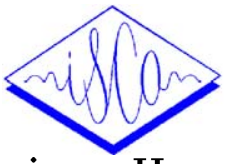

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

Diphone Concatenation using a Harmonic plus Noise Model of Speech

Yannis Stylianou, Thierry Dutoit, and Juergen Schroeter

AT&T Labs-Research
180 Park Ave, PO Box 971
Florham Park, NJ 07932-0971
email: [styliano, dutoit, jsh]@research.att.com

10.21437/Eurospeech.1997-222

ABSTRACT

In this paper we present a high-quality text-to-speech system using diphones. The system is based on a Harmonic plus Noise (HNM) representation of the speech signal. HNM is a pitch-synchronous analysis-synthesis system but does not require pitch marks to be determined as necessary in PSOLA-based methods. HNM assumes the speech signal to be composed of a periodic part and a stochastic part. As a result, different prosody and spectral envelope modification methods can be applied to each part, yielding more natural-sounding synthetic speech. The fully parametric representation of speech using HNM also provides a straightforward way of smoothing diphone boundaries. Informal listening tests, using natural prosody, have shown that the synthetic speech quality is close to the quality of the original sentences, without smoothing problems and without buzziness or other oddities observed with other speech representations used for TTS.

1. INTRODUCTION

Many current Text-To-Speech (TTS) systems are based on diphone or larger unit concatenation. There are various methods of concatenating diphones including TD-PSOLA[7], MBROLA[4], LPC[8], and sinusoidal coders[6].

TD-PSOLA is currently one of the most popular concatenation methods. Although it provides, in general, good quality speech synthesis and much greater naturalness than earlier systems, it suffers from some weaknesses. The principal problem concerns spectral mismatch at segment boundaries. Due to its time-domain, non-parametric representation of the signal, TD-PSOLA offers very limited smoothing possibilities. The preparation of a new database is also a time-consuming task; pitch marking is not a completly automated process and units have to be chosen carefully so as to minimize spectral mismatch during synthesis. MBROLA tries to overcome concatenation problems in the time domain by resynthesizing voiced parts of diphones with constant phase, constant pitch[4] and by linear smoothing between pitch periods at segment boundaries. This unnatural processing of the diphones is the main source of the slight buzziness in the speech produced by MBROLA. Note that this kind of processing can be applied only on voiced frames; discontinuities on unvoiced frames, if any, are not smoothed. The sinusoidal approach[6] is a non-pitch-synchronous approach where both voiced and unvoiced frames are represented by a sum of sinusoids. Similarly, hybrid harmonic/stochastic (H/S) synthesis of Dutoit [3] is non-pitch-synchronous, in which unvoiced components are modeled as a sum of narrow-band noises. In both cases, proper phase concatenation was reported to be critical. Macon performed concatenation by making use of the glottal closure instants, a process which is not always successful[6] while Dutoit found that when propagating the phase differences the resulting quality was not good.

A new model for speech has been proposed[11] [9] based on a Harmonic plus Noise (HNM) representation of speech. HNM has shown the capability of providing high-quality prosodic modifications[11] without the buzziness and tonal quality encountered in previously reported methods. All these properties of HNM had only been tested, to date, on straight copy synthesis of speech sentences. In this study, we present the extension of HNM to diphone concatenation. A previous version of HNM[5] has also been tested at CNET[1] where pitch marks were locked around glottal closure instants. Here we present a different version of HNM[11], where explicit pitch marks are not required.

The fully parametric representation of speech using HNM provides a straightforward way of smoothing diphone boundaries. Around each concatenation point the HNM parameters are smoothed within the region of quasi-stationarity (i.e., for unvoiced frames, the region where the LP gain does not vary too much, and for voiced frames, the region where the maximum voiced frequency remains approximately constant).

