Speech Production Based on Lossy Tube Models: Unit Concatenation and Sound Transitions

Karl Schnell, Arild Lacroix

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

Abstract
The discrete time tube model is well established in speech analysis and synthesis providing a simplified modeling of the vocal tract. The standard lossless tube model is extended by introducing frequency dependent losses. In this contribution it is shown how the lossy vocal tract model can be used for speech production. For that purpose the vocal tract areas of this model can be estimated from speech signals by an optimization algorithm. The estimated model parameters can be used successfully for resynthesis. Furthermore speech units are concatenated by transitions of the vocal tract areas. For transitions from one sound to another a nonlinear area transition is proposed which improve the transition in contrast to a purely linear transition of lossy and lossless models. The investigations show that the lossy tube model is advantageous compared to the lossless standard tube model for speech analysis and speech production.

1. Introduction
For a categorization of models for speech production two features are important. The physical correctness of the model and the resulting speech quality of the output signals. Articulatory speech production models simulate the articulators and their movements directly. Since this is a complex process, it is difficult to produce good audio quality. In comparison simple models like the LPC-model produce moderate audio quality by previous speech analysis in an easy way. A major advantage of the articulatory speech production systems is the close relationship to the human speech production system. In this contribution the standard lossless tube model is extended to yield a more realistic physical modeling of the vocal tract concurrently with a good speech quality.

2. Lossy Tube Model
The tube model consists of a concatenation of uniform acoustic tube elements describing the propagation of forward $x^+$ and backward $x^-$ acoustic traveling wave components. The tube elements are realized by a time delay only in the lossless case. For an introduction of frequency dependent losses the delays are multiplied by a damping system $V(z)$. If a tube length is determined corresponding to $z^{-l/2}$ the relationship between left and right wave quantities of two lossy uniform tubes is depicted in fig. 1 above. An equivalent structure of two adjacent tube elements with the factor $Vz^{-1}$ in the lower and upper path is depicted in figure 1 below. This structure is favorable due to the avoidance of fractional delays.

\[
V(z) = \frac{E(z)}{D(z)} = \frac{(1-0.9151\cdot z^{-1} \cdot 0.9925)}{1-0.9165\cdot z^{-1} + 0.00329 \cdot z^{-2}}
\] (1)

describes the losses caused by yielding walls affecting the low frequencies and viscous friction and heat conduction affecting the high frequencies. The parameters of $V(z)$ are determined by an optimization algorithm for a sampling rate of 22 kHz [1]. The magnitude response of $V$ is shown in fig. 2 describing the damping by two elements.

The pole-zero system

\[
x' = \begin{bmatrix} x_1' \\ x_2' \\ \end{bmatrix} = \begin{bmatrix} \frac{-1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{-1}{2} \\ \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \end{bmatrix} + k(r_i) \begin{bmatrix} 1 \\ \frac{r_i \theta}{1 - r_i \theta} \\ \frac{r_{ii} \theta}{1 - r_{ii} \theta} \\ \end{bmatrix} k(r_i) \begin{bmatrix} 1 \\ \frac{r_i \theta}{1 - r_i \theta} \\ \frac{r_{ii} \theta}{1 - r_{ii} \theta} \\ \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \end{bmatrix}
\] (2)

\(k(r_i)\) describes the kind of sound waves (pressure, flow or...
power waves) while \( \eta \) represents a reflection coefficient. The upper index \( u \) or \( l \) of \( T_i \) indicates the upper or lower location of the lossy delay \( \vartheta = V \cdot z^{-\alpha} \) corresponding to fig. 1 below. The vocal tract model consists of \( N \) tube sections and a tube termination \( \alpha \cdot L(z) \) at the output of the system representing the reflection at the lips. The pole-zero system \( L(z) = F(z)/G(z) \) is realized by the lip-impedance model from Laine [x] adapted to the sampling rate with an additional real damping factor with \( 0.9 < \alpha \leq 1 \). The tube termination at the other end represents the input of the system and is assumed to be reflection free which simplifies the analysis and synthesis of speech. As discussed in [1] the calculation of the transfer function of the lossy tube model causes high powers of \( E \) and \( D \) resulting in a wide value range of the polynomial coefficients of the numerator and denominator of \( H(z) \). This fact causes numerical errors which can be avoided by using the terms

\[
\theta = z^{-1} \cdot V(z) \bigg|_{\vartheta = \varphi} = e^{i\omega}V(e^{i\omega}) = e^{i\omega} \frac{E(e^{i\omega})}{D(e^{i\omega})} \tag{3}
\]

at the interesting frequencies \( \omega = \varphi \). The transfer function can be evaluated at \( \omega = \varphi \) by

\[
H(e^{i\omega}) = \frac{\left(\theta_i\right)^N}{G(e^{i\omega})T_{i-1}(\theta_{\varphi}) + \alpha F(e^{i\omega})T_{i+1}(\theta_{\varphi})} \tag{4}
\]

without numerical errors. \( T_{i-1} \) and \( T_{i+1} \) are the two upper elements of the matrix \( T_i(\vartheta) = \prod_k^{N-1} T_{k+1}(\vartheta)T_k(\vartheta) \) depending on \( \vartheta \).

