PROSODIC MODIFICATIONS OF SYNTHETIC SPEECH

Igor SZÖKE, Master Degree Programme (5)
Dept. of Intelligent Systems, FIT, BUT
E-mail: xszoke00@stud.fit.vutbr.cz

Supervised by: Ing. Filip Orság

ABSTRACT

This article presents a method for speech prosodic modification. Parametric harmonic and noise model (HNM) is used to describe a speech signal. The algorithm for prosodic modifications consists of two parts. The analysis part parameterizes the speech signal. We spread speech into a sequence of the pitch-marks, each of which is represented by a vector of parameters. The next part is synthesis. Prosodic modifications are done by remapping of the pitch-marks. Modified signal can be resynthesized from interpolated parameters.

1 INTRODUCTION

Prosodic modifications are needed for a high-quality speech synthesis. PSOLA algorithm is widely used, but there are some disadvantages in it. We tried to create a parametric model of speech signal to overcome this disadvantages. Our goal is to find acceptable parametric description of speech, which is useful for prosodic changes without decreasing of the quality of the output. Prosody is a part of the speech acoustic. It consists of intonation (changing of the fundamental frequency), rhythm (changing of speed) and stress (changing of volume). OLA based methods are the most frequently used methods for prosody modifications. OLA means OverLap and Add, e.g. PSOLA (Pitch Synchronous OLA) is one type of OLA. These methods are simple and efficient, but they are problematic in case of modifications where substantial pitch/duration changes are requested and in unvoiced parts of speech (e.g. fricatives)[1].

2 SPEECH SIGNAL MODEL

Good model is needed for a good parametric description of speech. This model must be quite simple and must describe all important speech features. That is why we are inspired by the theory of human speech production. One suitable theory is presented in [2]

and it is called Formant and Resonance theory. This theory says, that speech has two components. One of them is a **noise component**, the other one is a **harmonic component**. The harmonic component originates from the resonances of fundamental frequency in cavities (i.e. mouth, nose and throat). The noise component originates from the air friction against obstacles. These components have different characteristics, so it is better to split them and process them separately (in PSOLA, they are processed together).

A model which can describe both components is called **Harmonic and Noise Model** (HNM) presented by Ioannis Stylianou in [1]. In the frequency domain the voiced parts of speech can be split into harmonic and noise components. The harmonic component $h(t)$ is represented by the lower frequencies, and the noise $n(t)$ by the higher ones. A separating frequency $F_m(t)$, which separate both components is called *maximum voiced frequency*. The unvoiced parts of speech consists only from noise component. Speech signal is equal to sum of the harmonic and noise components:

$$s(t) = h(t) + n(t). \tag{1}$$

Harmonic component $h(t)$ consists of the harmonic multiplications of the fundamental frequency. Coding is done by using a sinusoidal model. The harmonic component could be defined as a sum of sinusoids with appropriate amplitudes and phases:

$$h(t) = \sum_{k=1}^{K(t)} a_k(t) \cos \phi_k(t), \tag{2}$$

where $a_k(t)$ and $\phi_k(t)$ is an amplitude and phase of k-th harmonic component in time t. $K(t) = F_m(t)/F_0(t)$ is a time function representing a number of harmonics in harmonics component of speech. $F_0(t)$ denotes the fundamental frequency. Phase is defined as a harmonic frequency varying in time:

$$\phi_k(t) = 2\pi k f_0(t), \tag{3}$$

where $f_0(t) = F_0/F_s$ is normalized fundamental frequency, and F_s is a sampling rate.

Noise component $n(t)$ can be described in two different ways. The first of them is coding of a spectral envelope using AR filter. Synthesis is done by filtering of a white noise using this filter. The second way is to use the same method as in case of the harmonic component. Noise component has no fundamental frequency because it is unvoiced. So we must set F_0 to a value, e.g. $100Hz$. Because the noise is a stochastic signal, the sinusoid phases are set randomly.

3 PROSODY MODIFICATION BASED ON HNM

Our method has two steps. The first one is a speech signal analysis. The second one includes a changing of the prosodic parameters and a speech signal synthesis.

3.1 ANALYSIS

The first parameter which must be estimated is the fundamental frequency F_0. Next, we have to decide which part of speech is voiced and which is unvoiced. Then we must

find analysis pitch-marks t_a^i. Pitch-marks are the beginnings of glottal periods. In the noise parts without any glottal activity we use the fixed frequency of $100Hz$. For the voiced parts we need to find the maximal voiced frequency F_m.

Harmonic component of speech signal is present in voiced part from $F_0(t_a^i)$ to $F_m(t_a^i)$. We must determine amplitudes $a_k(t_a^i)$ and phases $\phi_k(t_a^i)$ of k-th harmonic of $F_0(t_a^i)$ for all pitch-marks t_a^i. k is in $\langle 1, \lceil F_m(t_a^i)/F_0(t_a^i) \rceil \rangle$. Harmonic component is given by:

$$\hat{h}(t) = \sum_{k=-L}^{L} A_k(t_a^i) e^{j2\pi k F_0(t_a^i)(t-t_a^i)}, \tag{4}$$

where L is a number of harmonics in harmonic component, $A_k(t_a^i)$ is complex amplitude of the k-th harmonic with $A_{-k} = A_k^*$ ($*$ denotes a conjugation). A derivation of amplitudes is deduced in [1]. The result is:

$$A_k = \frac{\sum_{t=t_a^i-N}^{t_a^i+N} w^2(t) s(t) e^{-j2\pi k F_0(t_a^i)t}}{\sum_{t=t_a^i-N}^{t_a^i+N} w^2(t)}, \tag{5}$$

where $N = \lceil P(t_a^i) \rceil$, $s(t)$ is original speech signal and $w(t)$ is Hamming window. $P(t_a^i) = F_s/F_0(t_a^i)$ is pitch period.

