm{d}\omega \,, \tag{3}$$

which describes a spectral distance between the magnitude response $|H_k|$ of the tube model and the spectrum $|S'_k|$ corresponding to frame *k*. The transfer function of the tube model in Eq. (3) is calculated with adaptors which are equal to those used for the standard lattice filter. This is necessary for the error definition since the adaptors for power waves introduce a factor of the transfer function which is unfavorable for the estimation. The error definition (3) represents an inverse filtering approach in the frequency domain. Since the segments s'_k are finite signals, the integral in Eq. (3) can be represented by a sum with discrete frequencies. The error *e* is minimized by a gradient-based optimization algorithm. The gradient is approximated by error differences of small variations of individual reflection coefficients. Since the transfer function $H_k(\boldsymbol{r}_k, \mathrm{e}^{\mathrm{j}\omega})$ is a function of the parameter vector $\boldsymbol{r}_k$ of the *k*-th frame, the spectral error $e_k(\boldsymbol{r}_k)$ is a function of the reflection coefficients, too. The approximation of gradient is defined by

$$\nabla e_k = (\nabla_1 e_k, \nabla_2 e_k, \ldots)^{\mathrm{T}} \quad \text{with}$$
$$\nabla_i e_k = e_k(\ldots r(i-1), r(i) + \varepsilon, \ldots)^{\mathrm{T}} - e_k(\ldots r(i-1), r(i), \ldots)^{\mathrm{T}};$$

ε is a small constant about 10^{-8}. One iteration of the gradient algorithm is defined by a step in the direction of the negative normalized gradient with the adaptive step size δ:

$$\boldsymbol{r}_k^{j+1} = \boldsymbol{r}_k^{j} - \delta \cdot \nabla e_k / \max(|\nabla e_k|) \,; \tag{4}$$

the superscript with *j* indicates the iteration number and the function max() yields the maximum value. The step size $\delta = \sum_{l=1}^{7} c_l \cdot d_l$ is a parameterized function with the variables $c_l \in \mathbb{N}$ which are determined by

$$\arg\min_{c_l} e_k(\boldsymbol{r}_k^{j} - \delta \cdot \nabla e_k / \max(|\nabla e_k|)) \tag{5}$$

with the constraints $\delta = \sum_{l=1}^{7} c_l \cdot d_l$ and $|r_k(i)| \leq 0.99$. For the minimization of Eq. (5), the step size δ is, firstly, increased repeatedly by d_1 until the error is equal or greater in comparison to the previous error value or if $|r_k(i)| > 0.99$ is valid. Then, the next smaller d_l is used for increasing δ to minimize the error. The iteration is finished if the smallest d_l is reached. Here, the values of d_l are defined by $d_1 = 0.05$ and $d_{l+1} = d_l / 5$ for $l = 2 \ldots 7$.

2.1.1. Joint analysis of frames

To ensure a smooth trajectory of parameter vectors, the frames are analyzed jointly. The joint analysis is realized by an exchange of interim results between adjacent frames during optimization. For that purpose, the parameter vectors of an individual iteration are averaged with the vectors of adjacent frames. For example, if the *j*-th iteration yields the vectors $\boldsymbol{r}_k^{j}$, then these vectors are updated by a mean of vectors including

those of the neighboring frames. This averaging can be performed in different parameter descriptions. For that purpose, the vectors are transformed into the desired description $\boldsymbol{\psi}_k^j$, then the averaging is performed, and, finally, the averaged parameter vectors $\tilde{\boldsymbol{\psi}}_k^j$ are transformed back into reflection coefficients:

$$\begin{aligned} \boldsymbol{r}_k^j &\rightarrow \boldsymbol{\psi}_k^j \\ \tilde{\boldsymbol{\psi}}_k^j &= a_0 \cdot \boldsymbol{\psi}_k^j + \sum_{i=1}^{W} a_i \cdot (\boldsymbol{\psi}_{k-i}^j + \boldsymbol{\psi}_{k+i}^j) \\ \tilde{\boldsymbol{\psi}}_k^j &\rightarrow \tilde{\boldsymbol{r}}_k^j . \end{aligned} \tag{6}$$

