TD-PSOLA VERSUS HARMONIC PLUS NOISE MODEL IN DIPHONE BASED SPEECH SYNTHESIS

Ann Syrdal, Yannis Stylianou, Laurie Garrison+, Alistair Conkie and Juergen Schroeter

AT&T Labs-Research, SIPS 180 Park Avenue, Florham Park, NJ 07932
email : [syrdal, styliano, adc, jsh]@research.att.com
+ AT&T, 101 Crawfords Corner Rd, Holmdel, NJ 07733
email : lfg@zippy.ho.att.com

ABSTRACT

In an effort to select a speech representation for our next generation concatenative text-to-speech synthesizer, the use of two candidates is investigated; TD-PSOLA and the Harmonic plus Noise Model, HNM. A formal listening test has been conducted and the two candidates have been rated regarding intelligibility, naturalness and pleasantness. Ability for database compression and computational load is also discussed. The results show that HNM consistently outperforms TD-PSOLA in all the above features except for computational load. HNM allows for high-quality speech synthesis without smoothing problems at the segmental boundaries and without buzziness or other oddities observed with TD-PSOLA.

1. INTRODUCTION

The goal of speech synthesis is to enable a machine to transmit orally information to a user in a man machine communication context [1]. However, in spite of the long history of speech synthesis, no one speech synthesis system available today is able to produce speech that could be characterized as natural or completely pleasant. In order to improve the speech quality of current text-to-speech (TTS) systems in terms of naturalness, three areas must be addressed [1]: 1) improved linguistic analyses, 2) improved prosody modeling, and 3) improved speech synthesis models. While all the above areas are equally important, this paper will investigate only the third.

There has been a considerable amount of research effort directed at the problem of speech representation for TTS. The advent of Linear Prediction (LP) has had its impact in speech coding as well as in speech synthesis [2]. However, the buzziness inherent in LP degrades perceived voice quality. Other synthesis techniques based on pitch synchronous waveform processing have been proposed such as TD-PSOLA [3]. TD-PSOLA is currently one of the most popular concatenation methods. Although TD-PSOLA provides good quality speech synthesis it has limitations which are related to its *non-parametric* structure; spectral mismatch at segmental boundaries and tonal quality when prosodic modifications are applied on the concatenated acoustic units. MBROLA [4] tries to overcome the TD-PSOLA concatenation problems by resynthesizing voiced parts with constant phase and constant pitch. This artificial processing is the main source of MBROLA's problems, like buzziness. Sinusoidal approaches have also been proposed for speech synthesis [5], [6]. They perform concatenation by making use of glottal closure instants a process which is not always successful [5], resulting in poor quality because of phase mismatch at segment boundaries. Formal or informal listening tests have been reported from many researchers in order to compare the above speech representations for text-to-speech. In [6], pitch-synchronous LPC was compared with a pitch-asynchronous sinusoidal model [7]. A preference for the sinusoidal model became clear. In [4], LPC, TD-PSOLA, and a pitch-asynchronous hybrid harmonic/stochastic (H/S) representation were compared with MBROLA. The conclusion was that MBROLA is comparable to TD-PSOLA while the H/S representation comes third followed by the LPC approach. Another test was carried out at CNET [8] comparing TD-PSOLA and another hybrid harmonic/stochastic (H/S) representation. This H/S representation was a modified version of the Harmonic plus Noise Model, HNM, proposed in [9] in the sense that the model used to this experiment required pitch marks to be locked at glottal closure instants. This was not a requirement in [9]. The results showed that while the quality of the synthetic speech produced by both systems was quite similar, the naturalness of the unvoiced sounds was noticeably better with the hybrid model than with TD-PSOLA.

A speech model has been proposed [9],[10] based on a pitch-synchronous Harmonic plus Noise (HNM) representation of speech. HNM has shown the capability of providing high-quality prosodic modifications [10] without buzziness and tonal quality encountered in previously reported methods. Recently, HNM has been proposed for diphone concatenation [11] and informal listening tests have shown that HNM-based synthetic speech is of high quality. Note that HNM does not require pitch marks unlike other pitch-synchronous speech representations.

In order to select a speech representation for our next generation TTS, it was decided to compare TD-PSOLA, the most popular to date concatenation method, with HNM. In this paper, we present results from a formal listening test comparing TD-PSOLA versus HNM. Small-scale TTS diphone inventories were recorded using pre-selected professional speakers. Two type of inventories were recorded for each speaker: a series of nonsense words and a series of English sentences. Because only the speech representation was under investigation, prosody from naturally spoken sentences was used. Synthetic sentences were rated for intelligibility, naturalness and pleasantness.

The first part of the paper is devoted to a brief description of the two speech representations used in the formal listening: TD-PSOLA and the extension of HNM to diphone concatenation. It is followed by the description of the formal listening test. Results and discussion are given in the third part of the paper.

