Low Resource TTS Synthesis Based on Cepstral Filter with Phase Randomized Excitation

Guntram Strecha, Matthias Eichner

Institut für Akustik und Sprachkommunikation
Technische Universität Dresden
{guntram.strecha,matthias.eichner}@ias.et.tu-dresden.de

Abstract

In this paper we present the acoustic synthesis of a low resource Text-To-Speech (TTS) system based on a 7th order cepstral filter. The excitation signal is designed in frequency domain by a two parameter model. This model is able to generate the excitation signal for both, voiced and unvoiced segments. The sets of filter coefficients represent the speech units and are stored in a compressed form in the inventory of the TTS system. An inventory which is normally used by a concatenative synthesis system is transformed to obtain the inventory for the proposed system. The compression method consists of a lifter and an interpolation technique to describe the temporal progression of the cepstral features. Additional spectral warping is applied to prefer lower frequency components in order to preserve the spectral structure in the compression step. This warping method offers the possibility to change the voice characteristics of the synthesized speech without additional computational or algorithmic efforts. We integrated the proposed synthesis method in our multilingual TTS system and achieved high quality speech syntheses using sampling rates up to 32 kHz with an average bit rate of 14 $^{\text{kBit}}/_{\text{s}}$ and an inventory compression rate of 36:1.

1. Introduction

Improving naturalness of Text-To-Speech (TTS) systems is an important research task. At the same time the consumption of algorithmic resources has to be considered to enable various (mobile) multimedia applications.
TTS systems based on concatenating pre-recorded speech units achieve a considerable audio quality. The naturalness increases as the speech units pooled in the inventory are chosen larger and the amount of alternative units increases [1]. This approach requires a large amount of pre-recorded and stored speech data, which is still not practicable for the use in mobile multimedia devices.
A possible solution for decreasing memory consumption is the compression of the inventory containing the speech segments. In [2] we investigated the acoustic synthesis with encoded diphone inventories using the Adaptive Multi-Rate codec. With this ACELP (Algebraic Code Excitated Linear Prediction) based speech codec compression rates up to 18:1 at 8 kHz and 1:26 at 16 kHz are achievable. Continuing the work of [2, 3] and [4], where the baseline system is published, this paper describes the acoustic synthesis using a cepstral coded inventory.

For the naturalness of parametric synthesis a mixed excitation in voiced parts seems to be essential [5, 6]. In contrast to Helmholtz [7] and contemporary researchers, Schroeder [8] and thereafter Patterson [9] proved that the timbre of sounds not solely depends on the number and relative strength of its partial tones, but also on the their initial phases. Taking this into account the excitation signal of the cepstral filter is generated regarding the phase as parameter (section 2.2).
Section 2 gives a brief overview of the cepstrum basics, including cepstral analysis (section 2.1) and synthesis from cepstral parameters (section 2.2). The integration into our synthesis system and the applied inventory coding and compression methods are described in section 3.

2. Cepstrum basics

2.1. Cepstral analysis

The real cepstrum is defined as:

$$c(n) = F^{-1}\left\{\left|\ln\left(F\left\{s(n)\right\}\right)\right|\right\}, \tag{1}$$

where F denotes the N point discrete Fourier serie, $s(n)$ the windowed k-th speech frame and $c(n)$ the N cepstral coefficients. As $c(n)$ is symmetric at $^{N}/_{2}$ with

$$\ln\left(\left|F\left\{s(n)\right\}\right|\right) = c(0) + 2\sum_{n=1}^{N-1} c(n)\cos\left(n\omega T\right) \tag{2}$$

the minimum phase transfer function $\bar{H}(z)$, which approximates the envelope of $|F\{s(n)\}|$, is given by

$$\bar{H}(z) = e^{c(0)} e^{2\sum_{n=1}^{N_0-1} c(n)z^{-n}} = \beta e^{2C(z)}, \tag{3}$$

where N_0 is the cut off quefrency of the lifter. In contrast to the Linear Predictive Coding (LPC) the cepstral analysis models peaks (formants) just as valleys (antiformants), while the LPC models only the peaks. Therefore the magnitude spectrum of the excitation signal of the synthesis filter has not to be chosen as carefully as for the LPC synthesis filter.

2.2. Cepstral synthesis

For the synthesis of the speech signal a filter is required to realize the transfer function $\bar{H}(z)$ given in (3). Following [10], the synthesis filter consists of two nested filters. As shown in figure 1 the outer IIR filter approximates the exponential function and the inner FIR filter performs the transfer function $C(z)$.