The updated vectors $\tilde{\boldsymbol{r}}_k^j$ are used for the starting vectors $\boldsymbol{r}_k^{j+1}$ of the next iteration. The use of the averaging (6) imposes dynamic constraints and helps prevent divergence between neighboring frames during parameter optimization. The averaging (6) is performed in prescribed iterations with the numbers $j \in J$; J is a set of iteration numbers. The averaging can be performed every n-th iteration denoted by $J_n = \{n, 2n, 3n, ...\}$. The iterations without averaging allow that the vectors of the frames can evolve apart from each other a little bit. An irregular arrangement of the numbers in J can be suitable allowing a more unconstrained parameter evolution for the first iterations. For that purpose, the set $J_{ir} = \{40, 50, 52, 54, ..., 68, 70\}$ is used.

2.1.2. Independent analyses of frames

In comparison to the joint analysis, the frames can be analyzed by minimizing the errors $e_k(\boldsymbol{r})$ for each frame independently. Since there are no iterations with averaging, this independent analysis is denoted by the null set $J_0 = \{\}$.

2.1.3. Analysis of speech units

Both the joint and the independent analysis terminate after a prescribed number of iterations. In the following sections, analysis results are shown with a total iteration number of 70. The averaging by Eq. (6) is performed in the description of logarithmic areas. The values for the averaging in Eq. (6) are $W = 1$, $a_0 = 3/7$, and $a_1 = 2/7$ describing a weighted mean of adjoining frames, which emphasizes the middle frame.

3. Analysis of Diphones

The estimation procedure in the preceding section is used for speech analysis and synthesis. The sampling rate of the speech signals is 16 kHz. For re-synthesis, words are analyzed, whereas diphones are analyzed for synthesis. The diphones are from the diphone database de1 [15] for German from a female speaker. The analysis of the diphones yields the corresponding parameter vectors representing the diphones. To demonstrate the effect of the losses by the lossy tube model, also analysis and synthesis is performed with the lossless tube model represented by the standard lattice filter with power waves. In this case, the parameter estimation is performed as described in the preceding section, however, using the transfer function of the lossless model for the error definition of Eq. (3). The lossy model can be converted into the lossless tube model by the substitutions $\vartheta(z) := z^{-1}$ and $\alpha \cdot L(z) := -1$. If the estimation is performed with the lossless tube model by the independent analysis without averaging, the estimation results are comparable to those of the common linear prediction approach. For the analysis, the diphones are segmented into overlapping frames with the length of 625 samples and an overlap of 125. Figures 2-5 show the estimated logarithmic areas and the corresponding magnitude responses $|H_k|$ of each analyzed frame s_k' of diphones. Fig. 2 shows the results from the independent analysis of the frames without averaging using J_0 . It can be seen that the estimated areas and magnitude responses fluctuate from frame to frame, especially in the case of the lossy tube model. These discontinuities of the model parameters decrease usually the quality of the synthesized speech and can be reduced by averaging during the optimization, which can be seen from Figs. 3 and 4. The iteration set J_2 is used for the results of Fig. 3, whereas the iteration set J_{ir} is used for the results of Fig. 4. The differences between the results using J_2 or J_{ir} are relatively small for the lossy tube model and almost negligible for the lossless tube model. The effect of J_{ir} is a slightly stronger emphasis on the temporal details in comparison to J_2 . Here, a compromise should be made between smoothness and detailed approximation. It should be noted that temporal details can be caused by different effects: on the one hand, resonance movements by articulation which should be preserved and, on the other hand, by fluctuations of the excitation and by block-wise processing which should be ignored. Besides different uses of the averaging, the results obtained by the lossy and the lossless tube model show generally some differences in the estimated areas and magnitude responses. The areas estimated by the lossy tube model are more prominent in comparison to those of the lossless model; additionally, the shapes of the area configurations differ between the lossy and the lossless case. For example, for the fricative /v/ of the diphone /a-v/ in Fig. 5, the corresponding areas near the lips show more an open mouth for the lossless model, whereas the areas of the lossy model show more a closed mouth, which is more realistic. In general for the majority of the voiced sounds of the diphone database, the estimated logarithmic areas show reasonable vocal tract cavities. For example, from the figures 2-5 it can be seen that the estimated areas of the vowel /a/ shape a large front cavity ranging to the lips. For the sound /j/, the back cavity can be recognized. Due to the fact that the transfer function is more sensitive to the relationship of areas than to the absolute areas themselves, the logarithmic areas can be estimated more reasonably than the absolute areas. The assessment of the areas can be performed by regarding vocal tract areas from literature obtained from x-ray or NMR; however, this comparison can be used only for a rough assessment since these vocal tract shapes are from other subjects and the vocal tract configurations differ, in general, by coarticulation and by individual representations of the phonemes. An important pre-processing step for obtaining reasonable area configurations is an appropriate pre-emphasis [16]. Here, the repeated adaptive pre-emphasis seems to be suitable for that task.

