ENHNCEMENT INFLUENCE OF THE APERIODICITY COEFFICIENTS IN SPEECH SYNTHESIS

Marius Adrian COTESCU[1], Inge GAVĂŢ[2]

Lucrarea prezintă un studiu asupra îmbunătățirii calității vorbirii sintetice parametrice folosind coeficienţii de aperiodicitate extraşi prin metoda STRAIGHT de analiză a vorbirii. În acest scop au fost construite trei voci sintetice pentru limba engleză, folosind corpusul de sinteză ARCTIC_SLT şi setul de programe HTS, exemplificând trei modalități de abordare a generării secvențelor de coeficienți de aperiodicitate, unul clasic şi două propuse de autori. Cele trei voci au fost evaluate de un lot de ascultători în vederea comparării naturaleței şi similarității cu o voce naturală.

This paper presents an attempt to enhance the naturalness of synthetic parametric voices using the aperiodicity coefficients extracted by STRAIGHT analysis. Three synthetic English voices were built from the ARCTIC_SLT database using the HTS toolkit, with different approaches in treating the aperiodicity coefficients, one classical and two proposed by the authors. The three voices were evaluated by a panel of ten non-native English speakers to compare the naturalness and similarity to a natural voice.

Keywords: speech synthesis, text-to-speech, STRAIGHT analysis

1. Introduction

Speech synthesis is offering an important communication channel between the machine and its human user. It allows the user to receive messages on multiple levels and to process them in a parallel manner. This is very useful in environments where the user has to focus on visually or physically intensive tasks, but information should still be transmitted to him. Moreover, together with speech recognition, it enables the deployment of speech driven interfaces, which can provide a natural means for a user to exchange information with a machine. This can be extremely useful for the blind and visually challenged people, but it can also be implemented for automated public information systems.

There are two main techniques used by machines to render speech. One is based on selecting fragments of speech recorded from a human speaker and

[1] Eng., Depart. of Applied Electronics and Information Engineering, University POLITEHNICA of Bucharest, Romania. e-mail: mcotescu@lpsv.pub.ro

[2] Prof., Depart. of Applied Electronics and Information Engineering, University POLITEHNICA of Bucharest, Romania

stitching them to form a new message. This method is called *concatenative* speech synthesis, and it produces very natural sounding voices, especially for systems that only have to render a limited vocabulary. The other technique used to produce synthetic speech relies on using sets of *parameters* extracted from recorded utterances to reconstruct new waveforms corresponding to the desired message. This second method had the tendency to produce less natural speech, often characterized as “robotic” or “metallic”. Recent developments in speech processing, such as the STRAIGHT technique [1] and trajectory hidden Markov models (HMM) [2] have almost eliminated these problems, such that the current output of parametric synthesizers is very close to natural speech. There is still no synthetic voice that could pass as natural, though.

The current synthetic voices are lacking in two aspects: the prosody model and the similarity to the original speaker. Current generated prosody relies on HMMs to render the pitch contour and sound durations. The combination of statistics and a limited alphabet used to described the phrasing, leads to a repetitive and monotonous prosody, which, although close to the model in short phrases, sounds unnatural and tiring when used for longer texts or phrases. Our concern is, however, the study of the particularities of the natural speaker and their integration with the synthetic voice for enhanced naturalness. In this work we will focus on the characteristics of the excitation generator for voiced sounds. It is already known that the aperiodicity coefficients extracted by STRAIGHT analysis are a good way to capture some of the excitation source characteristics [3], and they have been previously used to enhance the naturalness of parametric voices [4] [5] [6]. We will show that natural extracted sequences of aperiodicity coefficients can be used to enhance synthetic voices.

In this paper we are going to present a study on the effect of the aperiodicity coefficients extracted by STRAIGHT analysis on the naturalness of synthetic parametric speech. The following section describes the tools and techniques used to built the synthetic voice. The third section presents our experiments and results, while the final section draws the conclusions and shows some possible future work.

