DCT-BASED AMPLITUDE AND FREQUENCY MODULATED HARMONIC-PLUS-NOISE MODELLING FOR TEXT-TO-SPEECH SYNTHESIS

Kris Hermus[§], Hugo Van hamme[§], Werner Verhelst[†], Sufian Irhimeh[‡], and Jan De Moortel[‡]*

[§]Dept. ESAT, Katholieke Universiteit Leuven, Belgium
{kris.hermus,hugo.vanhamme}@esat.kuleuven.be

[†]Dept. ETRO, Vrije Universiteit Brussel, Belgium
{werner.verhelst}@etro.vub.ac.be

[‡]Nuance Communications Belgium
{sufian.irhimeh,jan.demoortel}@nuance.com

ABSTRACT

We present a harmonic-plus-noise modelling (HNM) strategy in the context of corpus-based text-to-speech (TTS) synthesis, in which whole speech phonemes are modelled in their integrity, contrary to the traditional frame-based approach. The pitch and amplitude trajectories of each phoneme are modelled with a low-order DCT expansion. The parameter analysis algorithm is to a large extent aided and guided by the pitch contours, and by the phonetic annotation and segmentation information that is available in any TTS system. The major advantages of our model are : few parameter interpolation points during synthesis (one per phoneme), flexible time and pitch modifications, and a reduction in the number of model parameters which is favourable for low bit rate coding in TTS for embedded applications. Listening tests on TTS sentences have shown that very natural speech can be obtained, despite the compactness of the signal representation.

Index Terms— Speech synthesis, speech coding, speech analysis, speech processing

1. INTRODUCTION

In corpus-based TTS, speech is synthesised by concatenating natural speech segments that are looked up in a large segment database. Especially for embedded applications, the memory resources are low, and a compact representation and encoding of the database is of crucial importance. The lower the number of segments in the database, the more concatenation points and the more frequently the temporal and spectral properties of the segments will have to be modified. In this respect, a *model-based* coding strategy, that facilitates the creation of smooth concatenation and natural sounding speech modifications, is an extra asset.

In this paper, we describe a speech modelling approach that can fulfil the above requirements, and that is widely known to produce high quality speech that is almost indistinguishable from natural speech, namely harmonic-plus-noise modelling (HNM) [1]. In HNM, a series of sinusoids with harmonic frequencies for the voiced speech parts is combined with a synthetic noise signal for the unvoiced parts. HNM can rely on the existence of a broad range of powerful and flexible coding strategies for the parameters of a sinusoidal model (SM), and on computationally efficient sinusoidal speech synthesis algorithms, which limit the hardware cost for end-user devices.

*Kris Hermus is a Postdoctoral Research Fellow of the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT-Flanders), project 'ESSCorTeSS'.

Apart from the above mentioned intrinsic advantages of HNM, the application of this model *in a TTS context* opens some nice perspectives for *the way we can extract the SM parameters*, thereby significantly extending the possibilities of a standard speech modelling and coding application. First, the speech modelling and coding of the segment database of a TTS system occurs only once and is performed off-line, such that powerful parameter extraction algorithms can be applied, that either further reduce the number of parameters of the representation for a given speech quality, or increase the speech quality for a given number of model parameters. Second, the sinusoidal analysis can benefit from an accurate pitch labelling that is always at hand in a TTS system. Third, an accurate phonetic transcription of the speech utterances is available, which can be exploited to adapt the parameter extraction to the phoneme at hand (e.g. type of voicing, transitional speech).

Another major advantage of speech modelling in a TTS framework is the availability of an *accurate speech segmentation* In this paper, we fully exploit this property. We are doing away with the traditional frame-based approach in sinusoidal analysis and synthesis, and model *whole phonemes* instead. The given segmentation guarantees that the changes of the signal characteristics are minor and/or only slowly time-varying within one signal segment (phoneme). Discontinuities will mainly occur at the signal boundaries. As such, we are able to construct an accurate signal representation with a limited number of parameters, that remains valid for a longer time duration. Due to coarticulation effects, we can expect smooth variations of both the amplitudes of the speech harmonics and of the pitch within one phoneme. This will motivate the use of pitch and amplitude modulation of the sinusoidal components. This so-called *long-term modelling* approach has already been described for traditional sinusoidal modelling, e.g. in [2] (phase/pitch) and in [3] (amplitude). The use of long-term modelling *significantly reduces the need for parameter-interpolation* during synthesis, since only one such interpolation will be needed per phoneme, instead of one interpolation every (say) 10 ms.