In addition to the differences in terms of area functions, the magnitude responses estimated by the lossy tube model imply resonances of expanded bandwidths in comparison to the lossless model, which tends to underestimating of bandwidth. One reason for that may lie in the more realistic modeling of vocal tract acoustics by the lossy model.

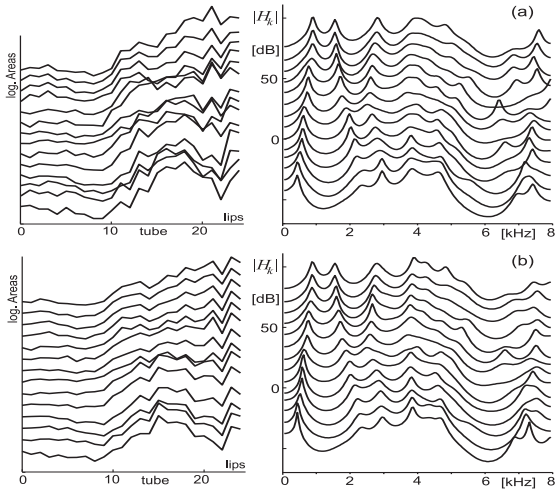

Figure 2: Estimated log. areas and magnitude responses of diphone /j-a/ by optimization without averaging using J_0, (a) for lossy tube model and (b) for lossless tube model.

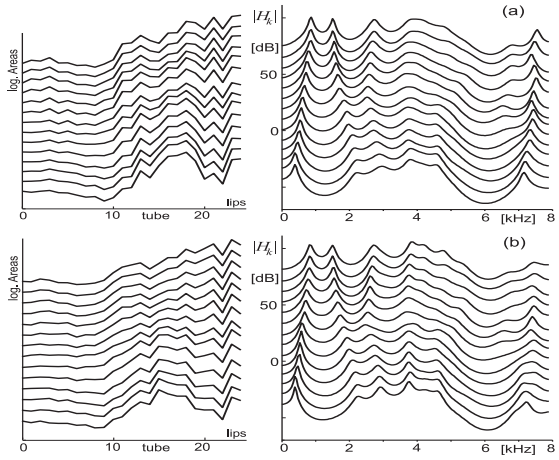

Figure 3: Estimated log. areas and magnitude responses of diphone /j-a/ by optimization with averaging using J_2, (a) for lossy tube model and (b) for lossless tube model.

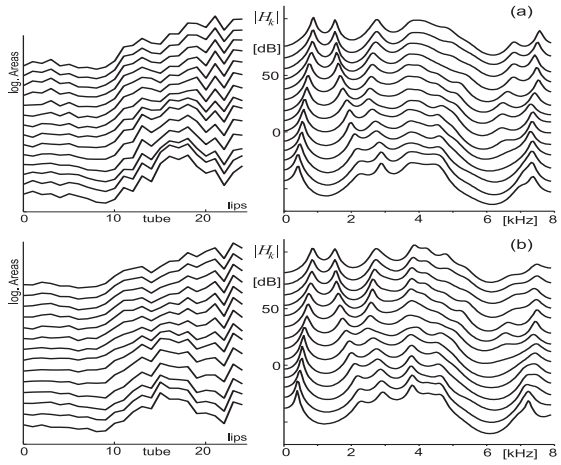

Figure 4: Estimated log. areas and magnitude responses of diphone /j-a/ by optimization with averaging using J_{ir}, (a) for lossy tube model and (b) for lossless tube model.