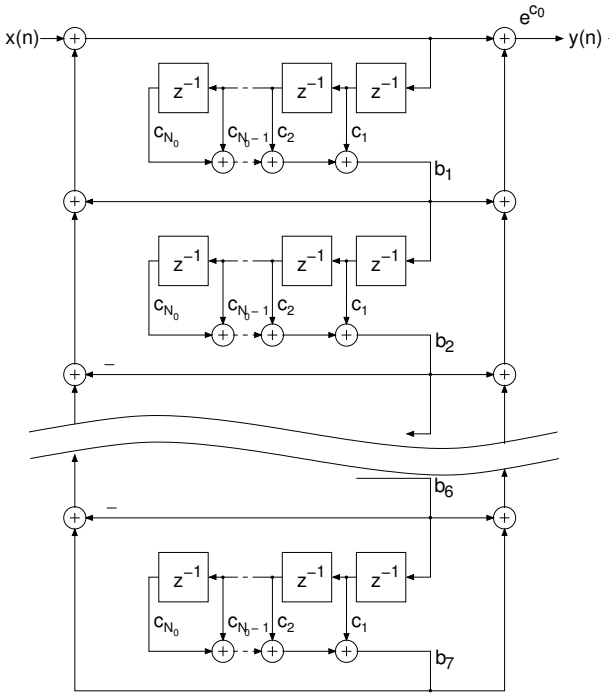

Figure 1: Cepstral synthesis filter with 7th order Padé approximation (direct form II realization).

To realize the exponential function of (3) the M-th order Padé approximation [11]:

$$\begin{aligned} e^{2x} &\approx \frac{a_0 + a_1 x + a_1 x^2 + \cdots + a_M x^M}{a_0 - a_1 x + a_2 x^2 - \cdots + a_M (-1)^M x^M} \\ &= \frac{\sum_{m=0}^{M} a_m x^m}{\sum_{m=0}^{M} a_m (-1)^m x^m} \end{aligned} \tag{4}$$

with:

$$a_m = \frac{(2M-m)!M!}{(2M)!(M-m)!m!} 2^m \tag{5}$$

is used. The approximant x, which corresponds to the filter delays of the outer IIR filter, is replaced by the inner filter $C(z)$.

$$x = C(z) = \sum_{n=1}^{N_0-1} c(n) z^{-n} \tag{6}$$

The coefficients of the outer filter b_m, shown in figure 1, finally result in:

$$b_m = \frac{a_m}{a_{m-1}} = \frac{2(M-m+1)}{(2M-m+1)m}, \quad 0 < m \leq M \tag{7}$$

3. The System

3.1. Inventory coding

Similar to the AMR encoding procedure outlined in [2] the speech segments of the diphone inventory are processed separately. Each of the segments j is splitted into K_j frames spanning two speech periods. Two consecutive frames overlap by one speech period. The cepstral coefficients $c_k(n)$ are calculated from the windowed k-th frame $s_k(n)$ according to (1). We apply a spectral warping to get a higher resolution in the lower frequency components where the important spectral properties of the speech signal are located. The warping is done by a filter, which realizes the transfer function of a first order allpass (equation 9). The amount of spectral warping can be adjusted by the warping factor λ. For $\lambda = 0.57$ and a sampling rate of 16 kHz the frequency transformation approximates the Mel-scale[12].

$$\tilde{C}_k(\tilde{z}) = \sum_{n=0}^{N_0-1} \tilde{c}_k(n) \tilde{z}^{-n} \approx \sum_{n=0}^{N-1} c_k(n) z^{-n} = C_k(z) \tag{8}$$

$$\text{with} \quad \tilde{z}^{-1} = \frac{z^{-1} - \lambda}{1 - \lambda z^{-1}} \quad \text{and} \quad N_0 \leq N. \tag{9}$$