2. DCT-BASED AM-FM HNM

Given the phonetic annotation and segmentation of a TTS corpus, we split every speech utterance into segments, each containing one single phoneme. For a particular phoneme segment $s[k]$ $(0 \leq k \leq N\text{-}1)$ with sampling frequency f_s, we propose the following harmonic-plus-noise (HNM) model :

$$\hat{s}[k] = \underbrace{g_h(k).\sum_{i=0}^{L(k)} a_i(k)\sin\left(2\pi i\int_0^k f_0[t]dt + \phi_{0,i}\right)}_{\text{harmonic part}} + \underbrace{g_n(k).n[k]}_{\text{noise part}}$$

Sinusoidal part The choice for a *harmonic* sinusoidal model in which the frequencies of all sinusoids are a multiple of the time-varying fundamental frequency is in correspondence with speech production, and it significantly reduces the number of model parameters compared to a traditional sinusoidal model in which the frequencies are not constrained. The parameters of the harmonic part are :

- $L(k)$ time-varying number of harmonics, determined by the voicing cut-off frequency (see below) at sample k
- $a_i(k)$ normalised amplitude contour of i^{th} harmonic (DCT-based)
- $f_0(k)$ DCT-based representation of the pitch contour
- $\phi_{0,i}$ phase offset of i^{th} harmonic
- $g_h(k)$ gain function of the harmonic part

The pitch and amplitude modulation is based on a DCT expansion, which has some clear advantages : ability to model smooth transitions, good interpolation/modification characteristics, and low sensitivity to coding errors.

Noise part The noise part is represented by a randomly generated noise signal $n[k]$, with time-varying bandwidth $[L(k)f_0(k), f_s/2]$, with a spectral envelope modelled by a time-varying auto-regressive (AR) model, and with a gain contour $g_n(k)$.

The spectral envelope parameters will be used for both the harmonic and the noise part, such that the noise gain function $g_n(k)$ contains the only parameters that are exclusively used for the noise part.

3. EXTRACTION OF MODEL PARAMETERS

Number of harmonics The number of harmonics in the SM depends on the voicing type of the phoneme. For unvoiced phonemes, the sinusoidal part is zero and the HNM reduces to a synthetic noise signal. For voiced phonemes, the number of harmonics is determined by a so-called voicing cut-off frequency (VCO) estimation algorithm. The VCO is defined as the frequency that separates a low-frequency harmonic part from a high-frequency noise part. The number of voiced harmonics is then simply given by the VCO divided by the local pitch frequency, rounded to the closest integer. Numerous estimation algorithms for the VCO exist, e.g. [4, 1]. In the context of this work, we derived a new VCO estimation algorithm (see [5]) that yields very smooth and accurate VCO contours (figure 1). For high quality speech we model the VCO contour with a low-order DCT expansion. If a lower quality is sufficient, one can use a fixed phoneme-dependent VCO that is interpolated at the frame boundaries to obtain smooth transitions.

DCT-based pitch modelling For the extraction of the pitch contour and the phase offsets of the harmonics (see below), we start from the available pitch labelling information. The pitch contours are usually very accurate, since they were performed off-line, and sometimes manually verified. In order to further increase its accuracy and to guarantee smoothness of the pitch contour, we propose the following optimisation strategy.

Let $s[k]$ $(0 \leq k \leq N-1)$ be a voiced phoneme and $L(k)$ be the number of harmonics. We now define the following constant-amplitude, DCT-based frequency modulated, sinusoidal representation :

$$\hat{s}[k] = \sum_{i=0}^{\bar{L}} a_i \cos\left(2\pi i \int_0^k f_0[t]dt\right) + \sum_{i=1}^{\bar{L}} b_i \sin\left(2\pi i \int_0^k f_0[t]dt\right)$$

with $\bar{L} = \max(L(k))$, and with the DCT-based pitch model of order

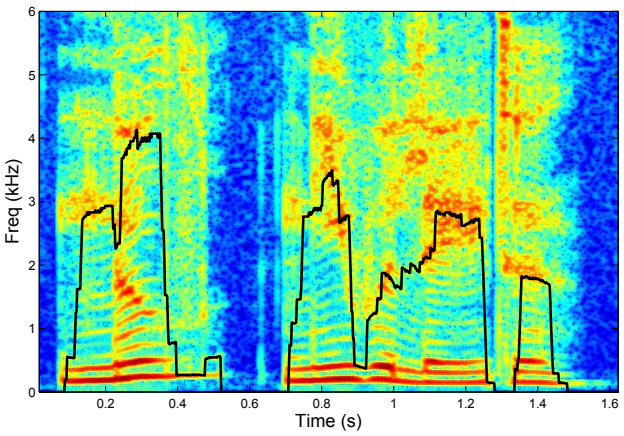

Fig. 1. Spectrogram of a short speech utterance, with automatically annotated VCO contour.

