PHASE-MISMATCH-FREE AND DATA EFFICIENT APPROACH TO NATURAL SOUNDING HARMONIC CONCATENATIVE SPEECH SYNTHESIS

Zbyněk Tychtl

Department of Cybernetics, University of West Bohemia in Pilsen
Univerzitní 8, 306 14 Pilsen, Czech Republic
email: tychtl@kky.zcu.cz

ABSTRACT

This paper proposes our innovative approach to speech signal representation and generation in the harmonic/noise based speech synthesis. The problem with harmonic/noise synthesis arises when it is required to achieve the high-quality synthesis on the low-resource devices. It is because of the necessity to manipulate a large speech unit databases and unavailability of a method for an efficient *phase data representation*. The other implementations with artificial phases (linear, minimal, zeroed, etc.) produce the synthesized speech with unsatisfactory quality. In proposed approach we use phase data derived from real speech signals to reach natural sounding synthesis. We choose, so-called, *representative phase vectors*, that are only stored to the speech unit database. We reached a dramatic reduction of demands for the database storage space. It corresponds with the rate (number of voiced speech *units*):(number of voiced speech *frames*) stored in the database. Our proposed method also ensures the phase coherence in the synthesized speech signal automatically.

1. INTRODUCTION

For years, we have been developing concatenative TTS speech synthesis system with huge statistically prepared triphone-based speech unit database. For speech signal generation the time-domain concatenative approach is applied. We achieved a very high quality of a synthesized speech signal, but there are still perceivable artifacts well known from time-domain based concatenative speech synthesis systems. We found the sinusoidal coding [1] and harmonic plus noise modeling techniques [2] very promising in our goal of reaching high quality of synthesized speech. Moreover, using a frequency domain modeling it is possible to reach the higher compression rate for speech unit database storage (except the phase components, which is also subject of this paper).

In contrary to time-domain methods the model based methods of synthesis benefit from capability to smooth spectral transitions between concatenated units via interpolation of the model parameters. In our effort to build the high quality high-end speech synthesis system, we also tried model-based approaches for speech signal generation like LPC and residual excited LPC (called RELP). We found all these methods producing number of unacceptable artifacts. On the other hand, the model-based methods are useful for efficient speech unit representation, which is an advantage in a development of the synthesis systems for the embedded devices (handhelds, phones, etc.). In conjunction with our high-quality time-domain system we also tried [3] the approach similar to MBROLA [4] where we off-line re-synthesized speech unit database with the constant preset pitch-frequency and variously preset artificial phases. Regardless of the promising results in case of re-synthesis we observed higher number of disruptive artifacts in final synthesized speech (using so modified unit database) than in speech generated with the original (unmodified) database.

We found the harmonic approach to be capable to produce high quality synthetic speech. The level of synthetic quality reached is strongly constrained by the quality of the speech unit database. We found this method to be quite sensitive to accurate determination of the pitch-frequency and the placement of the phonetic unit boundaries. As we will discuss later, it is necessary to ensure the coherence of the phase components during the synthesis stage. It is not generally easy task. Since this model-based method uses frequency domain representation of speech it is hopeful for future possible extensions of speech modifications and refinements in achieving even higher speech naturalness. In [5] the method based on the estimation of *centres of gravity* of speech signals for the phase mismatches removal is offered. It works in a way of shifting speech waveforms relatively to the centre of gravity. After all, due to our big effort continuously pursued to speech unit database development, we need neither pitch-frequency refinement nor phase correction by signal shifting.

The proposed method offers phase coherent segment concatenation free of known phase mismatches. It also offers the efficient way of speech unit database storage.

2. SPEECH FLUENCY AND PHASE COMPONENTS

In a speech recognition task the phase component in parametric speech modelling is usually omitted. This is not true for task of speech synthesis - at least when the high-quality synthesis is demanded. In various approaches to speech synthesis the phase components are implicitly or explicitly manipulated by different manners.

The only implicit phase manipulation, employed in the common time-domain approaches, is caused by the "pitch-synchronous" processing of the waveforms. It means that signal is processed synchronously with the glottal closure instants, which occur with fundamental (pitch) frequency in the voiced frames. During the signal generation the mentioned instants, called "pitch-marks", drive the process of speech frame concatenation. If the pitch-marks determination is reliable and if in the database the units with requested prosodic parameters (mainly F_0) are available, then such units could be concatenated without phase mismatches. It is, however, difficult to satisfy to have all the speech units with all possible prosodic parameters available in the database even in the case of synthesis system with enough resources accessible. The frames, which will be concatenated, are usually firstly modified to match desired prosodic parameters. If required synthetic F_0 does not match original F_0 of the unit (which usually happens), then the unit is re-processed. It is first cut up to weighted frames of length of two local pitch periods, then, they are mutually shifted to match required synthetic pitch-period and, finally, re-concatenated. Frame weighting reduces amplitude discontinuity. Since the "inner-frame" spectral features stay

unaffected using this approach, the phase incoherence problem arises not only on the concatenation points of the units but also on the inter-frame level. This is usually perceived as artefacts disturbing the notion of speech fluency.

