PITCH MODIFICATION OF SPEECH RESIDUAL BASED ON PARAMETERIZED GLOTTAL FLOW WITH CONSIDERATION OF APPROXIMATION ERROR

Karl Schnell

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

ABSTRACT

Pitch modification plays an important part in the field of speech generation. For this task various approaches exist using time signals like PSOLA and related methods. Further approaches exist by using a parameterized model of the glottal flow or its derivative. In this contribution a hybrid approach is proposed based on a partly parameterized model of the glottal flow plus the consideration of the error of the model approximation; both of them are obtained from the residual of speech. By the consideration of the approximation error a high naturalness of the synthetic speech can be achieved; concurrently the use of the glottal flow model allows the alteration of pitch in a large range without appreciable decrease of speech quality.

1. INTRODUCTION

It is well known that voiced speech contains besides its periodic structure also non-periodic components as fluctuations and noise. Parameterized models of the voiced excitation like the LF-model [1] or polynomial model [2, 3] cannot represent the finer points of voiced excitation. These approximations of the glottal flow or its derivative yield not best possible speech quality but allow great range of pitch modification and, furthermore, the model parameters can be directly interpreted often. In contrast to that PSOLA techniques enable higher speech quality by using time signals of voiced speech; but can fail in the case of greater pitch modification. It can be advantageous to modify the pitch of the residual like in LPC-PSOLA, however, drawbacks are still remaining.

The proposed method modifies also the pitch of the residual, however, it uses a polynomial glottal flow model for an approximation of the low-pass filtered residual; the region of the glottal closure is treated separately. To retain the details of the original speech, the approximation error between the filtered residual and the polynomial model is considered; this is realized by processing of the error and subsequently adding to the new pitch modified period. In this way the finer points of the excitation like noise can be preserved yielding realistic relationships of noisy and periodic components of voiced speech, valid for the whole frequency range.

In contrast to that other approaches divide the frequency range in two parts for harmonic and noisy characteristics like the harmonic plus noise model (HNM) [4].

2. DECOMPOSITION OF GLOTTAL FLOW

Since the pitch modification is applied to the residual, at first a linear prediction is applied to the speech signal *s*:

$$r = h_p * s \ .$$

h_p is the impulse response of the prediction error filter. The resulting residual signal r is not convenient for parametric modeling due to its irregular course. Therefore the residual is filtered by a de-emphasis (low-pass) with the transfer function:

$$H_{low} = \frac{1}{(1 - 0.99 \cdot z^{-1}) \cdot (1 - 0.95 \cdot z^{-1})} \qquad (1)$$

resulting the signal $g = h_{low} * r$ which is related to the glottal flow of voiced excitation. The glottal closure is represented by a negative peak in the residual r and by a bump in the estimated glottal flow g, which can be seen in fig. 1(a)-(b). The positions of the glottal closure in the glottal flow g can be marked automatically with the aid of the negative peaks of residual.

2.1. Polynomial approximation of glottal flow

The first step of the proposed approach is related to the polynomial approximation described in [2, 3] which was mainly applied for speech coding. In contrast to [2, 3] the polynomial approximation is here not applied to the whole pitch period since a polynomial description is not optimal for the complete glottal circle. The region of the glottal closure is an important incident and cannot be described in detail by polynomials; therefore this region is treated separately. A small segment of the glottal closure is stored by a time signal represented by the vector

$$\boldsymbol{y}^i = \left(y^i(1), y^i(2), \ldots y^i(L_y) \right),$$

the index i indicates the i-th pitch period. Here a range of five samples in front of the glottal closure and five samples behind are used; the complete length of the vectors $\boldsymbol{y}^i$ is $L_y = 10$. Fig. 1 shows the segmentation of periods. The segments between the $\boldsymbol{y}^i$ are denoted by the vectors

$$\boldsymbol{x}^i = \left(x^i(1), x^i(2), \ldots x^i(L_x^i) \right).$$

Since the period length can vary from period to period the lengths L_x^i of the vectors $\boldsymbol{x}^i$ depend on the index i of the period number. A full period is described by the composed vector

$$\boldsymbol{g}^i = \left[\boldsymbol{y}^i, \boldsymbol{x}^i \right] = \left(y^i(1), \ldots y^i(L_y), x^i(1), \ldots x^i(L_x^i) \right);$$

the bracket [,] describes here a concatenation of vectors.