We assume that the spectral envelope is modeled sufficiently accurate if only a slight lifter is applied to remove the fundamental frequency peak and no spectral warping is done ($N_0 \simeq N$, $\lambda = 0$). Further liftering ($N_0 < N$) results in a smoother envelope and reduces the number of parameters needed to describe the speech frame. The warping factor is chosen with respect to the number of of cepstral coefficients N_0 after liftering:

$$\lambda = 1.0 - \frac{N_0}{N}. \tag{10}$$

Now, every frame $s_k(n)$ of a speech segment j in the inventory is represented by N_0 coefficients. To further compress the inventory, the resulting $K_j \times N_0$ matrix $\tilde{\mathbf{C}}$ is approximated along the rows k using Tschebysheff polynomials of order M

$$\mathbf{T} = \begin{pmatrix} T_0 \\ T_1 \\ \vdots \\ T_M \end{pmatrix} \quad \text{with} \quad \begin{aligned} T_m(x) &= \cos(m \arccos(x)), \\ x &= \frac{2k}{K_j} - 1 \end{aligned} \tag{11}$$

where the $M \times N_0$ approximation matrix $\hat{C}$ is found by:

$$\mathbf{T}\hat{\mathbf{C}} \approx \tilde{\mathbf{C}} \tag{12}$$

$$\hat{\mathbf{C}} = \left(\mathbf{T}^{\mathbf{T}}\mathbf{T}\right)^{-1} \mathbf{T}^T \tilde{\mathbf{C}} \tag{13}$$

For each segment the elements of $\hat{\mathbf{C}}$ are stored j to the inventory with an accuracy of 16 bit. Therefore, the size of the inventory does not depend on the size of the segments anymore, but:

1. the amount J of speech segments (units) included in the inventory,
2. the amount N_0 of cepstral coefficients $c(n)$, i. e. to the order of the cepstral analysis,
3. the approximation order M,
4. the size of inventory description information (period markers, phoneme lookup table, etc.).

The American-English inventory used in our experiments consists of $J = 1784\,\text{units}$. Using $N_0 = 21$ cepstral coefficients and an approximation order of $M = 5$ the size without inventory description is $N_0 M J \cdot 2$ byte=366 kbyte (with description 398 kbyte). Table 1 gives an overview of currently tested 32 kHz inventories with their uncoded and coded sizes.

Language	No. of units (J) (diphones)	Size / kByte uncoded	coded
US-English	1,784	16,927	398
English	1,466	13,792	327
German	1,176	9,051	262
French	1,014	7,491	226
Spanish	685	4,933	153
Italian	1,097	8,505	239
Mandarin	(Syllables) 1,644	32,747	374
Dutch	1,573	14,563	351

Table 1: Listing of currently tested inventories with their uncoded (PCM, 16 bit, 32 kHz) and cepstral coded sizes (32 kHz, $N_0 = 21$, $M = 5$).

3.2. Acoustic synthesis

In our synthesis system, the acoustic module receives the output stream of the previous TTS modules to synthesize speech. The stream consists of the target phonemes annotated with the target phoneme durations and prosodic targets (F_0 contour and intensities). Using the cepstral coded inventory the acoustic synthesis is done by applying the cepstral filter (fig. 1) as mentioned in section 2.2.
Due to the approximation (13) there is no interdependence between the period length and the update rate of the filter coefficients $c_t(n)$, i. e. at each time $t = \frac{n}{f_s}$ a set of cepstral coefficients can be estimated by the equations (11-12) and

$$x = \frac{2n}{N_t} - 1, \quad 0 \le n \le N_t \tag{14}$$

where N_t is the target length of the processed segment. Therefore, the control of the phoneme durations is a lot easier than with Overlap-And-Add (OLA) based methods. Once the warped cepstrum $\tilde{c}_t(n)$ is estimated, the inverse warping is done with eq. (8-9) and setting λ to $-\gamma\lambda$.
At this processing step voice conversion can be done quite easily by setting $\gamma \neq 1$. Values between $0 < \gamma < 1.0$ are related to shifting formants towards lower frequencies compared to the original formant frequencies. This changes voice characteristics to sound more like a male voice, while $\lambda > 1.0$ results in more female characteristic. Additionally, the excitation signal of voiced segments has to be adjusted to match the typical fundamental frequency of the new voice.