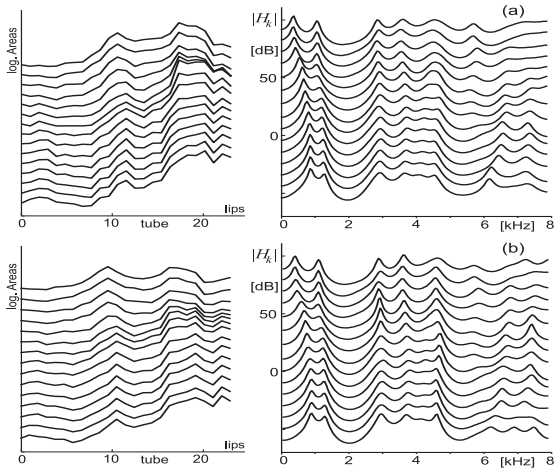

Figure 5: Estimated log. areas and magnitude responses of diphone /a-v/ by optimization with averaging using J_{ir}, (a) for lossy tube model and (b) for lossless tube model.

4. Synthesis of speech

The estimated areas and pre-emphasis coefficients obtained from the speech units are used for synthesis. For that purpose, the tube model is controlled by the parameter vectors successively. A de-emphasis filtering controlled by the pre-emphasis coefficients precedes the filtering of the tube model. To adapt the speech units to the required phoneme durations, parameter vectors can be doubled or omitted. The quality of the acoustic synthesis depends, aside from estimation of parameter vectors, also on the excitation of the tube system and on the concatenation of the model-based diphones. The diphone concatenation is performed by a parameter transition between the boundary vectors of the diphones to be concatenated, and is also treated in [12]. The excitation of the tube model is different for voiced and unvoiced sounds. For unvoiced fricatives, the excitation is relatively unproblematic and can be realized by noise. It is well known that the realization of the voiced excitation is more problematic due to its complicated structure and its impact on the speech quality; the voiced excitation has harmonic and non-harmonic components and its noisy components are non-stationary within a speech period. The use of an impulse train is the easiest way to implement a voiced excitation, however, with the disadvantage of introducing buzziness into synthesis [17]. To yield a more naturally sounding excitation, in [18] analyzed speech segments are used repeatedly for the voiced excitation. Related to that approach, here, a pitch-modified residual of an individual utterance of the schwa-sound is used for all voiced sounds, which avoids unnaturally sounding effects like the buzziness, for the most part. The pitch modification algorithm is based on a decomposition and parameterization of the residual signal in a low-pass filtered description, which is sketched in Fig. 6. The low-pass filter causes a smooth waveform related to the glottal flow. Each period of the low-passed residual *g* is decomposed into a small region *y* including the glottal closure instance and the remaining part *x*. The segments of the glottal closure instances are taken over unchanged, whereas the adaptation to the new period length is performed in the remaining parts. The remaining parts *x* are approximated by a polynomial which

models the smooth contour of the waveform. Additionally, the approximation error is considered, which contains also noisy components. The modification of the lengths of the remaining parts is performed differently for the polynomial model and for the error of approximation, namely, by interpolation for the polynomial model and by a specific OLA-technique (overlap and add) for the error of approximation. After the modification of length, the parts are composed and the resulting signal is filtered by a high-pass filter. The pitch modification algorithm is explained in detail in [19]. For the realization of the excitation during synthesis, a sufficient number of adjacent periods between 15 and 30 of the schwa-sound is used one after the other. If the last period is reached, a period of the beginning is used randomly between the first and fifth period; in this way the repetition is irregular. Since the excitation is