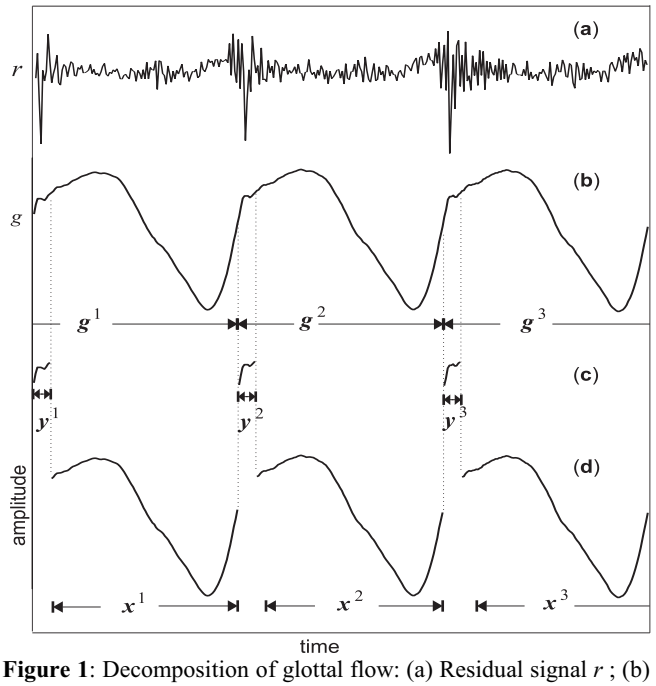

Figure 1: Decomposition of glottal flow: (a) Residual signal r ; (b) estimated glottal flow signal g and glottal flow periods $\boldsymbol{g}^i$; (c) segments $\boldsymbol{y}^i$ of glottal closure; (d) remaining segments $\boldsymbol{x}^i$.

The vectors $\boldsymbol{x}^i$ are approximated by a polynomial expansion $\hat{\boldsymbol{x}}^i$ with:

$$\hat{x}^i(n) = \sum_{k=0}^{N} p_k^i \cdot \left(\frac{n}{L_x^i}\right)^k , \quad n = \left(1 \ldots L_x^i\right) . \tag{2}$$

The polynomial coefficients p_k^i are determined in a mean square sense. The order N is chosen with $N = 7$. Fig. 2 shows an example of the polynomial approximation for $\boldsymbol{x}^i$.

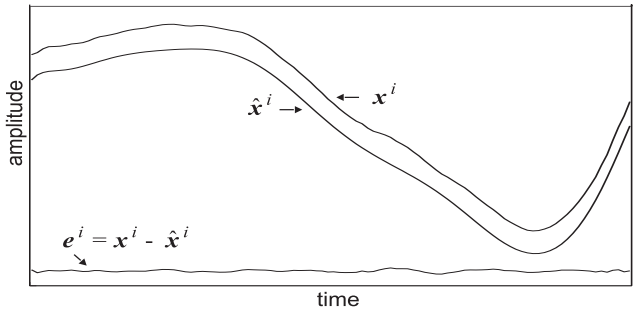

Figure 2: Polynomial approximation $\hat{\boldsymbol{x}}^i$ of $\boldsymbol{x}^i$ and error $\boldsymbol{e}^i$.