2. TWO CANDIDATES FOR DIPHONE CONCATENATION

2.1. TD-PSOLA

The Time Domain Pitch Synchronous OverLap Add method, TD-PSOLA [3], relies on the speech production model described by the sinusoidal framework [7]. However, the parameters of this model are not estimated explicitly and for this reason TD-PSOLA is also referred to as "null" model [4]. The "analysis" process consists of extracting short-time *analysis signals* by multiplying the speech waveform by a sequence of time-translated analysis windows. The analysis windows are located around glottal closure instants and their length is proportional to the local pitch period. During unvoiced frames the analysis time instants are set at a constant rate. During the "synthesis" process a mapping between the synthesis time instants and analysis time instants is determined according to the desired prosodic modifications [3]. This process specifies which of the short-time analysis signals will be eliminated or duplicated in order to form the final synthetic signal.

2.2. HNM

HNM is based on a pitch-synchronous harmonic plus noise representation of the speech signal [10]. The spectrum is divided into two bands; the low band 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 frequency content of the high band is modeled by a time-varying AR model;its time-domain structure is represented by a piecewise linear energy-envelope function. The noise part, $n(t)$, is therefore assumed 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 $w(t)$:

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

A time-varying parameter refered to as *maximum voiced frequency* determines the limit between the two bands. During unvoiced frames the maximum voiced frequency is set to zero.

The first step of the HNM analysis consists of estimating pitch and maximum voiced frequency based on a time-domain pitch detector [12]. Then, harmonic amplitudes and phases are estimated by minimizing a weighted time-domain least-squares criterion. For the noise part, the spectral density function of the speech signal is modeled by an all-pole filter by use of a standard correlation-based method [13]. The variance of the speech signal is estimated as the gain of this filter. The analysis windows are set at a pitch-synchronous rate during the voiced portions of speech and at a constant rate during the unvoiced frames. Note that HNM *does not use* pitch marks locked on glottal closure instants in contrast to TD-PSOLA; however, the distance between two analysis time instants is one local pitch period and the analysis window is two local pitch periods long.

The second step of the HNM analysis consists of estimating a continuous spectral and phase envelope per voiced frame. The spectral envelope is estimated from the harmonic amplitudes by a discrete regularized cepstrum technique described in [14] using a warped frequency scale (Bark scale) [10]. The phase envelope is obtained by the phase unwrapping algorithm described in [10], under the constrain of a smooth "phase slope". Thus, an HNM voiced frame is fully described by its fundamental frequency, the number of harmonics, the discrete cepstrum coefficients, the phase envelope, the reflection coefficients of the AR filter and the gain of this filter (LP gain). An HNM unvoiced frame is only represented by the AR filter and its gain.

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 technique associates synthesis time instants with analysis time instants [10].

After the determination of synthesis instants, harmonic amplitudes and harmonic phases are retrieved by sampling the spectral and phase envelope respectively, at the harmonic frequencies of the target fundamental frequency. Then, HNM parameters have to be smoothed around diphone boundaries. The number of frames used in the smoothing process depends on the variance of the number of harmonics for voiced frames and on the variance of the LP gain for unvoiced frames. The phoneme boundaries inside each diphone define the maximum number of frames for smoothing. Finally, there is no smoothing at the boundary between unvoiced and voiced frames. Spectral amplitudes, LP gain and reflection coefficients are smoothed around concatenation points by a simple linear interpolation procedure. Phase smoothing is not so straightforward. First the phase offset is estimated between a diphone (left diphone) and its successor (right diphone) based on the cross correlation of two sinusoids which have the same amplitude and same frequency while having 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 of the first harmonic in the first frame of the right diphone. Next a 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).

3. QUALITY ASSESSMENT

For the purpose of the formal listening test, six professional female voices were recorded at a 16kHz sampling rate. Two types of diphone inventories were recorded for comparison: 1) a series of nonsense words which contained the diphones required to synthesize the test materials and 2) a series of English sentences which also contained the required diphones. The phonetic segmentation and alignment of both inventories was first performed automatically with Entropics Aligner software, whose output was subsequently verified and hand-corrected if obvious inaccuracies affecting target segments were found. A relatively minimal set of phones was used for speaker audition purposes.

The two synthesis techniques that were used to generate the TTS test materials were: 1) TD-PSOLA as it was implemented at AT&T Labs-Research and 2) a research implementation of HNM as it was presented in the previous section. Both methods used the same input and prosody, which was modeled on naturally spoken tokens of the test sentences recorded from each speaker. Table 1 shows the mean fundamental frequencies of the speakers and their standard deviations.