The excitation signal of voiced sections differs from the one of unvoiced sections mainly by the phases ϕ_i of the i-th frequency f_i. Setting $\phi_i = 0$ for all i a pulse train arises, assigning equally distributed random values $0 \le \phi_i \le 2\pi$ produces white noise.
Exciting the minimum phase system with a pulse train corresponding to the target F_0 contour results in short responses depending on the number of cepstral coefficients N_0. The energy is concentrated at the beginning of the synthesized period and the speech signal sounds distinct voiced.
The level of voicing could be seen as the randomness of phases ϕ_i between the pitch periods. Enhancing the excitation signal at voiced parts with noise at higher frequencies reduces the gravelly of the voice. For each voiced period of the excitation signal the enhancement is applied to the phases ϕ_i:

$$\phi_i = \begin{cases} 0 & i = 0 \\ \mathrm{rand}(r(i)) & 0 < i \le \frac{I}{2} \\ -\phi_{I-i} & \frac{I}{2} < i < I \end{cases} \tag{15}$$

$$r(i) = \frac{2\pi}{1 + e^{-\alpha\left(\frac{4(i-\beta)}{I} - 1\right)}}, \quad 0 < i \le \frac{I}{2}, \tag{16}$$

where I is the order of the inverse Discrete Fourier Transformation (equals to the length of the excitation period) and rand generates uniform distributed random values in the range $[0, r(i))$. Examples of $r(i)$ with different values of α and β are plotted in figure 2.
With equation (15) and (16) the excitation signal $e(n)$ is calcu-

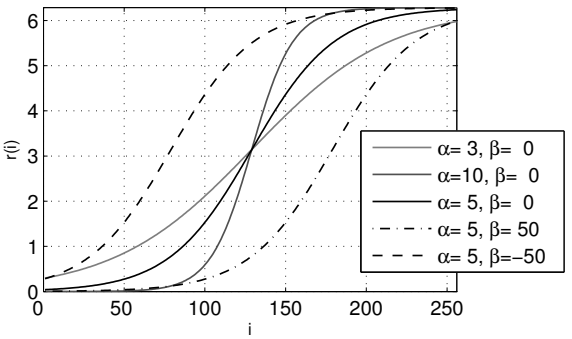

Figure 2: Function $r(i)$, which gives the range of the uniform distributed random values ϕ_i.

lated as:

$$e(i) = F^{-1}\left\{e^{j\phi_i}\right\}, \quad 0 \le i \le I = N_p, \tag{17}$$

with N_p the length of excitation period given from the target F_0 contour. Voiceless periods are generated in the same way, setting α to a high and β to a negative value.
For reducing calculation time I is set to a multiple of the power of 2 and using the inverse Fast Fourier Transformation (FFT). In this case the excitation has to be adjusted (truncated or zero-filled) to fit the target period length N_p. Figure 3 shows the spectrogram of an example of an excitation signal.

3.3. Results

For the evaluation of the acoustic synthesis a preliminary listening test was performed. To avoid the influence of the other synthesis modules (text processing, prosody generation) resynthesis (vocoding) of three English sentences was used for the MOS test. The 25 listeners had to judge 12 sentences in total (three sentences at four compression rates). For the resynthesis the three sentences, spoken of a female US-English speaker, were manual labeled and automatic pitch marked. For each sentence a pseudo inventory were generated at four different compression rates (see table 2).

4. Conclusion

We applied a cepstral filter to the acoustic synthesis, which is embedded in a low resource TTS system. The special coding

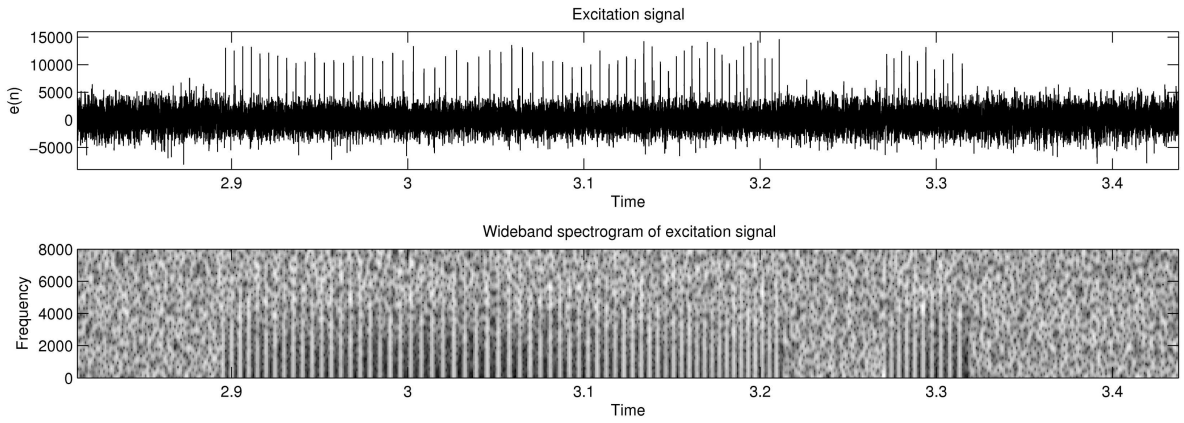

Figure 3: Section of an excitation signal used for cepstral synthesis with corresponding spectrogram.