2.1. Approximation error

The error vector $\boldsymbol{e}^i$ is the subtraction of segment $\boldsymbol{x}^i$ and the polynomial approximation $\hat{\boldsymbol{x}}^i$:

$$\boldsymbol{e}^i = (e^i(1) \ldots e^i(L_x^i)) = \boldsymbol{x}^i - \hat{\boldsymbol{x}}^i .$$

The original glottal flow period is $\boldsymbol{g}^i = [\boldsymbol{y}^i, \hat{\boldsymbol{x}}^i + \boldsymbol{e}^i]$. Since the error signal still contains information about fluctuations and noise of speech, the error signal is considered for the subsequent synthesis. The importance of this consideration is obvious for voiced fricatives, however, also for vowels the details of speech signals are relevant for the naturalness of speech. For synthesis the error signal $\boldsymbol{e}^i$ has to be adapted to the modified pitch. For that purpose the characteristic features of these segments should be taken into account. Besides the region of glottal closure the region of glottal opening is interesting. This glottal opening can be observed in the residual often by an impulse or a peak-shaped region. Since this region should be preserved unaltered as good as possible, each signal $\boldsymbol{e}^i$ is stored by three segments. Two segments $\boldsymbol{e}_l^i$ and $\boldsymbol{e}_r^i$ for the left and for the right part of $\boldsymbol{e}^i$, describing the neighborhood of glottal closure, and one segment $\boldsymbol{e}_c^i$ for the center part, describing the glottal opening with neighboring regions:

$$\begin{aligned} \boldsymbol{e}_l^i &= \left(e^i(1), e^i(2), \ldots e^i(L_x^i/2 - m)\right) \\ \boldsymbol{e}_c^i &= \left(e^i(m), \ldots e^i(L_x^i - m)\right) \\ \boldsymbol{e}_r^i &= \left(e^i(L_x^i/2 + m), \ldots e^i(L_x^i)\right). \end{aligned} \tag{3}$$

m indicates the number of left and/or right sample values which are ignored for the individual parts. A proper value is about $m = 0.1 \cdot L_x^i$.

2. COMPOSITION OF GLOTTAL FLOW

For synthesis the vectors $\boldsymbol{y}^i$, $\boldsymbol{e}_l^i$, $\boldsymbol{e}_c^i$, and $\boldsymbol{e}_r^i$ and the coefficients p_k^i of the analyzed glottal flow periods $\boldsymbol{g}^i$ are used to generate new periods $\boldsymbol{g}_{\text{new}}^i$ with the required period length. The sample vector $\boldsymbol{y}^i$ of the glottal closure is used unaltered for the new period $\boldsymbol{g}_{\text{new}}^i$. The pitch modification of the period is performed in the remaining region. For that purpose the polynomial is evaluated by

$$\hat{x}_{\text{new}}^i(n) = \sum_{k=0}^{N} p_k^i \cdot \left(\frac{n}{L_{x\text{new}}^i}\right)^k , \quad n = \left(1 \ldots L_{x\text{new}}^i\right) \tag{4}$$

describing an interpolation of the argument. The segment of the polynomial vector $\hat{\boldsymbol{x}}^i$ is stretched or shortened to the new vector $\hat{\boldsymbol{x}}_{\text{new}}^i$ with the length $L_{x\text{new}}^i$. By analogy to the length modification of $\boldsymbol{x}^i$ the error vector has to be adapted to the new length $L_{x\text{new}}^i$, too. For that purpose in a first step the error signals $\boldsymbol{e}_l^i$, $\boldsymbol{e}_c^i$, and $\boldsymbol{e}_r^i$ are processed by bounding and windowing. After that the processed error signals are combined resulting the error signal $\boldsymbol{e}_{\text{new}}^i$ with the required period length. For the first step the error signals $\boldsymbol{e}_l^i$, $\boldsymbol{e}_c^i$, and $\boldsymbol{e}_r^i$ are truncated or expanded to the vectors $\tilde{\boldsymbol{e}}_l^i$, $\tilde{\boldsymbol{e}}_c^i$, and $\tilde{\boldsymbol{e}}_r^i$ with the new length $L_{x\text{new}}^i$:

$$\begin{aligned} \tilde{\boldsymbol{e}}_l^i &= \left(e_l^i(1), e_l^i(2), \ldots e_l^i(L_{x\text{new}}^i)\right) \\ \tilde{\boldsymbol{e}}_c^i &= \left(e_c^i(\Delta L_x^i/2 + 1), \ldots e_c^i(L_x^i/2), \ldots e_c^i(L_x^i - \Delta L_x^i/2)\right) \\ \tilde{\boldsymbol{e}}_r^i &= \left(e_r^i(\Delta L_x^i + 1), e_r^i(\Delta L_x^i + 1), \ldots e_r^i(L_x^i)\right). \end{aligned} \tag{5}$$

$\Delta L_x^i = L_x^i - L_{x\text{new}}^i$ is the difference between the original length and the new length. If error vectors are expanded, indices of $\boldsymbol{e}_l^i$, $\boldsymbol{e}_c^i$, and $\boldsymbol{e}_r^i$ are outside of the vectors; in this case the corresponding

values are set to zero: $e_l^i(n) = e_c^i(n) = e_r^i(n) := 0$ in eq. (5). Then the vectors of (5) are weighted with the windows $\boldsymbol{w}_l^i$, $\boldsymbol{w}_c^i$, and $\boldsymbol{w}_r^i$ which can be seen in fig 3. The values of the windows are constantly equal to one in the left, center, and right region of $\boldsymbol{w}_l^i$, $\boldsymbol{w}_c^i$, and $\boldsymbol{w}_r^i$, respectively. Besides these regions the windows consist of regions which are constantly zero, in between a linear transition exists for continuity; the length of the transition is t_1. The overlapping between the windows $\boldsymbol{w}_l^i$, $\boldsymbol{w}_c^i$, and $\boldsymbol{w}_r^i$ with values greater zero can be determined by t_2 which can be seen in fig. 3. The value of t_2 is in the range of $0 \le t_2 \le t_1$; $t_2 = 0$ indicates no overlapping whereas $t_2 = t_1$ stands for full overlapping.

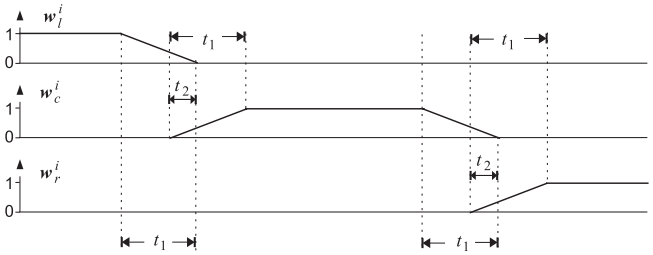

Figure 3: Window functions $\boldsymbol{w}_l^i$, $\boldsymbol{w}_c^i$, and $\boldsymbol{w}_r^i$ with parameters t_1 and t_2.

The vectors $\tilde{\boldsymbol{e}}_{wl}^i$, $\tilde{\boldsymbol{e}}_{wc}^i$, and $\tilde{\boldsymbol{e}}_{wr}^i$ are the error signals of (5) processed by a element-by-element multiplication with the windows $\boldsymbol{w}_l^i$, $\boldsymbol{w}_c^i$, and $\boldsymbol{w}_r^i$, respectively:

$$\begin{aligned} \tilde{e}_{wl}^i(n) &= w_l^i(n) \cdot e_l^i(n) \qquad \text{for} \quad n = 1 \ldots L_{x\text{new}}^i \\ \tilde{e}_{wc}^i(n) &= w_c^i(n) \cdot e_c^i(n) \qquad \text{for} \quad n = 1 \ldots L_{x\text{new}}^i \\ \tilde{e}_{wr}^i(n) &= w_r^i(n) \cdot e_r^i(n) \qquad \text{for} \quad n = 1 \ldots L_{x\text{new}}^i . \end{aligned} \tag{6}$$