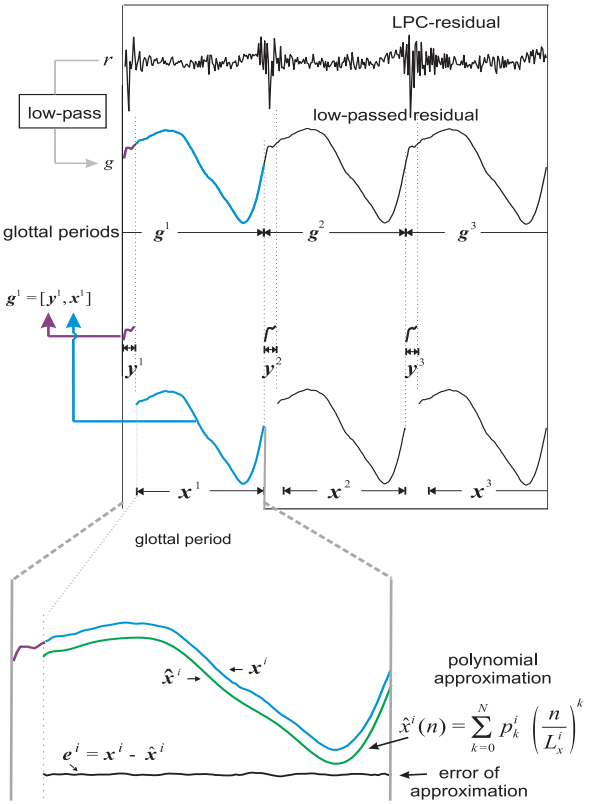

Figure 6: Sketch of the decomposition of the low-passed residual signal for pitch modification.

independent of the analyzed speech units, the acoustic synthesis needs only the estimated model parameters, which is favorable for data reduction. Fig. 7 shows spectrograms of synthesized speech signals of the German word "jawohl" [javo:l] by concatenation of the parameter vectors obtained from the diphones; the lossy tube model is used for analysis and synthesis. Figure 7(a) results from the synthesis with the excitation using the pitch-modified residual of the schwa-sound, whereas Fig. 7(b) results from the synthesis with impulse train excitation. The impulse train causes a harmonic structure in the whole frequency range, whereas the residual-based excitation reduces the harmonicity in the higher frequency range comparable to natural speech. A perceptive evaluation shows that the synthesized speech signals with the residual-based excitation sound less peaky with significantly

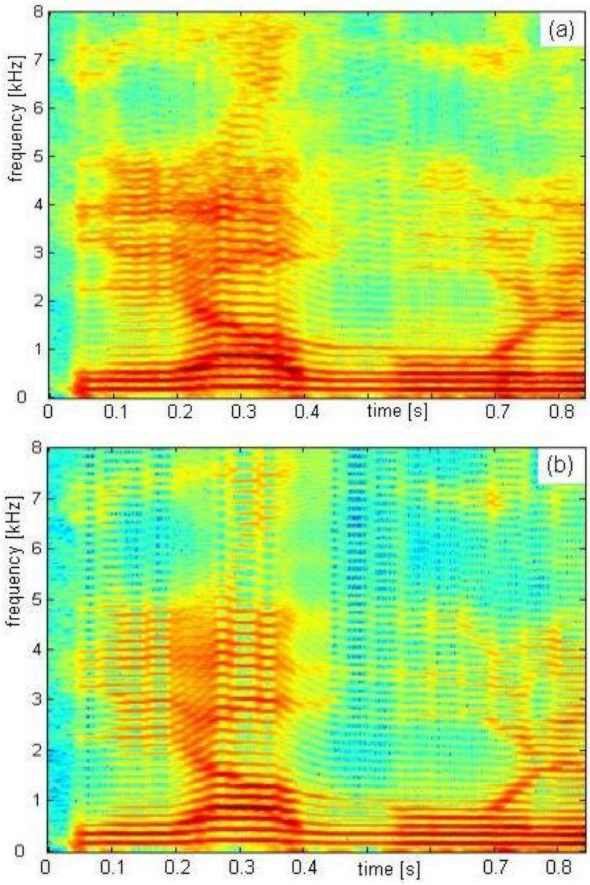

Figure 7: Spectrograms of synthesized word [javo:l] by model-based diphones with the lossy tube model: (a) with residual-based excitation obtained from the schwa-sound and (b) with impulse train excitation.