Speaker	Mean F0 (Hz)	S.D. F0 (Hz)
1	214	55
2	150	38
3	196	60
4	217	46
5	188	46
6	231	56

Table 1: Prosody characteristics of the speakers.

Three sentences were included in the test:

Two boy scouts stood watch outside.
I'm waiting for my pear tree to bear fruit.
We must complete every task.

All test sentences were equated for level.

Naturally spoken versions of the three test sentences were subjected to one of two *modulated noise reference unit* MNRU reference conditions, Q10 and Q35. Q10 served as a low-end reference point with MOS scores similar to those previously found for a low-end commercial 16kbps ADPCM encoded voice mail system. Q35 served as a high-end reference whose MOS scores are typically equivalent to very high quality telephone speech.

Speech samples were presented in two different modes: 1) in the wide bandwidth condition, speech signals were low-pass filtered by a brickwall filter set to 6.5 kHz and presented to listeners via headphones (ITU specifications) and 2) in the telephone bandwidth condition, speech signals were filtered for a nominal telephone bandwidth form 300 Hz to 3300 Hz and presented to listeners via AT&T Trad100 telephone receivers.

Independent subjective ratings of each test sentence for intelligibility, naturalness and pleasantness were made. For each test trial, listeners were presented a 5-point (MOS like) rating scale from which to select their judgments using a touch sensitive screen. For each of the three types of ratings a familiarization session preceded testing during which listeners were presented speech samples representing the full range of variation along the dimension being rated, and they were given practice in using the rating scale.

Listeners were 41 adults who were inexperienced in listening to or evaluating text-to-speech synthesis. The group was composed of 7 males and 34 females. No listeners reported any known hearing impairments. Listeners were tested in four groups of from 8 to 11 individuals.

For one half of each test session, speech signals were presented over headphones (wide bandwidth), and for the other half, they were presented through the telephone handsets (telephone bandwidth). The order of the two bandwidths was counterbalanced across the four test sessions, so that wide bandwidth was presented first for two groups, and telephone bandwidth was presented first for the other two. For each bandwidth, the three types of ratings (intelligibility, naturalness, and pleasantness) were blocked; that is, all the speech signals were presented for intelligibility ratings during one interval of time, naturalness ratings for all the signals were collected during another time interval, and pleasantness ratings during a third interval. Blocking of type of rating was done to avoid subjects' confusion over what quality they were rating in a given trial. The order of the rating types and of the speech signals within a rating block were randomized. The counterbalancing and randomization of the order of test items among test blocks and across groups was intended to control possible order effects in the test, such as learning or fatigue effects, by evenly distributing them among test items.

A total of 936 ratings were collected from each of 41 listeners, totaling 38,376 observations for the entire experiment. Repeated measures Analyses of Variance (ANOVAs) were performed on the data. There were significant main effects of speaker, synthesis method, and inventory, plus interactions.

Figure 1 compares mean ratings among Q35 (plus-mark), Q10 (star-mark), HNM (circle-mark) and TD-PSOLA (x-mark).

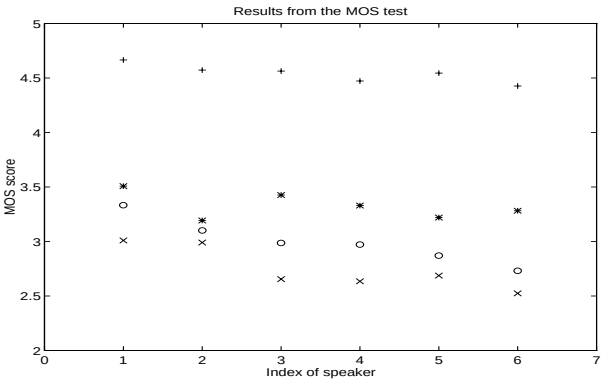

Figure 1: Average of all ratings (Intelligibility, Naturalness, Pleasantness) per speaker for Q35(+), Q10(*), HNM(o), and TD-PSOLA(x).

In more details, for Q35 (high-quality natural speech), Naturalness and Intelligibility ratings were equivalent, and they were significantly higher than Pleasantness ratings.

Lower-quality natural speech (Q10) had the following ordering: Naturalness $>$ Intelligibility $>$ Pleasantness. Synthetic sentences were rated higher for Intelligibility than for Naturalness or Pleasantness, which were equivalent.

HNM was consistently rated about 0.25 points higher than TD-PSOLA in Intelligibility, Naturalness and Pleasantness. Finally, the type of inventory from nonsense words versus from sentences has a smaller difference for HNM (0.10) than for TD-PSOLA (0.19).

It is worth noting that the diphone inventories were prepared twice because TD-PSOLA had serious quality problems with the first instance of the database. However, the quality of the HNM-based synthetic speech signals practically were equivalent for both databases.