2. Speech Processing and Synthesis System

The speech analysis stage involved in speech synthesis aims to extract accurate features from natural recordings that are then used to build precise models for the synthesis module. All synthesis modules use a source-filter model [7] to synthesize speech. In order for the synthesis module to be able to produce a large variety of sounds and to take full advantage of the ability to modify parameters without affecting the quality of speech, the analysis techniques should be able to separate as much as possible the source signal from the vocal tract filter. In the same time, the coding method should provide a compact representation of the analysis frame's speech content keeping as much of the initial information.

After analyzing the recordings the extracted parameters were used to train HMMs using the HMM-based Speech Synthesis System (HTS) toolkit. Details of the process are described further.

2.1. STRAIGHT analysis

One analysis method that produces accurate separation of the excitation signal from the vocal tract filter is STRAIGHT. It relies on treating the speech signal's spectrogram as a continuous surface sampled by the windowing process in the time domain and by the harmonic structure of the excitation source in the frequency domain. The STRAIGHT analysis extracts three components: the fundamental frequency, the power spectrum of the vocal tract filter, and an aperiodicity coefficient.

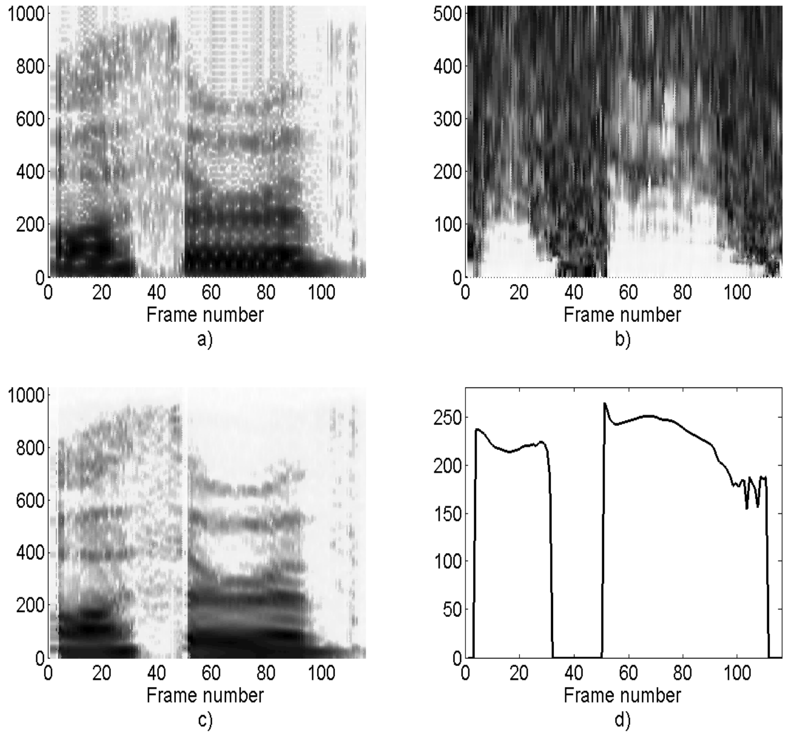
The method considers that a periodic signal $s(t) = s(t+n\tau_0)$, with a fundamental period τ_0, is thought to provide information of the surface for every τ_0 in the time domain and every $f_0 = 1/\tau_0$ in the frequency domain. The goal of the analysis is to recover the surface $S(\omega, t)$ using this partial information.