reduced buzziness in comparison to synthesis with impulse train excitation. In general, the residual-based excitation yields a more natural timbre. It should be reiterated that the residual-based excitation is independent from the analyzed speech units. Figure 8 shows a segment of the synthesized

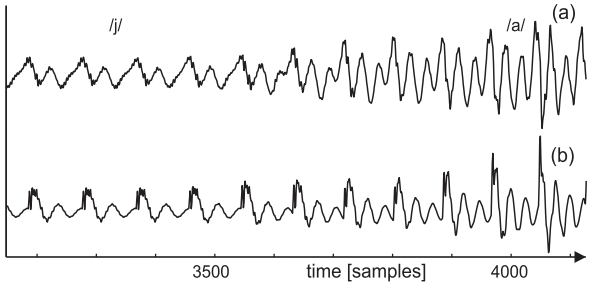

Figure 8: Segment of synthesized speech of [javo:l] by model-based diphones with the lossy tube model: (a) with residual-based excitation obtained from the schwa-sound and (b) with impulse train excitation.

speech signal. It can be seen that the synthesized speech signals with the residual-based excitation produces waveforms like natural speech, whereas the impulse train causes unnatural peaks in the synthesized speech waveforms.

4.1. Re-synthesis of words

To assess the influence of the lossy tube model without concatenation effects, whole words are analyzed and synthesized, too. The analyses reveal that the main spectral difference between the lossy and lossless model is that the estimated bandwidths are often narrower in the case of the lossless model. In Fig. 9, the resulting magnitude responses of the estimated areas are depicted for the German word “Julia” [jUlja] uttered by a male speaker; the averaging during the

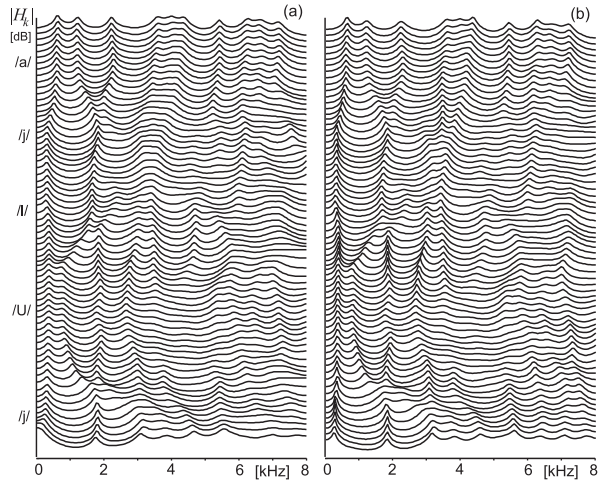

Figure 9: Estimated magnitude responses of word [jUlia] by optimization with averaging: (a) lossy tube model and (b) lossless tube model.

optimization is chosen with J_{ir}. Underestimating of bandwidth is usually caused by overemphasizing the harmonics of voiced speech. One effect of too small bandwidths concerning synthesis or re-synthesis with pitch modification is that artifacts can occur, for example, if the shifted harmonics don’t match the resonances with small bandwidths. Hence, the bandwidths should be not too small. In comparison to bandwidth expansion methods [5-6], the investigations show that the use of the lossy tube model implies an avoidance of bandwidth underestimating inherently.

5. Conclusions

The parameter estimation and the realization of an acoustic synthesis for a model-based approach based on tube models is discussed. The results show that independent estimation without averaging causes, especially for the lossy tube model, fluctuations from frame to frame decreasing the synthesis quality. To yield a continuous trajectory of parameter vectors, the proposed joint analysis of frames is favorable. The main differences of the estimation results relating to the type of tube model is that the estimated areas of the lossy model are more prominent and the bandwidths of the corresponding resonances are expanded in comparison to the lossless model. The voiced excitation can be can be realized by a repeated use of a pitch-modified residual segment.

6. References

[1] Hunt, A.J. and Black, A.W., "Unit Selection in a Concatenative Speech Synthesis System using a Large Speech Database", *Proc. ICASSP, Atlanta 1996*, pp. 373-376.