Once the number of frames for interpolation has been specified, the set of parameters related to the noise part (reflection coefficients and LP gain) as well as harmonic amplitudes if the frames are voiced, are smoothed using simple linear interpolation (in effect, overwriting a few data frames). Since HNM is not locked on glottal closure instants, phase problems will arise during concatenation. In order to cope with phase mismatches two strategies have been adopted: the first strategy is based on a combination of minimum phase and of the phase of an all-pass filter, while the second uses the original phase with on-the-fly correction of phase offsets.

The first part of the paper is devoted to the description of HNM. It is followed by the description of the diphone

synthesis system based on HNM and the presentation and comparison of the two proposed solutions for the phase problem. Results and comments on the application of HNM for TTS in English and French using female and male speakers are given in the third part of the paper.

2. DESCRIPTION OF HNM

HNM[11] is based on a harmonic plus noise representation of the speech signal. The harmonic part accounts for the quasi-periodic component of the speech signal; the noise part models its non-periodic components, which include friction noise and period-to-period variations of the glottal excitation.

The spectrum is divided into two bands. The time-varying maximum voiced frequency determines the limit between the two bands. In the lower band, the signal is represented solely by harmonically related sinewaves with slowly varying amplitudes, and frequencies.

$$h(t) = \sum_{k=1}^{K(t)} A_k(t) cos(k\theta(t) + \phi_k(t)) \quad (1)$$

with $\theta(t) = \int_{-\infty}^{t} \omega_0(l)dl$. $A_k(t)$ and $\phi_k(t)$ are the amplitude and phase at time t of the k-th harmonic, $\omega_0(t)$ is the fundamental frequency and $K(t)$ is the time-varying number of harmonics included in the harmonic part.
The upper band, which contains the noise part, is modeled by an AR model and it is modulated by a time-domain amplitude envelope. The noise part, $n(t)$, is therefore supposed to have been obtained by filtering a white gaussian noise $b(t)$ by a time-varying, normalized all-pole filter $h(\tau, t)$ and multiplying the result by an energy envelope function $w(t)$:

$$n(t) = w(t)\ [h(\tau, t) \star b(t)] \quad (2)$$

The first step of HNM analysis consists of estimating pitch and maximum voiced frequency based on a time-domain pitch detector[10]. Using the stream of the estimated pitch values, the position of the analysis instants are set at a pitch-synchronous rate (regardless of the exact position of glottal closure). In voiced frames, the amplitudes and phases of the sinusoids composing the harmonic part are estimated by minimizing a weighted time-domain least-squares criterion. This time-domain technique combined with the relatively short duration of the analysis frame in the voiced parts of the signal (two pitch periods) provides a very good match between the estimated harmonic part and the original speech signal. The noise part is modelled by an all-pole filter estimated from 40ms of signal located around the center of the analysis frame.

Synthesis is also performed in a pitch-synchronous way using an overlap and add process. The harmonic part is synthesized directly in the time-domain as a sum of harmonics (Eq.1). The noise part is obtained by filtering a unit-variance white Gaussian noise through a normalized all-pole filter. If the frame is voiced, the noise part is filtered by a high-pass filter with cutoff frequency equal to the maximum voiced frequency. Then, it is modulated by a time-domain envelope synchronized with the pitch period. This modulation of the noise part was shown[5] to be necessary in order to preserve the naturalness of some speech sounds, such as voiced fricatives.

Note that HNM does not use pitch marks locked on glottal closure instants in contrast with other pitch-synchronous methods such as TD-PSOLA; however, the distance between two analysis/synthesis time instants is one local pitch period and the analysis window is two pitch periods long.

Thanks to the pitch-synchronous scheme of HNM, time-scale and pitch-scale modifications are quite straightforward[11].

3. USING HNM FOR TTS

Synthesis using HNM can be divided into an off-line and an on-line process. During the off-line process a diphone segmented speech database is analyzed using the HNM analysis module described in the previous section. A voiced frame is represented by its fundamental frequency, harmonic amplitudes and phases, the number of harmonics included in the harmonic part, reflection coefficients and the LP gain (the last two sets of parameters are for the noise part of voiced frames). Unvoiced frames are represented by reflection coefficients and the LP gain. Therefore, a set of several HNM frames is assigned to each speech unit. If the minimum phase approach is used, the original phases are omitted.