However, speech is not purely periodic, nor stable. Other errors are introduced by the estimation process of the fundamental frequency. All these aspects must be taken into account by the algorithm, so using the following representation is more dependable in modeling the non-stationary repetitive nature of speech waveforms

$$S(\omega,t) = \sum_{k=N} \alpha_k(t) \cdot \sin\left(\int_{t_0}^{t} k \cdot \left(\omega(\tau) + \omega_k(\tau)\right) \cdot d\tau + \Phi_k \right), \tag{1}$$

where $\alpha_k(t)$ represents the time varying amplitude of the k-th harmonic component, $\omega_k(\tau)$ represents the time varying fundamental frequency of the k-th component, and Φ_k represents the initial phase at t_0. The equation implies that the speech signal is an almost harmonic sum of sinusoids frequency modulated by $\omega_k(\tau)$) parameters and amplitude modulated by $\alpha_k(t)$) parameters. It is well known that the partial information provided by the $\alpha_k(t)$ parameters can be used to reconstruct the surface $S(\omega,t)$, representing the vocal tract transfer function.

Apart from extracting a spectrogram that is free of interferences from the signals periodicity, the STRAIGHT analysis module extracts two more important features of the spoken signal: a very precise pitch contour (using the TEMPO method), and the aperiodicity coefficients, which are computed by dividing the spectrum obtained by the interpolation of the peaks to the spectrum obtained by the interpolation of the valleys of the original power spectrum [3]. The aperiodicity coefficients are closely correlated to the bandwidth of each harmonic component, and so to the frequency modulation factor ω_k in Equation 1.



a)

b)

c)

d)

Fig.1. Examples of normal spectrogram (a), aperiodicity coefficients (b), STRAIGHT spectrogram (c), and pitch contour extracted with TEMPO (d) for the word \author\

Figure 1 shows an example of a FFT extracted spectrogram (a) compared with a spectrogram extracted using STRAIGHT analysis (c). The spectrogram's frequency resolution is 7.8 Hz. In the normal spectrogram, the line pattern given by the harmonic components of the fundamental frequency is clearly visible. The STRAIGHT spectrogram however does not show any traces of the harmonics of the excitation signal, proving the better separation between source and filter in the STRAIGHT analysis.

In addition are represented the pitch contour (d) and the aperiodicity coefficients (b). The aperiodicity factor is smaller at low frequencies, and rises with the frequency for voiced sounds. Unvoiced sounds tend to have constant aperiodicity coefficients over the entire frequency range.

2.2. Vocal tract parameterization

The spectrogram obtained with STRAIGHT can be used by classic analysis methods to extract accurate parametric representations of the vocal tract filter. There are two methods generally used in speech analysis and coding: the linear prediction method [8], and the cepstral method [9]. The linear prediction method represents the signal's spectrum using the all-pole model, which prevents any representation using the zeros of the spectrum. On the other hand, the exponential-type transfer function obtained by the cepstral method has difficulties in reconstructing the sharp peaks of the spectrum. One approach to solving this problem is the generalized cepstral method [10]. It proposes a method of unifying the cepstral method and linear prediction, allowing for the spectrum model to be varied continuously from the all-pole type to the exponential type. The speech spectrum *H(e$^{j\omega}$)*, is modeled as follows, using the mel-generalized spectrum *c(m)*:

$$H(z)=\begin{cases}\left(1+\gamma\sum_{m=0}^{M}c(m)\cdot\tilde{z}^{-m}\right)^{1/\gamma}, & -1\le\gamma<0\\ \exp\left(\sum_{m=0}^{M}c(m)\cdot\tilde{z}^{-m}\right), & \gamma=0\end{cases} \tag{2}$$

where z^{-1} is given by an all-pass function as

$$\tilde{z}^{-1}=\frac{z^{-1}-\alpha}{1-\alpha\cdot z^{-1}},\qquad |\alpha|<1. \tag{3}$$

For our experiment, we used a value for γ of 0, corresponding to the normal cepstral analysis. A value of 0.42 was chosen for α, which for the 16 kHz sampling rate used in the recordings, approximates well the mel frequency scale. The vocal tract models were trained using 39 mel-cepstral coefficients extracted using the above α and γ values.

2.3. Excitation source parameterization

Two of the three components extracted by STRAIGHT analysis contain information regarding the excitation source: the pitch contour, and the aperiodicity coefficients. The pitch contour dictates the intonation of the utterance, while the aperiodicity coefficients contain information about the particularities of the excitation generator.

