Hungarian Speech Synthesis Using a Phase Exact HNM Approach

Kornél Kovács[1], András Kocsor[2], and László Tóth[3]

Research Group on Artificial Intelligence
of the Hungarian Academy of Sciences and University of Szeged
H-6720 Szeged, Aradi vértanúk tere 1., Hungary
{[1]kkornel, [2]kocsor, [3]tothl}@inf.u-szeged.hu
http://www.inf.u-szeged.hu/speech

Abstract. Unnaturally sounding speech prevents the listeners from recognizing the message of the signal. In this paper we demonstrate how a precise initial phase approximation can improve the naturalness of artificially generated speech. Using the Harmonic plus Noise Model provided by Stylianou as a framework for a Hungarian speech synthesis, the exact initial phase extension of the system can be easily performed. The proposed method turns out to be more effective in preserving the sound characteristics and quality than the original one.

1 Introduction

The idea of artificially generated high quality speech signal has been present in science for a long time. We do not intend to review all the relevant literature, but there are some general features which help us to categorize the existing approaches into the following types: the articulatory model, the formant tracking mechanism, and the concatenation method which uses pre-recorded and analyzed natural speech signals to obtain the desired sound.

The Harmonic plus Noise Model is a well-known representative for concatenating speech synthesis. The synthesis part of HNM can generate prosodically modified speech signal using the parameters from the analysis step. The model provided by Stylianou [3] regards a speech signal as a sum of a voiced and an unvoiced noise part with distinct frequency bands, where the lower voiced part can be expressed as a sum of harmonically related sinusoids. The analysis step can determine the uppermost voiced frequency via a peak picking algorithm that is based on the estimation of the pitch period. Because the noise part can be also modelled as a sum of harmonically related sinusoids [3], the analysis part ends with the computation of sinusoid parameters in pitch synchronous time instants. Moreover, in the synthesis step prosodic modifications can be easily executed using this sinusoidal representation.

Using the zero-phase parameter estimation technique proposed by Stylianou we get convincing result. But, based on human listening tests we found that the initial phase of sinusoids have great importance on the naturalness of the speech. Taking into account the initial phase in the HNM framework the resultant method improves the naturalness of the speech signal quite significantly: the finally produced artificial speech sounds more natural than the speech originated from the basically implemented Stylianou system.

2 Harmonic approximation

Firstly, let us assume that the parameters of harmonics and the pitch period are nearly constant for a small time interval. This part of the model approximates the signal by a sum of harmonic sinusoids over a small interval. The signal is known in N time instants $\mathbf{t} = (t_1, \ldots, t_N)^T$ where the signal values are $\mathbf{s} = (s_1, \ldots, s_N)^T$.

The approximation procedure optimizes the amplitudes and phases of the following equation:

$$h(t) = a_0 + \sum_{k=1}^{L} a_k \cos(k\omega t + \psi_k), \tag{1}$$

where the $\mathbf{a}$ and ψ vectors contain the amplitudes and phases of the harmonic sinusoids. The number of harmonics L can be derived from the fundamental frequency and the maximal voiced frequency of the desired time instant.

The optimal parameters have values which minimize the square of the error between the original signal and the approximated one:

$$\epsilon = \sum_{t=t_1}^{t_N} W_{tt}^2 (s_t - h(t))^2, \tag{2}$$

where W is a diagonal matrix with properly chosen weights.

Stylianou makes use of equation (1) supposing that $\psi_k = 0$, which requires solving a set of linear equations when minimizing the error ϵ. To obtain this set of equations we use the vector form of (1) without initial phases:

$$\tilde{h}(t) = \mathbf{b}^T(t)\mathbf{a}, \qquad \mathbf{b}^T(t) = (1, \cos(1\omega t), \ldots, \cos(L\omega t))$$

With this type of harmonic approximation we can redefine equation (2) like so:

$$\tilde{\epsilon} = \sum_{t=t_1}^{t_N} W_{tt}^2 (s_t - \tilde{h}(t))^2 = \|W(\mathbf{s} - B\mathbf{a})\|_2^2, \tag{3}$$

where

$$B^T = (\mathbf{b}(t_1), \ldots, \mathbf{b}(t_N))$$

The error function is expressed by the quadratic form (3), whose minimum defines the amplitudes of the harmonic sinusoids with no initial phase:

$$B^T W^T W B\mathbf{a} = B^T W^T W \mathbf{s} \tag{4}$$

Our approach does not place any restrictions on the form of equation (1) as Stylianou did. Though, the approximation with non-harmonic sinusoids has been solved by Kocsor et al [1] in a locally optimal way, our approach can work out the parameters of harmonic sinusoid approximation in a globally optimal way by using the known angular frequency.