M_{FM} given by

$$f_0[k] = \sum_{j=1}^{M_{\text{FM}}+1} w_j \, f_{\text{dct,j}} \cos\left(\frac{\pi(2k-1)(j-1)}{2N}\right) \qquad (1)$$

and with $w_j = 1/\sqrt{N}$ for $j = 1$ and $w_j = \sqrt{2/N}$ for $2 \leq j < M_{\text{FM}} + 1$.

We now solve this optimisation problem in LS sense. The problem is highly non-linear in the pitch parameters $f_{\text{dct,j}}$, but linear in a_i and b_i for fixed $f_{\text{dct,j}}$. We first derive good starting points for $f_{\text{dct,j}}$ by taking the DCT expansion of the given pitch contour, and then refine this initial pitch estimate in a combined optimisation of a_i and $f_{\text{dct,j}}$ using the Levenberg-Marquardt (LM) update rule. See [6] for an analytical derivation of the solution of this optimisation problem. The iteration is continued until convergence is reached.

Phase offset The phase offsets $\phi_{0,i}$ are simply obtained from the above optimisation as

$$\phi_{0,i} = \tan^{-1}\left(\frac{b_i}{a_i}\right) \quad \text{for } k = 1 \ldots \bar{L}$$

Time-frequency AR modelling The assumption of constant amplitudes for the sinusoids within one frame becomes unacceptable for longer speech segments. In natural speech, each sinusoidal track has its own amplitude modulation (AM) factor, related to the time-evolving vocal tract filter.

An explicit DCT-based expansion of the AM of every harmonic is not feasible due to the high number of parameters involved. It is therefore common practice in SM to model the spectral envelope by means of auto-regressive (AR) modelling, and to obtain the harmonic amplitudes by sampling this spectral envelope at the pitch frequency and its harmonics. Motivated by the observation that within one phoneme *AR parameters* tend to *change smoothly over time* (figure 2), we apply Time-Frequency Auto-Regressive modelling (TFAR), which is a generalisation of standard AR modelling towards *time-varying* AR modelling. The properties of TFAR with an FFT-based expansion of the AR parameters have been studied in [7]. In this work, we use a low-order *DCT-based* expansion to model the temporal variations in the AR parameters.

Mathematically, the TFAR problem of order (M_{AM}, K) is expressed as follows. Find the parameters $p_{m,l}$ that minimise

$$\sum_{k=0}^{N-1} e^2[k]$$

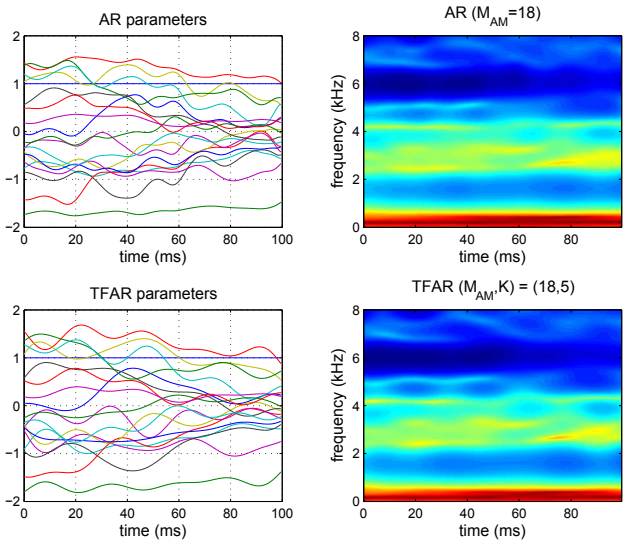

Fig. 2. Classical AR model (top) and TFAR model (bottom) for a realisation of the phoneme /e/. Left : prediction coefficients versus time. Right : corresponding spectrograms.

subject to

$$s[k] = -\sum_{m=1}^{M_{\text{AM}}} p_m(k) s[k-m] + e[k] \qquad (2)$$

with

$$p_m(k) = \sum_{i=1}^{K} w_i \, p_{m,i} \, \cos\left(\frac{\pi(2k-1)(i-1)}{2N}\right)$$

As before, w_i is given by $1/\sqrt{N}$ $(i=1)$ or by $\sqrt{2/N}$ $(2 \leq i \leq K)$.

In the above expressions, we observe both time $(s[k-m])$ and frequency $(\cos(\pi(2n-1)(i-1)/2N))$ shifts of the input signal $s[k]$, hence the name *time-frequency AR*.

The order K is adapted to model the spectral envelope with sufficient detail without capturing the modulation due to the source signal.

The solution of the TFAR problem of order (M_{AM}, K) requires the estimation of $M_{\text{AM}}.K$ prediction parameters $p_{m,l}$. Even though the set of equations in (2) can be solved as a simple linear LS problem, more refined ways to solve it exist [7].