The intonation model was trained using the logarithm of the extracted fundamental frequency values. The source's particularities were modeled by the mean value of the aperiodicity coefficient over 5 sub-bands: 0 – 1 kHz, 1 – 2 kHz, 2 – 4 kHz, 4 – 6 kHz, and 6 – 8 kHz. A set of aperiodicity coefficients extracted from a random analysis window containing voiced speech and the sequence of aperiodicity coefficients corresponding to the longest stretch of voiced sounds were saved for use in the synthesis stage.

2.4. Text-to-speech system

We have used the HMM–based Speech Synthesis System (HTS) toolkit developed by the HTS working group at the Nagoya Institute of Technology, in Japan. It is based on the popular Hidden Markov Models Toolkit (HTK) developed by the Cambridge University Engineering Department as a portable toolkit for building and manipulating hidden Markov models. HTS extends the capabilities of HTK to build and train HMMs, to be able to generate observation sequences using the trained models that can be used to synthesize speech waveforms.

Hidden Markov Models (HMMs) had been used for speech recognition and synthesis for a long time, producing very good results. However, in speech synthesis, the classic HMM model is unable to generate smooth observation sequences using just the mean values corresponding to each state. A series of articles by Tokuda and colleagues [2] [11] [12] presents a method to produce maximum likelihood observations that took the natural dynamics of speech into account. In [11], Tokuda et al presents a system which uses delta coefficients as a constraint on what observations can be generated. This can be easily extended to cases which use acceleration and higher order constraints also. Further description of the core techniques used by the HTS toolkit can be found in [13] [14] [15].

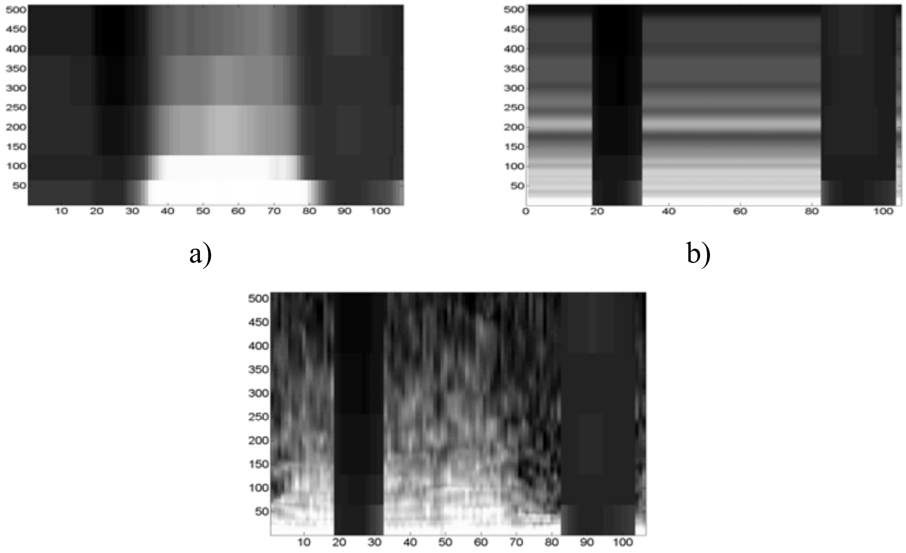
3. Experiment and Results

Our work focused on evaluating effects on synthetic voices of different strategies to built sequences of aperiodicity coefficients extracted by STRAIGHT analysis. We know about the aperiodicity coefficient that it is a characteristic of the excitation source [3]. The experiment aimed to show the existence of aperiodicity coefficients sequences that can be used to enhance the quality of synthetic parametric voices.