Noise component of speech signal is present in the voiced parts from $F_m(t_a^i)$ to $F_s/2$ and in the unvoiced parts from $F_0^{'}$ to $F_s/2$. Computation of amplitudes of the noise component, the same method as for harmonic component is used. $F_0^{'} = 100Hz$ and phases are set randomly.

3.2 SYNTHESIS

Synthesis is the second step of our method. At first, we won't change speech prosody. Modification is done by changing the pitch-marks t_a^i into t_s^i (synthetic pitch-marks) and by changing the fundamental frequency $F_0(t_a^i)$ into $F_0^{'}(t_s^i)$. Marks $t_a^i = t_s^i$ and $P(t_a^i) = P'(t_s^i)$ are relevant for synthesis without any modifications. Synthetic speech is sum of synthetic harmonic and noise components $\hat{s}(t) = \hat{h}(t) + \hat{n}(t)$. Synthetic component is equal to sum of all harmonics:

$$\hat{h}(t) = \sum_{k=0}^{L(t_s^i)} a_k(t_s^i) \cos(\phi_k(t_s^i) + k2\pi F_0(t_s^i)t) \quad for \quad t \in \langle 0, N \rangle, \tag{6}$$

where N is synthetic window length, i.e. $N = \langle P'(t_s^i) \rangle = t_s^{i+1} - t_s^i$. We would obtain amplitudes and phases discontinuities if we used only this method, so we have to use an amplitude and phase interpolation.

The heart of the prosodic modifications is a calculation of the synthetic pitch-marks t_s^i and their mapping onto the original pitch-marks t_a^i by using virtual pitch-marks t_v^i. Virtual pitch-marks are $t_v^i = D^{-1}(t_s^i)$, where $D(t)$ is a time warping function:

$$D(t) = \int_0^t \beta(\tau) d\tau, \tag{7}$$

where $\beta(t)$ is a coefficient of the speech rate. Equation for the synthetic pitch-marks is:

$$t_s^{i+1} - t_s^i = \frac{1}{t_v^{i+1} - t_v^i} \int_{t_v^i}^{t_v^{i+1}} \frac{P(t)}{\alpha(t)} dt, \tag{8}$$

where $\alpha(t)$ is a coefficient of the intonation. Each pitch-mark t_s^i is related only to one pitch-mark t_v^i which is between pitch-marks t_a. Last step is to interpolate parameters of t_v^i from two closest t_a^L and t_a^R. From this parameters we compute t_s^i (see Figure 1). After this, we can synthetize modified speech signal from t_s pitch-marks.

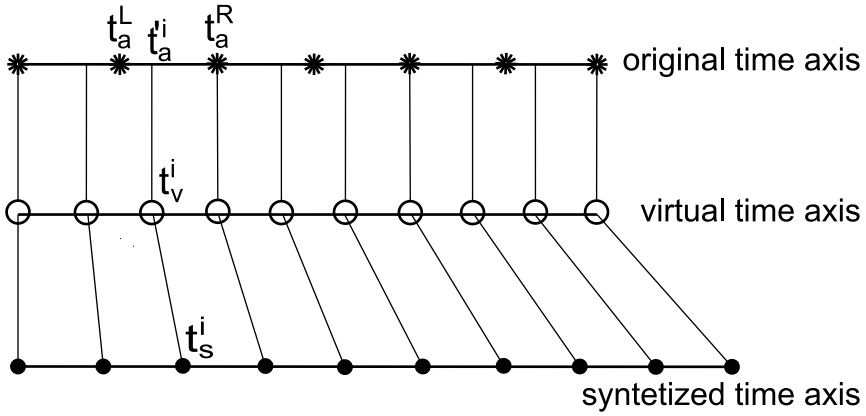

Figure 1: Schema of the pitch and duration prosodic modification. Pitch modification factor is $7/6$ (modified speech has little bit higher fundamental frequency). Duration modification factor is $6/5$ (modified speech is little bit slower). You can see mapping t_s^i onto $t_a^{\prime i}$ which are computed using interpolation between t_a^R and t_a^L.

4 CONCLUSION

The results of experiments showed, that the quality of speech synthesis without the prosody modifications was excellent. Synthesis with modification was very good. There were some errors, which were caused by a not 100% reliable pitch-marks detection. Synthesis with α lower than 0.6 was good, but speech sounded roughly. This is hard to improve given analysis used. As we are not able to compute full impulse response of cavities. Some examples of an original and modified utterances are available on the web [3].

REFERENCES

[1] Stylianou, I.: Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification, [PhD], ENST Paris, 1996

[2] Hála, B., Sovák, M.: Hlas řeč sluch, Praha, SPN 1962, ISBN 16-901-62

[3] Examples: http://igi.darkfuture.cz/speech/EEICT/EEICT.html