### 3. Resynthesis and Concatenation of Speech

The parameters of the lossy vocal tract model can be estimated from speech signals. Prior to the parameter estimation the speech signal is filtered by an adaptive preemphasis to eliminate the influence of excitation and radiation. The preemphasis is performed by a repeated linear prediction of the low resonances and the corresponding vocal tract areas can be estimated from the spectrum by minimizing the error

\[
e(S,H) = \frac{1}{\pi} \int_0^\infty S(\omega) \frac{H(e^{i\omega})}{|H(e^{i\omega})|^2} d\omega \rightarrow \min \tag{5}
\]

representing an inverse filtering process in the spectral domain. For a subsequent resynthesis the filtered speech signal \( s \) is segmented into overlapping frames. In the case of voiced speech each frame contains few speech periods. The periods are marked at the zero crossings and the overlapping of the frames extends over one period. The analysis yields for each frame one vector \( a = (a_0,a_1,\ldots,a_{N-1},a_N)^T \) representing the logarithmic areas of the vocal tract model. Since the ratios of the areas are more relevant for the spectral behavior as the areas itself, the description of the logarithmic vocal tract areas is chosen. The parameters of the tube model are controlled by the resulting time sequence \( a(\vartheta) \) of area vectors. In the case of voiced speech the tube model is excited by an impulse train plus high-pass filtered noise. The resynthesized examples show that for a good audio quality it is important that the vocal tract area configurations are sufficiently smooth in time. The smoothness of the vocal tract area movement can be ensured by joint estimation of adjacent periods. For that purpose the analysis of the signal frame sequence is processed in a particular way. At first each frame is analyzed by the optimization algorithm with only few iterations separately. Afterwards the resulting coefficients are averaged with the coefficients of the adjacent frames. These averaged coefficients represent the initial parameter vector for the next iteration of the optimization algorithm. After several iterations the resulting coefficients are averaged again and so on. This processing guarantees that the parameters converge in a solution with a smooth area movement in time. Another improvement of the speech quality can be achieved by constraining the reflection coefficients during the optimization to absolute values smaller than 0.985 or 0.99. Figure 4 (a) and (b) show the spectrograms of the utterance ‘value’ of the original speech and resynthesis with modified fundamental frequency. It can be seen that the spectrograms are similar.

### 3.2. Concatenation of Speech Units

The analysis of different utterances and the resulting vocal tract area configurations can be used to produce new speech utterances. In the following example the utterance ‘value’ is produced with the aid of area configurations of diphones. For that purpose the area sequences of the diphones [vae], [æel], [li] and [ju:] are obtained by the analysis of the words ‘vacuum’, ‘place’, billiards, and you. For synthesis the area sequences are concatenated. To achieve good results, it’s important that the area transitions between the area configurations of the diphones are continuous. Therefore a linear area transition...
within the adjacent periods of the diphone boundaries are processed to smooth the area transition from one diphone to another which can be seen in fig. 5. The magnitude responses at the concatenation position are plotted by thick lines. The spectrogram of the synthesized word is depicted in fig. 4 (c).

**Figure 4:** Spectrograms of the utterance ‘value’: (a) original speech, (b) resynthesis, (c) synthesis by concatenated area sequences of diphones.

4. Sound Transitions

In the previous section the analyzed speech units are reproduced and concatenated by the lossy tube model. The analyzed speech units contain the sound transitions and the resulting movements of the formants. Now transitions from one sound to another sound, which are not included in the units, are modeled by a transition of the vocal tract area configuration of the model. The sound transition is performed by a sequence of vectors \( \mathbf{a}(t) \) containing the logarithmic areas of the vocal tract model. The whole transition starts with an area configuration \( \mathbf{a}_1 \) and ends with the configuration \( \mathbf{a}_2 \).

The vector \( \mathbf{a}_1 \) describes the initial sound while \( \mathbf{a}_2 \) represents the following sound.

4.1. Linear Area Transitions

For a linear area transition the vector sequence is given by

\[
\mathbf{a}(t) = (1 - f(t)) \cdot \mathbf{a}_1 + f(t) \cdot \mathbf{a}_2
\]

for time indices \( t = 0 \ldots T \). The function \( f(t) \) describes the transition in time. The simplest case is \( f(t) = t/T \) which is suitable for a visual presentation. However, a more realistic development in time is given e.g. by a sigmoidal function.

4.2. Nonlinear Area Transitions

The linear area transition is a simplification of the actual vocal tract area alterations which are results of the movements of the articulators mainly described by the tongue shape and their relationship to the vocal tract areas. This process can be modeled by complicated articulatory models which can describe the tongue and their muscles e.g. [3]. This approach cannot be applied in this case since only the vocal tract areas are available and not the articulators themselves. Therefore two nonlinear transitions are realized which can be combined.