4. DISCUSSION AND CONCLUSIONS

Results from the formal listening test show that HNM is a very good candidate for our next generation TTS. The score for HNM is consistenly higher than for TD-PSOLA in intelligibility, naturalness and pleasantness. The segment quality of synthetic speech is high, without smoothing problems and without buzziness observed with TD-PSOLA. An important point is that HNM is a pitch-synchronous system which does not require glottal closure instants as is the case with TD-PSOLA.

Other differences between TD-PSOLA and HNM (which basically justify the results form the formal listening test) are discussed

below.

TD-PSOLA was, so far, used for the low-cost high-quality prosodic modifications that this system can provide. However, TD-PSOLA eliminates/duplicates short-time waveforms extracted from the original speech signal by windowing. Although this process is very simple and the computational load is very low, this approach introduces a tonal noise quality because of the repetition of segments; an artificial long-time autocorrelation term in the output signal, perceived as some sort of periodicity (the problem is more noticeable during unvoiced frames and fricative voiced frames, of course).

Because of the non-parametric scheme of TD-PSOLA, limited smoothing possibilities are offered. This is an important issue in concatenative speech synthesis. Also, because its non-parametric scheme, TD-PSOLA does not allow complex modifications of the speech signal, such as increasing the degree of friction, or changing the amplitudes and phase relationships between pitch harmonics.

Comparing TD-PSOLA and HNM regarding computational cost, it is clear that HNM has a much higher complexity than TD-PSOLA. Actually, this is the only drawback of HNM versus TD-PSOLA. However, the HNM implementation presented in this paper is running in real time on an SGI Indy machine. Expecting the machine power to increase in the future, HNM complexity will not be at all a problem.

HNM, in contrast with TD-PSOLA, is a full-parametric pitch-synchronous harmonic plus noise representation of the speech signal. More explicitly, this means:

1. Smoothing diphone (or any other kind of units) boundaries is a simple and flexible procedure.
2. Prosodic modifications are quite straightforward and of high-quality [10].
3. Different prosody and spectral envelope modification methods can be applied to the harmonic and the noise part, yielding more natural-sounding synthetic speech.
4. Compression of a speech database. Preliminary results have shown that a bit rate of less than 6 kb/s is possible for wideband speech coding based on HNM.
5. HNM has also been tested on a voice conversion task [15] with very promising results. The possibility of voice conversion is important in TTS systems as a mean to create the desired variety of voices while avoiding recording a multitude speakers.

5. REFERENCES

[1] L. R. Rabiner, "Applications of Voice Processing to Telecommunications," *Proc. IEEE*, vol. 82, pp. 199–228, February 1994.

[2] R. Sproat and J. Olive, "An Approach to Text-To-Speech Synthesis," in *Speech Coding and Synthesis*, pp. 611–633, Elsevier, 1995.

[3] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, pp. 453–467, Dec 1990.

[4] T. Dutoit, "High quality text-to-speech synthesis : A comparison of four candidate algorithms," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 565–568, 1994.

[5] M. W. Macon, *Speech Synthesis Based on Sinusoidal Modeling*. PhD thesis, Georgia Institute of Technology, Oct 1996.

[6] M. Crespo, P. Velasco, L. Serrano, and J. Sardina, "On the Use of a Sinusoidal Model for Speech Synthesis in Text-to-Speech," in *Progress in Speech Synthesis*, pp. 57–70, Springer, 1996.

[7] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 744–754, Aug 1986.

[8] 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*, (New Paltz, USA), pp. 111–114, Sept 1994.

[9] J. Laroche, Y. Stylianou, and E. Moulines, "HNS: Speech modification based on a harmonic + noise model.," *Proc. IEEE ICASSP-93, Minneapolis*, Apr 1993.

[10] Y. Stylianou, J. Laroche, and E. Moulines, "High-Quality Speech Modification based on a Harmonic + Noise Model.," *Proc. EUROSPEECH*, 1995.

[11] Y. Stylianou, T. Dutoit, and J. Schroeter, "Diphone Concatenation using a Harmonic plus Noise Model of Speech," *Proc. EUROSPEECH*, 1997.

[12] Y. Stylianou, "A pitch and maximum voiced frequency estimation technique adapted to harmonic models of speech.," *IEEE Nordic Signal Processing Symposium.*, Sept 1996.

[13] S. M. Kay, *Modern Spectral Estimation*. Englewood Cliffs, New Jersey: Prentice Hall, 1988.

[14] O. Cappé and E. Moulines, "Regularization techniques for discrete cepstrum estimation," *IEEE Signal Processing Letters*, vol. 3(4), pp. 100–102, April 1996.

[15] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion.," *Accepted to IEEE Proc. on Speech and Audio Processing.*, 1996.