The methods were proposed [4], which off-line process the speech units in the database before its use in time-domain speech synthesis. The voiced frames can be transformed to spectral domain, reharmonized to constant F_0 and transformed back to time-domain. The phase components can also be replaced by some artificial phase (zeroed, minimum, etc.), which finally causes highly unnatural sounding [6]. It still does not prevent the phase mismatch effect to appear, since the units need be modified to have required synthetic F_0 during signal generation.

Using approaches like classical LPC or RELP (residual excited LPC) the phase mismatches are not so apparent. Even though the RELP can contain the original phase component in its residual excitation signal, it also does not eliminate the phase mismatch problem, because it is not clear how to jointly modify both the residual excitation and the model parameters. The pulse train excited LPC system produces artificially sounding speech.

Even in the sinusoidal based approaches like HNM [2] the phase mismatch appearance is not ensured. In [5] the method for the linear phase mismatches removal was proposed. It is based on the estimation of *centres of gravity* of speech signals. It works in a way of shifting speech waveforms relatively to the centre of gravity. So, in fact, it acts like a substitution of a demand of analysing the signals synchronously with glottal closure instants. It eliminates the necessity of determination of the glottal closure instants. Instead of it another time instants (instants of centres of gravity) are determined. Those new instants are used instead of pitch-marks in the same role, to reposition consequent frames. This method can be very useful when glottal closure instants (pitch-marks) cannot be reliably determined (which is not our case; see next chapter).

We have experimentally found that for the best naturalness of synthesized speech it is necessary to use the true phase components (not artificial ones) derived from real speech signals regardless of the synthesis method used. Moreover, to avoid the phase mismatches completely, the original prosodic features of concatenated units would have to match the required synthetic prosodic parameters. This requirement leads to methods building huge speech unit databases where more instants of each speech unit are stored (with various prosodic features). The effective unit selection algorithm is then employed during synthesis to choose the units best fitting required prosody. The approaches with huge unit database are not suitable for implementation on low-resources devices (phones, handhelds, simple one-purpose devices, etc.).

Despite mentioned aspects we stayed aspired to use the harmonic/noise-based approach for its capability of speech modification in the spectral domain. So we had to deal with an impossibility to have a huge unit database at our disposal. Moreover it is hard to compress the true phase components (which we considered to use). It does not apply to amplitudes. The efficient methods for the compression of amplitudes are available [8].

The basic idea of the proposed approach is otherwise to use the true phase components derived from real speech signals, but to store only minor number of them to the unit database. We reached the phase mismatch avoidance by manipulating the phase components by the same way as in the approach with constant artificial phases. One phase component is chosen from the speech unit database and it is then kept constant in all consequent voiced frames in the actually generated voiced part.

3. INITIAL SPEECH UNIT DATABASE

We have got speech corpus recorded professionally with use of electroglottograph for recording the glottal signal. The speaker was asked to try to speak monotonously. The corpus was checked by listeners and disposed of insufficient records. In the glottal signal we successfully detect the glottal closure instants (pitch-marks). So we can reliably determine the local pitch-frequencies as well as we can later rely on the phase-coherency in the consecutive frames during the pitch-synchronous analysis of the speech units.

Then the speech unit database was created from the corpus employing the HMM-based automatic segmentation [7]. As the phonetic units the triphones were chosen. Every unit in the database may be used to represent more phonetic units (triphones), which were clustered on the phonetic/acoustic basis. In spite of the effective tree-based clustering, and in spite of the fact that each unit has only one representation stored (that means for only one prosodic feature), the database may contain (for example) 6258 units. In the initial unit database the units are represented by their waveforms (which is for example 25MB).

4. NEW MODEL-BASED SPEECH UNIT DATABASE CONSTRUCTION

The new speech unit database is built by consequent unit-by-unit analysis of the initial database. Every speech unit represented by the waveform signal of particular triphone is analysed to obtain the set of harmonic/noise parameters, which are then stored to the new database.

4.1 Speech units parameterisation

The speech unit waveform is pitch-synchronously processed. At every pitch-mark the set of the parameters is determined.