At synthesis time, HNM frames are concatenated and the prosody of units is altered according to the desired prosody. Thanks to the pitch-synchronous scheme of HNM, a simple and flexible technique can be used for that purpose. A mapping between the synthesis and the analysis instants is determined, specifying which analysis instant should be selected for any given synthesis instant[11][2](p.255).

The amplitudes of the new harmonics are then obtained by sampling the spectral envelope defined by the original harmonic amplitudes. If the original phase is used, the phase of the new harmonics are obtained by sampling the phase envelope at the modified pitch-harmonics. Before that, the phase unwrapping technique described in[11] is used which guarantees a phase continuity in the frequency-domain as well as in the time-domain. This step is skipped if the minimum phase approach is used.

After the determination of synthesis instants and the re-estimation of the amplitudes and phases of the new harmonics (in case original phases are used), HNM parameters have to be smoothed around concatenation points (diphone boundaries). Spectral amplitudes, LP gain, and reflection coefficients are linearly interpolated. The number of frames used in the interpolation process depends on the variance of the number of harmonics for the voiced frames and on the variance of the LP gain for the unvoiced frames. The phoneme boundaries inside each diphone define the maximum interpolation range. Finally, there is no interpolation between unvoiced and voiced frames.

If the original phase is used then phase mismatches arise. The phase problem can be split into two sub-problems: phase offset and phase discontinuities. The phase offset is estimated from the maximum of the cross correlation function between two sinusoids which have the same amplitude and frequency but different phases, ϕ_l and ϕ_r, where ϕ_l is the phase of the first harmonic in the last frame of the left diphone and ϕ_r is the phase in the first frame of the right diphone. The resulting phase offset is used to correct the phase of the following diphone (a linear phase is added). This process is repeated (propagated) for all diphone boundaries. Even after this offset correction, a phase discontinuity remains, which is perceived as a background noise (increasing the noise level between harmonics). Using the unwrapped phase of the left and right diphone a simple technique can be applied; the phase difference is calculated and a weighted version of that difference is propagated towards only the following diphone, until the next boundary (last frame of the following diphone). This is somewhat different from what was tested in [3] (where no phase unwrapping and no time-varying decay was applied), which resulted in strong phase distortions after a few diphones had been processed.

When minimum phase is used the above problems do not exist. Since we use a pitch-dependent frame rate, minimum phase is well suited for our purpose. It is estimated on the fly using the amplitudes of the new harmonics. In order to increase speech quality, the phase of an all-pass filter, estimated on the new harmonic frequencies, is added to the minimum phase component. The all-pass filter is a second order filter with poles very close to the unit circle. The coefficients of this filter are held constants during synthesis. A similar technique has recently been proposed in[12] for low bit-rate sinusoidal coding.

4. RESULTS AND DISCUSSION

In this section, we present results obtained by using our HNM-based synthesizer in the context of prosody transplantation (i.e. using natural prosody) and we compare it with other approaches. Our test corpus currently includes one male voice for British English (Roger's voice, freely made available by CSTR, Edinburgh), one male voice for French (the one used for the French voice of MBROLA) and six in-house female voices for American English. All speech signals were sampled at 16kHz and diphone segmentation was performed manually or semi-automatically (and hand-checked). Pitch marking information was also extracted in order to be able to use TD-PSOLA for comparison. The HNM inventory file was built without any test on diphones concatenation costs (i.e., without any attempt to detect and replace bad diphone units).

Prosody was extracted from continuous speech and it was used as input to the HNM and TD-PSOLA diphone-based synthesizers. Informal listening tests using six listeners have shown that HNM exhibits higher quality than TD-PSOLA, by simultaneously providing smoothness and naturalness. In general, HNM performs clearly better than TD-PSOLA, particulary on voiced fricatives, breathy voices and unvoiced frames. Fig.1 presents the spectrograms of two synthetic signals obtained with TD-PSOLA (middle panel) and HNM (lower pannel) and of the original sentence (upper panel) from which the prosody has been extracted.