The new error vector $\boldsymbol{e}_{\text{new}}^i$ is the sum of these processed vectors:

$$\boldsymbol{e}_{\text{new}}^i = \tilde{\boldsymbol{e}}_{wl}^i + \tilde{\boldsymbol{e}}_{wc}^i + \tilde{\boldsymbol{e}}_{wr}^i$$

and the sum of $\hat{\boldsymbol{x}}_{\text{new}}^i$ and $\boldsymbol{e}_{\text{new}}^i$ describes the new vector

$$\boldsymbol{x}_{\text{new}}^i = \hat{\boldsymbol{x}}_{\text{new}}^i + \boldsymbol{e}_{\text{new}}^i$$

representing the original segment $\boldsymbol{x}^i$ with modified length. For pitch modification of the periods $\boldsymbol{g}^i = [\boldsymbol{y}^i, \boldsymbol{x}^i]$ the vector $\boldsymbol{x}^i$ is replaced by $\boldsymbol{x}_{\text{new}}^i$ resulting

$$\boldsymbol{g}_{\text{new}}^{\text{i}} = \left[\boldsymbol{y}^i, \boldsymbol{x}_{\text{new}}^i \right] = \left(y^i(1), \ldots y^i(L_y), x_{\text{new}}^i(1), \ldots x_{\text{new}}^i(L_{x\text{new}}^i) \right).$$

Fig. 4 shows glottal flow periods with different pitch modifications; the length of the original period is about 119 sample units. The glottal flow periods $\boldsymbol{g}_{\text{new}}^{\text{i}}$ are concatenated to the vector $\boldsymbol{g}_{\text{new}} = [\ldots \boldsymbol{g}_{\text{new}}^{\text{i-1}}, \boldsymbol{g}_{\text{new}}^{\text{i}}, \boldsymbol{g}_{\text{new}}^{\text{i+1}}, \ldots]$ representing the new glottal flow signal $g_{\text{new}}(n)$. A residual description can be obtained by filtering of $g_{\text{new}}(n)$ by a high-pass H_{high} which is the inverse system of H_{low}. The resulting signal

$$r_{\text{new}} = h_{high} * g_{\text{new}}$$

is a pitch modified version of the original residual *r*.

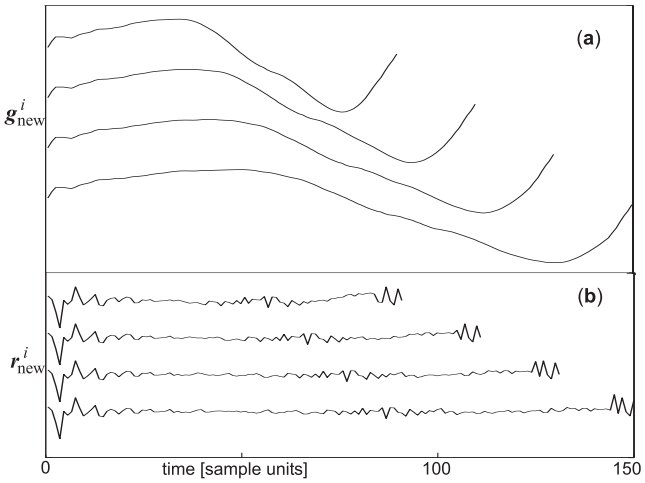

Figure 4: (a) Composed glottal flow periods $\boldsymbol{g}_{\text{new}}^{\text{i}}$ by different ranges of pitch modification; (b) corresponding residual representation.

For speech generation the modified residual is filtered by the impulse response h_p^{inv} of the inverse system of the linear prediction error filter:

$$s_{\text{new}} = h_p^{\text{inv}} * r_{\text{new}} .$$

Other systems can be also used to introduce the spectral envelope of the speech.