The unvoiced segments are processed at a constant rate (which substitutes the pitch-marks) to get the 10^{th} order LPC and a time-energy evolution in the frame (with a resolution of 1 or 2 ms).

The *voiced* segments are firstly pitch-synchronously cut up to the frames of the lengths of two pitch periods. In each frame the *maximal voiced frequency* $F_{\max}$ is determined. The $F_{\max}$ and the local fundamental frequency F_0^k determine the number of the F_0 harmonic frequencies below the $F_{\max}$:

$$L^k = F_{\max} / F_0^k , \tag{1}$$

where k denotes the index of the frame.

The frame is then analysed to obtain the model consisting of the vectors (of the length L^k) of the harmonic amplitudes $\boldsymbol{A}^k$ and the phases $\boldsymbol{\Phi}^k$. Also the set of parameters (LPC and the time-energy evolution), same as in the case of unvoiced frames, is determined. It describes the noise content in the frame. The description of the methods for determination of the parameters was proposed in [2] or alternatively in [1]. We use simpler one, where amplitudes $\boldsymbol{A}^k$ and phases $\boldsymbol{\Phi}^k$ are determined using the short time Fourier transform:

$$A_l^k = \left| \frac{\mathbf{X}^k\left(l\omega_0^k\right)}{\sum_i w(i)} \right| \quad \text{and} \quad \Phi_l^k = \arg\left(\mathbf{X}^k\left(l\omega_0^k\right)\right), \tag{2}$$

where $\mathbf{X}^k$ stands for vector of STFT of the k-th frame in analysed segment, w is the weighting window, l means the l-th component and ω_0 denotes the fundamental frequency.

Let's mention that not all of the parameters, especially the vectors of phases $\boldsymbol{\Phi}^k$, are stored to the new database.

4.2 The representative phase vectors

In our novel method, which is suitable for the low-resource solutions, only the minor number of the phase vectors is stored to the speech unit database. We store only the phase vectors, which we call the *representative phase vectors*. Moreover, such phase vectors are not only copies of the ones obtained by the analysis.

We propose choosing just *one* phase vector for the every uninterrupted sequence of the voiced frames in the unit. It is the best to choose one from the *spectrally stationary* region of the segment. The stationary region in the voiced segments is mostly found right in the middle of the segment. Let this phase vector be a starting version of the *representative phase vector*, which will be denoted $\mathbf{\Phi}_0$.

The $\mathbf{\Phi}_0$ vector is then extended by appending the additional elements to its end. The new elements are obtained by the analysis of the neighbouring frames in the analysed segment. It is performed by the following way. The segment is searched for the frames with the fundamental frequency lower then it was in the frame where the $\mathbf{\Phi}_0$ vector was preset. Analysing the frame with the lower fundamental frequency and/or with the higher $F_{\max}$ we obtain more harmonic elements. Those, which are above the length of the $\mathbf{\Phi}_0$, are appended to the end of the $\mathbf{\Phi}_0$. The purpose of this is to have the representative $\mathbf{\Phi}_0$ containing more elements. It will allow producing a synthetic speech with a "natural" phase with the lower fundamental frequency, which can be required by the prosodic rules during the synthesis.

4.3 Quantification of data storage demands

We propose the method of storing just one phase vector for every uninterrupted voiced sequence of frames (voiced segment) instead of storing the phase vector for every frame.

Performing this way a dramatic reduction of the database storage space demands is reached. It corresponds with the rate

$$\frac{(\text{number of voiced speech } \textit{segments})}{(\text{number of voiced speech } \textit{frames})}$$

stored in the database. Moreover, since we use quite short speech units (triphones), the voiced units do not contain more than one uninterrupted sequence of the voiced frames. So the number of the voiced speech *segments* can be then considered to be comparable to the number of the voiced speech *units*.

To give the quantified example, our database contains 6258 speech units. 5826 of them contain the sequence of the voiced frames. All the voiced segments consist of 95132 frames. Instead of 95132 phase vectors we store to the database just 5826 of their representatives. For this particular database we save 93.9% of the space required for the storage of the phase vectors.

The amplitude vectors can be efficiently quantized using for example the method described in [8].

5. SPEECH SIGNAL GENERATION

During the synthesis the speech signal is generated frame-by-frame, the size of which is one local pitch-period. The spectral parameters of the speech units picked from the speech unit database are modified in the frequency domain according to the required prosodic features. The fundamental frequency is preset every frame. The speech signal is generated summing two components:

$$s = s_h + s_n \,, \tag{3}$$

where s_h denotes the harmonic part and s_n denotes the noise part. When generating the voiced speech the both parts are used, but in the case of unvoiced speech the only s_n applies. The unvoiced parts s_n are generated by LPC filtering of the unit-variance white noise and by weighting the output by the time-domain energy contour. No LPC parameter interpolation is performed. The energy contour is smoothed by a linear weighting applied over the unit concatenations. The voiced part s_h is modelled as sum of harmonics

$$s_h^k(t) = \sum_{l=-L^k}^{L^k} \tilde{A}_l^k(t) e^{j\left(l\omega_0^k(t)+\Phi_{0l}^k\right)t} \,, \tag{4}$$

where k denotes the index of the frame and l is the index of the harmonic component (and of the elements of amplitude and phase vectors).