Fig.1 shows that diphones are concatenated properly by HNM, thereby providing smooth, continuous speech. This is not the case in general with TD-PSOLA, where discontinuities arise at diphones boundaries as a result of the phase and spectral mismatches mentioned in Section 3. Furthermore, during unvoiced and voiced fricative sounds, TD-PSOLA produces metallic quality speech when phoneme durations are increased. In contrast, prosodic modifications conducted by HNM are of high-quality, even for large pitch and duration modification. This is mainly because the harmonic part and the noise part are processed by different techniques.

Formal listening tests using numerous voices are currently being conducted in order to compare HNM and TD-PSOLA for TTS. Preliminary results show the superiority of HNM over TD-PSOLA. The overall quality, intelligibility and naturalness are very high for HNM.

HNM has also been compared with MBROLA[4]. We found that if only the minimum phase (without the all-pass filter phase) approach is used in HNM, the two systems provide almost the same quality. When the original phase is used with HNM, however, HNM-based synthetic speech is more "present" and more natural than that produced by MBROLA and does not exhibit the slight buzziness observed with MBROLA.

As with any other technique based on the hypothesis of quasi-stationary of the signal, however, we found that HNM has problems reproducing stops. A simple solution to this problem is to use the corresponding original waveforms directly for synthesis, as in TD-PSOLA and MBROLA.

The main cost for the enhanced quality of HNM lies in its computational complexity at synthesis time (as compared to the low complexity of TD-PSOLA and MBROLA). This problem, however, can be adressed to some extent by using techniques for real-time generation of signals from their spectral representation (as used in [3]).

5. CONCLUSION

In this paper we present the extension of HNM to diphone concatenation for high-quality text-to-speech synthesis. Informal listening tests, using natural prosody, have shown that HNM is a very good candidate for next generation TTS; the segmental quality of synthetic speech is close to the quality of the original sentences, without smoothing problems and without the buzziness or other oddities observed with other speech representations used for TTS. Because HNM does not require glottal closure instants, new voices can be easily integrated into TTS.

6. Acknowledgments

We would like to thank Mark Beutnagel, Alistair Conkie, and Ann Syrdal for many fruitful discussions during this work.