[2] Childers, D.G. and Wu, K., "Quality of Speech Produced by Analysis-Synthesis ", *Speech Communication, Vol. 9, No. 2, 1990*, pp. 97-117.

[3] Goodyear, C.C. and Wei, D., "Articulatory Copy Synthesis Using a Nine-Parameter Vocal Tract Model", *Proc. ICASSP, Atlanta 1996*, pp. 385-388.

[4] Sondhi, M.M. and Sinder, D.J., "Articulatory Modeling: A Role in Concatenative Text To Speech Synthesis" in *Text To Speech Synthesis: New Paradigms and Advances,* edited by Narayanan, S. and Alwan, A., Prentice Hall PTR, New Jersey, 2004, pp. 63-87.

[5] Tohkura, Y., Itakura, F., Hashimoto, S., "Spectral Smoothing Technique in PARCOR Speech Analysis-Synthesis" *IEEE Trans. ASSP, 26 (5), 1978*, pp. 587-596.

[6] Ekman, L.A., Kleijn, W.B., Murthi M.N., "Spectral Envelope Estimation and Regularization*", Proc. ICASSP, Toulouse 2006*, pp. 245-248.

[7] Shiga, A. and King, S., "Accurate Spectral Envelope Estimation for Articulation-to-speech Synthesis", *Proc. 5th ISCA Speech Synthesis Workshop, 2004*, pp. 19-24.

[8] Schroeter, J. and Sondhi, M.M., "Techniques for Estimating Vocal-Tract Shapes from the Speech Signal", *IEEE Trans. Speech and Audio Proc., 2(1), 1994*, pp. 133-150.

[9] Sorokin, V.N., Leonov, A.S., Makarov, I.S, Tsyplikhin, A.I., "Speech Inversion and Re-synthesis", *Proc. INTERSPEECH, Lisbon 2005*, pp. 3209-3212

[10] Gupta, S.K. and Schroeter, J., "Pitch-synchronous frame-by-frame and segment-based articulatory analysis by synthesis", *J.A.S.A., Vol. 94, 1993*, pp. 2517-2530.

[11] Schnell, K. and Lacroix, A., "Analysis of Lossy Vocal Tract Models for Speech Production", *Proc. INTERSPEECH, Geneva 2003*, pp. 2369-2372.

[12] Schnell, K. and Lacroix, A., "Model Based Analysis of a Diphone Database for Improved Unit Concatenation", *Proc. INTERSPEECH, Lisbon 2005*, pp. 2605-2608.

[13] Sondhi, M. and Schroeter, J.: "A Hybrid Time-Frequency Domain Articulatory Speech Synthesizer", *IEEE Trans. ASSP, 35(7), 1987*, pp. 955-967.

[14] Laine, U.K., "Modeling of lip radiation impedance in the z-domain", *Proc. ICASSP, Paris 1982*, pp. 1992-1995.

[15] Englert, F., "Acquisition of a Diphone Database for German", in *Forum Phoneticum 63, Speech Processing,* edited by Wodarz, H.-W., Hector-Verlag Frankfurt am Main, 1997, pp. 23–32.

[16] Wakita, H., "Estimation of Vocal-Tract Shapes from Acoustical Analysis of the Speech Wave: The State of the Art", *IEEE Trans. ASSP, 27(3), 1979*, pp. 281–285.

[17] Sambur, M.R., Rosenberg, A.E., Rabiner, L.R., McGonegal, C.A., "On reducing the buzz in LPC synthesis", *J.A.S.A., Vol. 63, 1978*, pp. 918-924.

[18] Matsui, K., Pearson, S.D., Hata, K., Kamai, T., "Improving Naturalness in Text-to-speech Synthesis using Natural Glottal Source", *Proc. ICASSP, Toronto 1991*, pp. 769-772.

[19] Schnell, K., "Pitch Modification of Speech Residual Based on Parameterized Glottal Flow with Consideration of Approximation Error", *Proc. ICASSP, Toulouse 2006*, pp. 737-740.