The amplitudes $\tilde{A}^k$ are obtained by harmonic spectral resampling of the $\boldsymbol{A}^k$ at the multiples of the synthetic (required) fundamental frequency. The signal is generated sample by sample as a sum of sine wave functions. The amplitudes are interpolated over the unit concatenations as well as on the concatenations of the frames.

The phases for the generation of the sine wave components are supplied from the representative phase vector $\mathbf{\Phi}_0$ from the actual unit. At the start of generation of the voiced segment every element in the vector of amplitudes is coupled with the element from the phase representative vector at the same position in the vector. In the successive frames (k) the phase Φ_l^k of each harmonic component (l) (each sine-wave) is copied from preceding synthesized frame using the following rule: If $F_0^k < F_0^{k-1}$ (synthetic F_0 decreases) then, for the Φ_l^k, the phase of the nearest higher (at the higher frequency) component from the preceding frame is used. If $F_0^k > F_0^{k-1}$ (synthetic F_0 increases) then the phase of the nearest lower (at the lower frequency) component from the preceding frame is used. The assignment of the components of the representative phase vector to particular frequencies is not fixed. Generally, the phase component Φ_l^k initially assigned (at the beginning) to the component at a frequency F_i is at the end of the segment assigned to the component at different frequency. This method of the phase assignment offers the efficient solution for reaching the phase-mismatch-free synthesis with the natural phase components. On the Fig. 1 the synthesis of one harmonic component over 3 consecutive frames is illustrated. In each frame the desired synthetic F_0 is changed and the synthetic amplitude is smoothly interpolated. It is easy to see that the amplitude is modified at the sample level and the phase of the component is kept constant over the whole segment.

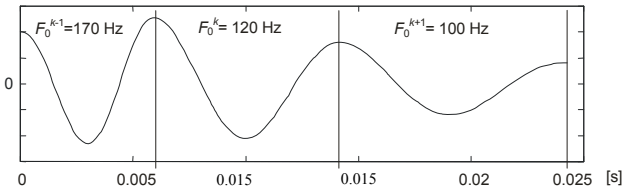

Fig. 1 An illustrative example: One harmonic component in the three consecutive frames with smoothly interpolated amplitudes and substituted constant phases.

All the harmonic components are processed by the described way. Each harmonic component in the 'following' frame is coupled to one component in the actual frame. The amplitudes are smoothly interpolated and the phases are 'adopted' from the actual frame. In other words, the phase component is set at the beginning of the voiced segment, and it is kept constant till the end of the voiced segment.