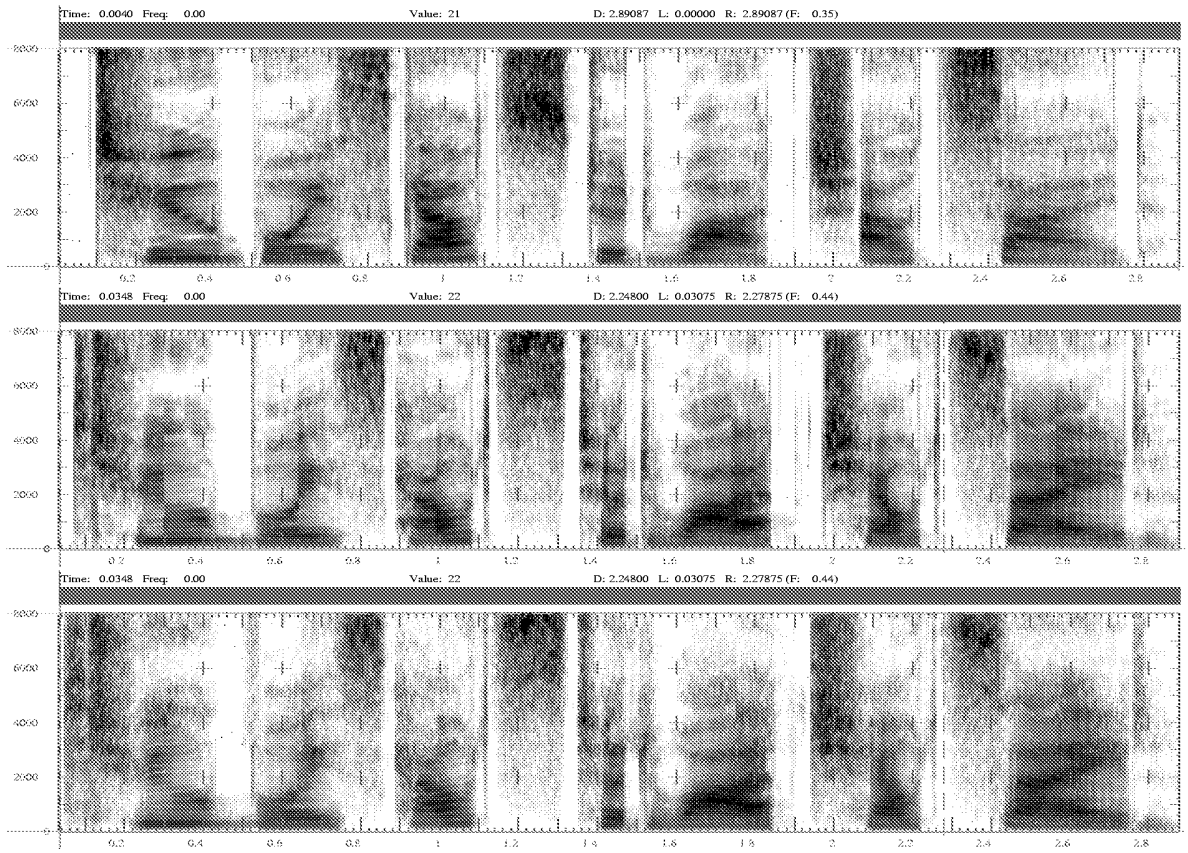

Figure 1: Spectrograms from a recorded natural version of a sentence (upper panel), a version synthesized with TD-PSOLA (middle panel) and a version synthesized with HNM (lower pannel). The sentence is: "Two boys scouts stood watch outside". Time is in seconds. None of the diphone units used with TD-PSOLA and HNM were cut from the recorded sentence.

7. REFERENCES

1. O. Boeffard and F.Violaro. Improving the robustness of text-to-speech synthesizers for large prosodic variations. In *Conf. Proc. of second ESCA-IEEE Workshop on Speech Synthesis*, pages 111–114, New Paltz, USA, Sept 1994.
2. T. Dutoit. *An introduction to text-to-speech synthesis.* Kluwer Academic Publishers, The Netherlands, 1997.
3. T. Dutoit and B. Gosselin. On the use of a hybrid harmonic/stochastic model for tts synthesis by concatenation. *Speech Communication*, 19:119–143, 1996.
4. T. Dutoit and H. Leich. Text-To-Speech synthesis based on a MBE re-synthesis of the segments database. *Speech Communication*, 13:435–440, 1993.
5. J. Laroche, Y. Stylianou, and E. Moulines. HNS: Speech modification based on a harmonic + noise model. *Proc. IEEE ICASSP-93, Minneapolis*, Apr 1993.
6. Michael W. Macon. *Speech Synthesis Based on Sinusoidal Modeling.* PhD thesis, Georgia Institute of Technology, Oct 1996.
7. E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5/6):453–467, Dec 1990.
8. R. Sproat and J. Olive. An Approach to Text-To-Speech Synthesis. In *Speech Coding and Synthesis*, pages 611–633. Elsevier, 1995.
9. Y. Stylianou. *Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification.* PhD thesis, Ecole Nationale Supèrieure des Télécommunications, Jan 1996.
10. Y. Stylianou. A pitch and maximum voiced frequency estimation technique adapted to harmonic models of speech. *IEEE Nordic Signal Processing Symposium.*, Sept 1996.
11. Y. Stylianou, J. Laroche, and E. Moulines. High-Quality Speech Modification based on a Harmonic + Noise Model. *Proc. EUROSPEECH*, 1995.
12. Xiaoqin Sun, Fabric Plante, Barry Cheetham, and Kenneth Wong. Phase modelling of speech excitation for low bit-rate sinusoidal transform coding. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 1691–1694, 1997.