In order to prove our hypothesis, we have trained a voice using the HTS toolkit and the ARCTIC_SLT speech synthesis database. Using the trained models, we have generated the pitch contour, cepstral coefficients, and aperiodicity coefficients for the five bands using the HTS toolkit. We then synthesized three sets of utterances: one set in the classical manner using all the generated parameters, the second using the aperiodicity coefficients extracted from one random frame of voiced speech for all the voiced synthetic sounds with the generated pitch contour and cepstral coefficients and a third one using the aperiodicity coefficients extracted from the longest continuous stretch of voiced sounds in the training database.

In Fig.2 we present the aperiodicity coefficients obtained in the three variants for the word “Alice”, and it is visible that the proposed two strategies ensure a larger variety of the aperiodicity coefficients than the classical method, so that an enhancement of the obtained synthetic voices is to be expected.

To evaluate the voices, a set of ten non-native English speakers, with no or little speech processing experience, were asked to rate the generated speech utterances.



a) b)

c)

Fig.2. Examples of generated aperiodicity coefficients for the word \Alice\: a) HTS generated sub-bands, b) using a set from a random window, c) using the longest stretch of voiced sound.

Table 1

Scores meaning for naturalness and similarity tests

Score	Meaning for the similarity test	Meaning for the naturalness test
1	*Sounds like a totally different person*	*Completely unnatural*
2		*Mostly Unnatural*
3	...	*Equally natural and unnatural*
4	...	*Mostly natural*
5	... *Sounds like exactly the same person*	*Completely natural*

The tests were divided in two sections: one aiming at measuring the similarity of the synthetic voices to the original speaker, and one measuring the perceived naturalness of the generated voices. The score meaning for each test are given in Table 1 on a five step scale. Each section was split into nine parts.

The similarity test section provides the person with four reference samples of the original voice and one other new sample for every part of the section. For each part, the person is asked to rate how similar the voice in the new example sounds to the voice in the 4 reference samples.

In the naturalness section, the person can listen to one sample of a voice (natural or synthetic) at a time, and he is asked how natural or unnatural the sentence sounds. The scores given by each person for every phrase generated by the three methods were compared to see which method was best appreciated.

Figure 3 shows the distribution of the winning method over all samples and participants. In the similarity test, voices generated in classical manner were preferred in 33% of the cases, the ones generated using the random sample in 30% of the cases, and the ones generated using the longest stretch in 37% of the cases. In the naturalness test, the HTS generated phrases were preferred in 30% of the cases, the ones generated using the single random sample were preferred in 34% of the cases, while the phrases generated the longest stretch of aperiodicity coefficients were preferred in 36% of the cases.

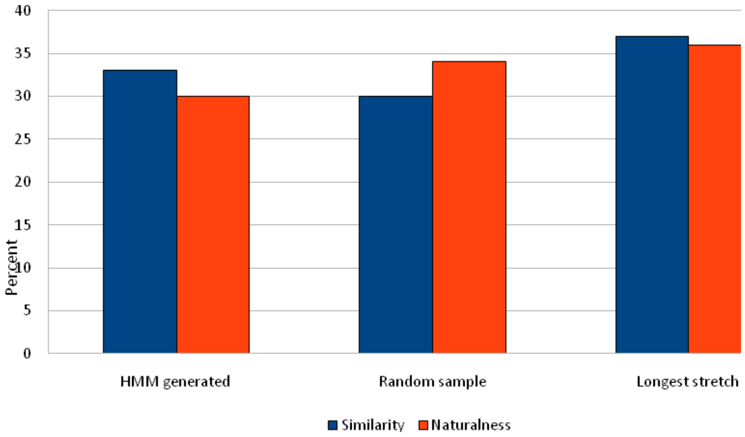


Fig. 3. Results of the similarity and naturalness tests

4. Conclusions

We built an English synthetic voice using the HTS toolkit and the ARCTIC_SLT database which we than modified to use samples of aperiodicity coefficients extracted by STRAIGHT from real speech. The initial voice was constructed using 39 cepstral coefficients extracted from STRAIGHT smoothed spectrograms to model the vocal tract, the mean value of the aperiodicity coefficients over five sub-bands and the logarithm of the fundamental frequency to model the excitation source. We have then modified it to use either a single random frame of aperiodicity coefficients extracted from real voiced speech, or the aperiodicity coefficients sequence extracted from the longest voiced stretch in the training database.