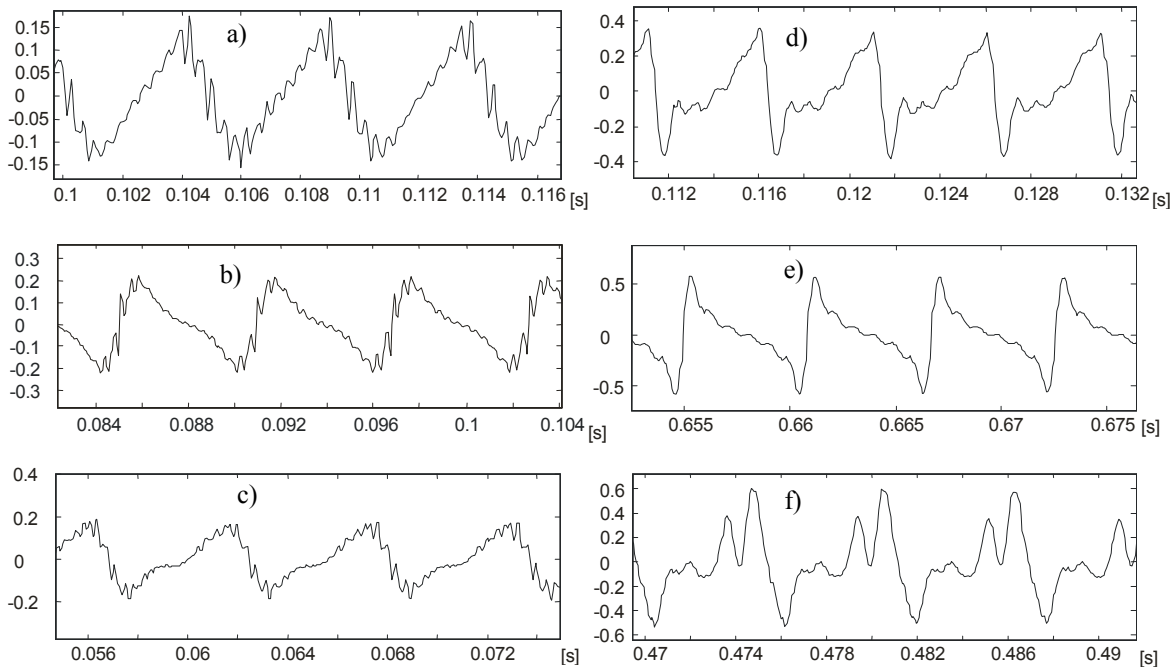

Fig. 2 The a) and d) are the original signals of the phonemes "j" and "o" respectively. The b) and e) are the synthesized signals of the both phonemes with the zeroed phases. The c) and f) are the synthesized signals with the substituted phase components, which were obtained from a different realization of phoneme "j".

This approach ensures the phase coherence in every uninterrupted voiced segment (the sequence of the voiced frames) with the use of the natural phases.

6. CONCLUSION

The Fig. 2 demonstrates the effect caused by the proposed method. It is evident that the shape of the wave is significantly changed when the phase component is substituted. The subjective listening tests confirm that the phase substitution is hardly perceivable especially in the short voiced segments.

The proposed method prevents the appearance of the artifacts caused by phase mismatches and incoherencies in the generated speech signal. The speech naturalness degradation is sometime perceivable in very long voiced sequences. In the shorter voiced segments the listener does not notice the phase interchange.

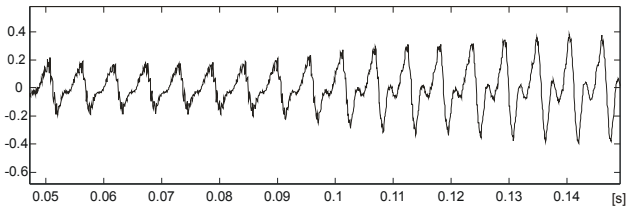

Fig. 3 The part of synthesized voiced segment with the use of phase substitution to preserve the local phase coherence at the concatenation of two different phonemes.

The method also offers very efficient solution, which helps to reduce the storage demands for the phase components by approximately 94%.

7. ACKNOWLEDGEMENT

This research was supported by the Grant Agency of Czech Republic, project No. GAČR 102/02/0124 and by the Ministry of Education of Czech Republic, project No. MSM 235200004.

REFERENCES

[1] R.J. McAulay and F. Quatiery, "Sinusoidal coding," in Speech Coding and Synthesis, W. Kleijn and K. Paliwal, Eds. New York: Marcel Dekker, 1991, ch.4, pp. 165-172.

[2] Y. Stylianou, "Applying the Harmonic Plus Noise Model in Concatenative Speech Synthesis," *IEEE Trans. Speech and Audio Proc.*, 9(1), pp. 21-29, 2001.

[3] Z. Tychtl, K. Matouš, and V. Mareš, "Czech Time-Domain TTS System with Sample-by-Sample Harmonically Pitch-Normalized Speech Segment Database," in *Speech Processing.* 12th Czech – German Workshop, Prague 2002, pp.44-46.

[4] T. Dutoit and H. Leich, "Text-to-speech synthesis based on a MBE re-synthesis of the segments database," *Speech Commun.*, vol 13, pp. 435-440, 1993.

[5] Y. Stylianou, "Removing phase mismatches in concatenative speech synthesis," in *Proc. 3rd ESCA Speech Synthesis Workshop,* Nov. 1998, pp. 267-272.

[6] Z. Tychtl and K. Matouš, "The Phase Substitutions in Czech Harmonic Concatenative Speech Synthesis", TSD 2003, Springer Verlag, LNAI 2807, pp. 333-340, Ceske Budejovice, September 2003.

[7] J. Matoušek and J. Psutka, "ARTIC: A New Czech Text-to-Speech System Using Statistical Approach to Speech Segment Database Construction," in *Proc. 6th Int. Conf. on Spoken Language Processing ICSLP2000*, vol. IV. Beijing, China, 2000, pp. 612-615.

[8] T. Erikson, H. Kang, and Y. Stylianou, "Quantization of the spectral envelope for sinusoidal coders," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* 1998, pp. 37-40.