In contrast to the linear transition an intermediate area configuration \( \mathbf{b} \) between the boundary configurations \( \mathbf{a}_1 \) and \( \mathbf{a}_2 \) is introduced. The intermediate configuration \( \mathbf{b} \) is obtained by nonlinear processing of \( \mathbf{a}_1 \) and \( \mathbf{a}_2 \). The complete transition consists of two linear transitions; one from \( \mathbf{a}_1 \) to \( \mathbf{b} \) and a following from \( \mathbf{b} \) to \( \mathbf{a}_2 \):

\[
\mathbf{a}_1 \rightarrow \mathbf{b} \rightarrow \mathbf{a}_2.
\]

The calculation of \( \mathbf{b} \) is achieved by a combination of two approaches considering a neutral configuration and horizontal area movements.

4.2.1. Neutral Sound Component

The first approach considers a movement to a central position. This position is related to a neutral sound e.g. the schwa-sound. The assumption is that the vocal tract areas come close to the neutral position during the sound transition. Therefore the intermediate configuration gets a component from the linear average \( \mathbf{a}_{12} = (\mathbf{a}_1 + \mathbf{a}_2)/2 \) and from the neutral area configuration \( \mathbf{n} \). The combination

\[
b'(k) = (1 - f_s(k)) \cdot \mathbf{a}_{12} + f_s(k) \cdot \mathbf{n}
\]

with \( 0 < f_s < 1 \) leads not yet satisfactory results. However, by a weighted combination with

\[
b(k) = (1 - f_s(k)) \cdot \mathbf{a}_{12} + f_s(k) \cdot n(k)
\]
the resulting transitions are improved. The weighting function \( f_k(k) \) depends on the position \( k \) of the areas and tends towards zero in the regions of \( k=0 \) and \( k=24-28 \). In the middle locations \( f_k(k) \) reaches the maximum value, so that the areas in the region of the lips and glottis termination are more influenced by the two configurations \( a_1 \) and \( a_2 \), representing the previous and next sound whereas the areas located in the middle are more influenced by the neutral configuration \( n \). The neutral configuration \( n \) can be obtained by an analysis of the isolated spoken schwa-sound although a zero vector is possible, too.

4.2.2. Horizontal Area Movement

The second approach considers horizontal area movements. For that purpose an area configuration between \( a_1 \) and \( a_2 \) is created by repeating or eliminating tube elements describing horizontal movements. In a first step the optimal path between the vectors \( a_1 \) and \( a_2 \) is calculated by dynamic programming or Viterbi search respectively which is depicted in fig. 6. The path can lengthen appropriate tube elements to match the two vectors. In contrast to the usual application of the dynamic programming the path varies space instead of time. The lengthening of \( a_1 \) and \( a_2 \), at particular locations \( \lambda_1 \) for \( a_1 \) and \( \lambda_2 \) respectively for \( a_2 \), results in the vectors \( \tilde{a}_1 \) and \( \tilde{a}_2 \). Therefore \( \tilde{a}_1 \) and \( \tilde{a}_2 \) are shortened by omitting suitable tube elements. To maintain the correlation between the two vectors, tube elements are eliminated in \( \tilde{a}_1 \) and \( \tilde{a}_2 \) jointly.

For the elimination every second position of \( \tilde{a}_1 \) and \( \tilde{a}_2 \) are used alternately until the correct length is reached. The resulting area configuration \( m \) is mixed linear to the intermediate area configuration \( b \) resulting in \( b' \), so that \( b \) consists of components of \( a_1, n \) and \( m \).

Figure 6: Viterbi search for adaption of \( a_1 \) and \( a_2 \).

**4.3. Comparison of Transitions**

For a comparison the resulting magnitude responses of the transitions between the plosive /d/ and the vowel /i/ are shown in fig. 7. The differences are caused by the kind of tube model and the transition processing. The top graph in fig. 7 shows the analysis of original speech [di] by a lossy tube model. The results by a linear transition with a standard lossless tube model (d) show the greatest difference in comparison to the analysis of the original transition (a). It can be seen that the bandwidths are smaller as in the case of the lossy model. The frequency responses of the middle configuration and of the intermediate area configuration \( b \) are plotted by thick lines. The intermediate area configuration realized by the nonlinear transition (b) is more comparable to the original middle configuration (a) as the linear transition of (c). Examples show that improvements of the transitions by the lossy model and the nonlinear area transition can be achieved.

Figure 7: Magnitude responses of the transition [di]: (a) obtained by analysis of original speech, (b) nonlinear area transition with lossy tube model, (c) linear transition with lossy tube model, (d) linear area transition with lossless standard tube model.

5. Conclusions

The investigations show that the lossy vocal tract model is suitable for speech production. Resynthesis of analyzed speech can be well performed as well as unit concatenation for the synthesis of new speech utterances. For sound transitions a nonlinear area transition is discussed. The advantages of the lossy vocal tract model in comparison to the standard lossless tube model result from a closer relationship to the acoustics of the human speech production system.

6. References