The original and modified synthetic voices were tested by ten non-native English speakers. The results showed that the random sample method has some advantages in naturalness, but not in similarity; the longest stretch method performs better, having an advantage in both the similarity and naturalness tests. Although not highly decisive, the results show that samples of aperiodicity coefficients can be used to enhance the quality of synthetic speech.

Better results we expect to obtain in the next step, by using samples of extracted aperiodicity coefficients for each phonetic unit, or applying one of the available unit selection synthesis algorithms [16] to generate the aperiodicity coefficients sequence.

Acknowledgement

This work was supported by the POSDRU/6/1.5/S/16 grant.

REFERENCES

[1] *Hideki Kawahara, Ikuyo Masuda-Katsuse, Alain de Cheveigne*, Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. Speech Communication, 27, pp.187-207, 1999

[2] *K. Tokuda, T. Kobayashi, S. Imai,* Speech parameter generation from HMM using dynamic features, in Proceedings of the International Conference on Acoustics Speech and Signal Processing 1995.

[3] *H. Kawahara, Jo Estill, O. Fujimura*, Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT, MAVEBA 2001, Sept.13-15, Firentze Italy, 2001.

[4] *V. Karaiskos, S. King, R.A.J. Clark, Catherine Mayo*, The blizzard challenge 2008, in Proc. Blizzard Challenge Workshop, Brisbane, Australia, September 2008

[5] *J.S. Andersson, J.P. Cabral, L. Badino, J. Yamagishi, R.A.J. Clark*, Glottal source and prosodic prominence modelling in HMM-based speech synthesis for the Blizzard Challenge 2009. In The Blizzard Challenge 2009, Edinburgh, U.K., September 2009

[6] *Junichi Yamagishi, Heiga Zen, Yi-Jian Wu, Tomoki Toda, Keiichi Tokuda*, The HTS-2008 system: Yet another evaluation of the speaker-adaptive HMM-based speech synthesis system in the 2008 Blizzard Challenge. In Proc. Blizzard Challenge 2008, Brisbane, Australia, September 2008

[7] *G. Fant*, Acoustic analysis and synthesis of speech with applications to Swedish. Ericsson Technics, 1–1959, pp 1–106.

[8] *F. Itakura, S. Saito*, Estimation of speech spectrum density and formant frequency by statistical method. Trans IEICE 1970; J53-A; pp 35-42

[9] A.*V. Oppenheim, R.W. Schafer*, Digital signal processing. Prentice-Hall; 1975

[10] *K. Tokuda, T. Kobayashi, R. Yamamoto, S. Imai*, Speech spectrum estimation with generalized cepstrum as parameters. Trans IEICE 1989, pp 1071-1076.

[11] *K. Tokuda, T. Masuko, T. Yamada*, An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features. In Proceedings of Eurospeech 1995

[12] *K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura*, Speech parameter generation algorithms for HMM-based speech synthesis. In Proceedings International Conference on Acoustics Speech and Signal Processing 2000

[13] *T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura*, Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis, Proc. of Eurospeech, pp.2347-2350, Sept. 1999

[14] *K. Tokuda, T. Mausko, N. Miyazaki, T. Kobayashi*, Multi-space probability distribution HMM, IEICE Trans. Inf. & Syst., vol. E85-D, no.3, pp.455-464, March 2002.

[15] *A.W. Black, H. Zen, K. Tokuda*, Statistical parametric speech synthesis, Proc. of ICASSP, pp.1229-1232, Apr. 2007.

[16] *R.A.J. Clark, K. Richmond, S. King*, Multisyn: Open-domain unit selection for the Festival speech synthesis system. Speech Communication, 49(4):317-330, 2007