3. SYNTHESIS EXAMPLES

In the following pitch modification is demonstrated by examples of a vowel and a voiced fricative; the sampling rate is 16 kHz. In Fig. 5 the schwa-sound [@] and in fig. 6 the voiced fricative [z] is treated. A segment of the natural speech signal is shown in graphs 5(a) and 6(a), respectively. The graphs 5(b)-(c) and 6(b)-(e) show segments of the synthesized sounds with different pitch modifications. The time signals show that by the parametric glottal flow model the shapes appear realistic; it should be mentioned that in the case of different period lengths the shape vary. In fig. 5 additionally short-term spectra of the original and pitch modified vowel are shown. The spectra in fig. 5 show that the relationships between harmonic and non-harmonic components of the pitch modified synthesis of 5(e)-(f) are comparable to these of the original speech 5(d). Besides the spectral characteristics the temporal noise appearance of the modified periods 6(b)-(e) is similar to that of the original in fig. 6(a). It is known that the temporal distribution of the noisy components of voiced speech is relevant for the perception and can be modeled e.g. by modulated noise [4]. Synthesis tests indicate that the realization of the noisy speech components with the aid of the approximation error yields better results as a realization by modulated noise. The influence of the parameter t_2 on the speech quality was for the performed synthesis examples not so strong as excepted; $t_2 = 0.6 \cdot t_1$ is chosen for the results of the figures.

Due to the consideration of the finer points of voiced speech, the pitch modified versions sound comparably to the original sounds, excepting the different fundamental frequency. In the case of similar fundamental frequency the original and synthesized signal

are practically undistinguishable. Synthesized signals with stronger pitch modification sound also naturally without appreciable artifacts. The difference between synthesized speech examples with and without the consideration of the approximation error depends on the sounds; especially for voiced fricatives the distinction is particularly prominent.

3.1. Usage of pitch modification algorithm

The pitch modification algorithm can be used to modify the pitch of speech units like diphones for speech synthesis. Since the modification of pitch is treated in the domain of glottal flow instead of the speech signal itself the segmented low-pass filtered residual of the speech units are necessary. The separate handling of excitation and vocal tract characteristics allows flexible synthesis structures. The filter for the spectral characteristics of the vocal tract can be realized by the standard LPC-model, however, more realistic vocal tract models like the lossy tube model can be used [5], too. The results of synthesis encourage also the use of the more sophisticated vocal tract models.