Sentence		1	2	3	MOS
Size of sentence in kByte (32 kHz, 16 bit, PCM)		258	241	271	
size of pseudo inventory in kByte	($N_0 = 11$, $M = 5$)	5.2	5.4	7.5	1.8
	($N_0 = 16$, $M = 5$)	6.7	7.0	9.9	2.5
	($N_0 = 21$, $M = 5$)	8.3	8.7	12.3	2.6
	($N_0 = 41$, $M = 5$)	14.5	15.3	21.8	3.3

Table 2: Listing of re-synthesis experiments with inventories at different compression rates with corresponding Mean Opinion Scores.

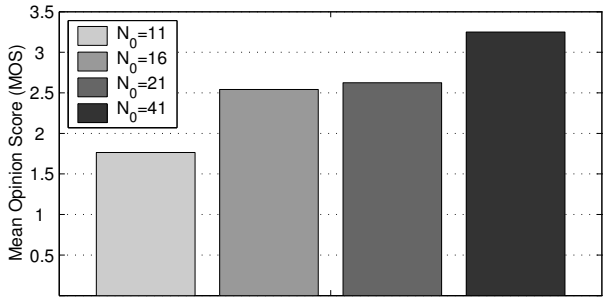

Figure 4: Results of the MOS listening test. Three US-English sentences where re-synthesized with four inventories at different compression rates (see table 2).

technique enables us to achieve very low bit rates at higher bandwidths. We improved the excitation of the cepstral filter to increase voice quality. Furthermore, the system can easily change voice characteristics of synthesized speech without additional computational costs.

5. References

[1] N. Campbell, "Prosody and the selection of source units for concatenative synthesis," in *Proc. ESCA-Workshop on Speech Synthesis, Mohonk (NY)*, 1994, pp. 61–64.

[2] Guntram Strecha, "Neue Ansätze zur Sprachsynthese mit kodierten Sprachsegmenten," in *Proc. 15. Konferenz Elektronische Sprachsignalverarbeitung, ESSV*, Cottbus, 2004, pp. 156–162.

[3] Guntram Strecha, Oliver Jokisch, and Rüdiger Hoffmann, "A Resource-Saving Modification of TD-PSOLA," in *Proc. Advances in Speech Technology, AST*, Maribor, 2003, pp. 151–155.

[4] R. Hoffmann, O. Jokisch, D. Hirschfeld, G. Strecha, H. Kruschke, and U. Kordon, "A multilingual TTS system with less than 1 megabyte footprint for embedded applications," in *Proc. ICASSP*, Hong Kong, 2003.

[5] A. V. McCree and T. P. Barnwell III, "A Mixed Excitation LPC Vocoder Model for Low Bit Rate Speech Coding," in *IEEE Trans. Speech and Audio Processing*, July 1995, pp. 242–250.

[6] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Mixed Excitation for HMM-based Speech Synthesis," in *Proc. EUROSPEECH*, Sep. 2001, vol. 3, pp. 2263–2266.

[7] H. Helmholtz, *Die Lehre von den Tonempfindungen als Physiologische Grundlage für die Theorie der Musik*, Friedrich Vieweg und Sohn, 1863.

[8] Manfred R. Schroeder, "Models of hearing," *Proceedings of the IEEE*, vol. 63, no. 9, pp. 1332–1350, Sept. 1975.

[9] Roy D. Patterson, "A pulse ribbon model of monaural phase perception," *J. Acoust. Soc. Amer.*, vol. 82, no. 5, pp. 1560–1586, Nov. 1987.

[10] Robert Vích, Jiří Přibil, and Zdeněk Smékal, "New cepstral zero-pole vocal tract models for TTS synthesis," in *EUROCON'2001, Trends in Communications, International Conference on*, Bratislava, July 2001, vol. 2, pp. 459–462.

[11] Jr. Baker, G. A. and P. Graves-Morris, *Pad Approximants*, Cambridge U.P.n, 1996.

[12] J. Smith and J. Abel, "Bark and ERB bilinear transforms," *IEEE Transactions on Speech and Audio Processing*, 1999.