Applying the trigonometrical relation

$$\cos(\alpha + \beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta$$

one can prove that the equation (1) can be re-expressed in vector form:

$$h(t) = \mathbf{g}^T(t)\mathbf{f},$$

where

$$\mathbf{g}^T(t) = (1, \cos(1\omega t), \ldots, \cos(L\omega t), -\sin(1\omega t), \ldots, -\sin(L\omega t))$$
$$\mathbf{f}^T = (a_0, a_1 \cos\psi_1, \ldots, a_L \cos\psi_L, a_1 \sin\psi_1, \ldots, a_L \sin\psi_L)$$

Using this notation:

$$\epsilon = \|W(\mathbf{s} - G\mathbf{f})\|_2^2, \qquad G^T = (\mathbf{g}(t_1), \ldots, \mathbf{g}(t_N))$$

The above equation shows how the error of the initial phase exact harmonic approximation (1) can be expressed in quadratic form with a unique minimum:

$$\mathbf{f} = (G^T W^T W G)^+ (G^T W^T W \mathbf{s}), \tag{5}$$

where $^+$ denotes the Moore&Penrose pseudo-inverse.

After obtaining $\mathbf{f}$, the amplitude and phase of each component can be computed by making use of the simple relations:

$$\psi_k = \arctan \frac{f_{1+L+k}}{f_{1+k}} \qquad a_k = \frac{f_{1+k}}{\cos\psi_k}$$

For the purpose of pitch scaling we need to interpolate the spectrum defined by vector $\mathbf{a}$ with a parametric curve like a cepstrum with real valued parameters. The phase envelope estimation of ψ must be determined as well when the phases have a monotonic character. The cepstrum interpolation with real valued parameters presumes that the interpolated values are non-negative, which can be achieved by using the following:

$$-A\cos(\omega + \psi) = A\cos(\omega + (\psi + (2k+1)\pi)) \qquad k \in \mathbb{Z}$$

3 Experiments

Before dealing with the quality of the synthetized speech we examine the solvability of the equations which provide the parameters of the different approaches. The short time signals are twice the pitch period, so the number of time instants included in the approximation depends on the sampling rate and pitch period. Experiences shows that the set of linear equations (4), and (5), become singular when the short time signal length is less than about 4 times the pitch period. To avoid using inverse, and to ensure that we find the best fitting harmonic approximation we employ the Moore&Penrose pseudo inverse[1] in (4) and (5). This can be used in both cases, because the parameters can be simply computed via a set of linear equations in each case.

[1] The pseudo inverse can be computed by the help of Singular Value Decomposition (SVD) which ensures that the computational cost of the pseudo inverse will be proportional to the rank of the matrix. It then means that the zero-phase and the precise initial phase approaches can generate the amplitudes and phases with about the same computational cost because the ranks of the coefficient matrices are nearly the same in both case.

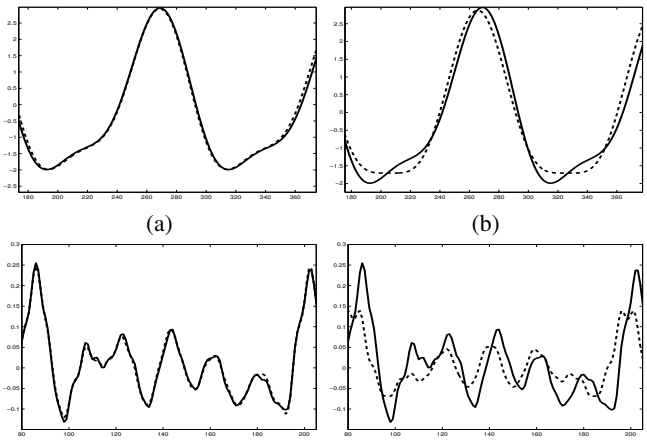

Fig. 1. Short time signals (*solid line*) and their approximations (*dashed line*). Both (a) and (b) display the same artificial harmonic signal and the same part of a Hungarian vowel 'a' is displayed in (c) and (d). Here (a) and (c) show the approximation with precise initial phases, while (b) and (d) show the corresponding zero-phase estimation.

In the artificial signal domain a comparison of the original and the synthetic signal was performed. The same short time frame of an artificial harmonic signal can be seen on Figs. 1 (a) and (b). It obviously seems that the approximation with precise initial phase describes the original signal much more accurately than the zero-phase version does. In the human speech domain the quality of the various synthesis models has been judged by informal listening. The series of testing done undoubtedly prove that the model with initial phase preserves much more detail of the original speech, which means a more natural and clear artificial signal. This difference appears more strikingly in the case of prosodic modification where the more inaccurate approximation of the zero-phase method leads to a metallic sounding signal. In Figs. 1 (c) and (d) we can see an example for a Hungarian vowel 'a' with precise and zero-phase approximation. The implemented models were tested on a segmented Hungarian speech database which makes it possible to have a text-to-speech system. Some speech signals and their prosodically modified versions can be accessed on the Internet (see [2]).

In conclusion, it is clear that the use of exact initial phase approximations is more beneficial for a speech synthesis system as the model is more realistic, and it allows for the possibility of modifying prosodic information.

References

1. KOCSOR, A., TÓTH, L., BÁLINT I.,: *On the Optimal Parameters of a Sinusoidal Representation of Signals,* Acta Cybernetica 14, pp. 315-330, 1999.
2. *Prosodically modified speech data,* http://www.inf.u-szeged.hu/speech
3. STYLIANOU, YANNIS *Harmonic plus Noise Model for Speech, combined with Statistical Methods, for Speech and Speaker Modification,* PhD Thesis, 1996.