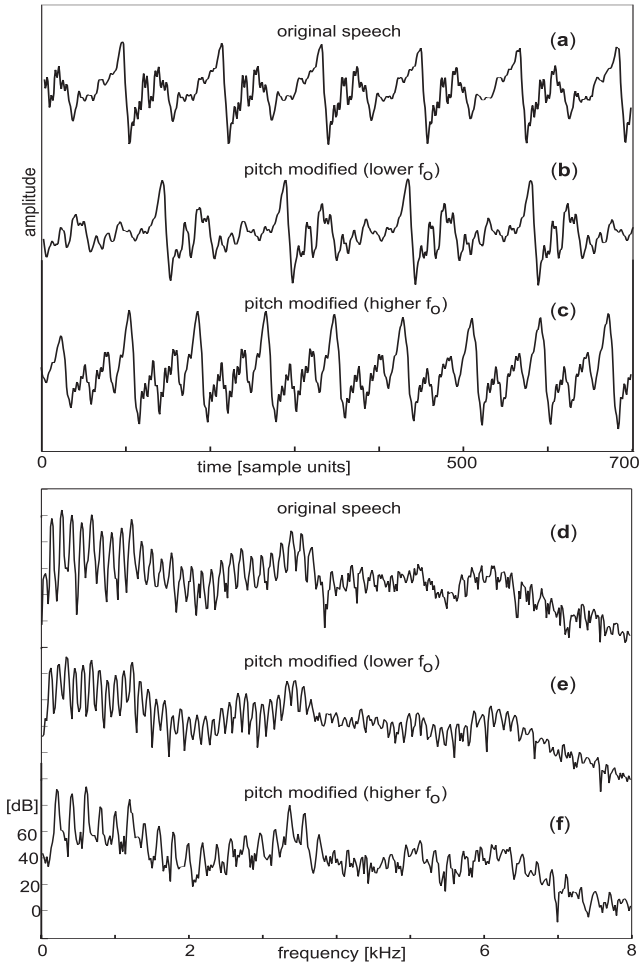

Figure 5: Original and synthesized schwa-sound [@]: (a) Time signal and (d) spectrum of the analyzed natural sound (period length about 118); (b) time signal and (e) spectrum of synthesized sound with period length 146; (c) time signal and (f) spectrum of synthesized sound with period length 81.

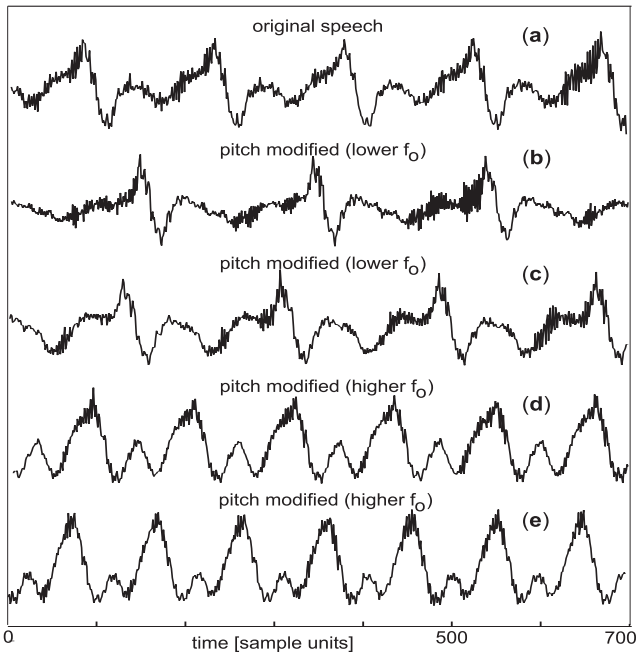

Figure 6: Time signals of original and synthesized voiced fricative [z]: (a) Analyzed natural sound, (b)-(e) synthesized sound with different pitch modifications.

4. CONCLUSIONS

A hybrid approach for pitch modification is proposed using a parameterized glottal flow model as well as time signals for glottal closure and approximation error. Due to the consideration of the glottal closure and the approximation error the finer points of voiced speech can be preserved for the modified speech. This maintains the frequency dependent relationships between harmonic and noise furthermore temporal noisy components are preserved yielding natural sounding synthetic speech.

5. REFERENCES

[1] G. Fant, J. Liljencrantz, Q. Lin, "A four parameter model of glottal flow", *STL-QPSR*, 2-3, pp. 119-156, (1985).

[2] D.G. Childers, T.H. Hu, "Speech synthesis by glottal excited linear prediction", *J. Acoust. Soc. Am.* 96(4), pp. 2026-2036, (1994).

[3] P.H. Milenkovic, "Voice source model for continuous control of pitch period", *J. Acoust. Soc. Am.* 93(2), pp. 1087-1096, (1993).

[4] Y. Stylianou, J. Laroche, E. Moulines, "High-Quality Speech Modification based on a Harmonic + Noise Model", *Proc. EUROSPEECH'95*, Madrid Spain, pp. 451-454, (1995).

[5] K. Schnell, A. Lacroix, "Speech Production Based on Lossy Tube Models: Unit Concatenation and Sound Transitions", *Proc. INTERSPEECH-2004 ICSLP*, Jeju Korea, pp. 505-508, (2004).