The main advantages of a TFAR instead of the classic AR approach is the reduction in number of model parameters, its intrinsic capability to model non-stationarity in speech[1], and the smoothness of the amplitude contours (no linear interpolation is needed). These benefits come with a significantly higher computational load during analysis, but this is not an issue in the current application. An example is given in figure 2, where in the upper panel a standard AR model of order 18 and small frame shift is compared in the lower panel to a TFAR of order $(18, 5)$ for a realisation of the phoneme /e/ with 100 ms duration. The parametrisation over time of the TFAR model seems appropriate to model the time-varying nature of the signal. Note that the order of the AR model can be made dependent on the phoneme (higher for vowels, lower for non-nasalised consonants) which further reduces the number of model parameters.

Gain function $\mathbf{g_h}[\mathbf{k}]$ The global gain contour $g_h[k]$ is modelled by a low-order DCT expansion. The parameters are found by LS fitting of the harmonic part $\sum_{i=0}^{L(k)} \hat{a}_i(k) \sin\left(2\pi i \int_0^k f_0[t]dt + \phi_{0,i}\right)$ to the original signal[2]. The amplitudes $\hat{a}_i$ are obtained by sampling the temporal envelope that was modelled with a TFAR model.

Let $\mathbf{D}$ be an $N \times (M_{\text{AM}}+1)$ matrix with the DCT expansion functions on its columns, and let $\mathbf{H}$ be an $N \times (M_{\text{AM}}+1)$ matrix with the signal $\sum_{i=0}^{L(k)} \hat{a}_i(k) \sin\left(2\pi i \int_0^k f_0[t]dt + \phi_{0,i}\right)$ on every column. The vector $\mathbf{q}$ of coefficients of the DCT expansion of $g_h[k]$ are then found by solving the set of equations $(\mathbf{D} * \mathbf{H})\mathbf{q} = \mathbf{s}$ in the least squares sense, with $*$ denoting the Hadamard (element-wise) product of two matrices and $\mathbf{s}$ the vector containing $s[k]$.
For phonemes with strong temporal dynamics (e.g. plosives or /r/), the above approximation will not be accurate enough. In this case, the energy envelope is updated every 2 ms, which is a common procedure.

Gain function $\mathbf{g_n}[\mathbf{k}]$ The gain of the noise part is obtained as a low-order DCT expansion of the short-term energy envelope of the high-pass (cut-off is VCO frequency) version of TFAR residual signal $e[k]$.

4. HNM SYNTHESIS

We first describe how the phoneme segments $\hat{s}[k]$ are resynthesised from the HNM analysis parameters.

Harmonic part First, the phase envelope $\phi(k)$ is obtained by considering k as a continuous variable and integrating[3] (eq. 1) over k. In a second step, we construct a set of unit-amplitude sinusoids, $\sin(i\phi(k) + \phi_{0,i})$, with $k = 0 \ldots (N\text{–}1)$ and $i = 0 \ldots \max(L(k))$. We then apply an amplitude modulation to each of the sinusoids, derived from the all-pole TFAR filter with filter coefficients $p_m(k)$ to obtain the correct spectral envelope. The summation of these AM modulated sinusoids is sent through a time-varying linear-phase low-pass (LP) filter with cut-off frequency $L(k)f(k)$. Finally, this signal is multiplied sample-by-sample by the gain contour.

Noise part The objective here is to generate artificial noise with the correct spectral and temporal characteristics and to properly combine it with the HSM signal.

We first generate a unit-variance fullband white noise signal $w[k]$, apply a time-varying linear-phase high-pass (HP) filter with cut-off frequency $L(k)f(k)$ (only for voiced speech; the LP and HP filters are matched and sum up to 1, and have a controllable transition zone to obtain smooth transitions from harmonic to noise spectral bands), send it through the all-pole TFAR filter, and finally apply the correct gain contour.

The synthesised phoneme is obtained by a simple summation of the noise part and the harmonic part.

Phoneme concatenation The synthesised speech sentence is now obtained by the concatenation of the resynthesised phoneme segments. Smooth phoneme to phoneme transitions can be obtained by either using a time-domain technique like WSOLA [8], or - more interestingly - by using direct parameter interpolation of the HNM parameters during synthesis. For more details on the latter, see the pioneering work of Mc. Aulay & Quatieri [9].

5. EVALUATION

Our HNM model was extensively evaluated on a TTS system with a donor database of a Dutch female voice. From this database, we selected a subset of 150 short speech recordings (10 minutes of speech)

[1] Although not proven, this may lead to a better spectral modelling in transitional speech segments.

[2] We can indeed use $s[k]$ and not a low-passed version of it, since the number of harmonics is low compared to the signal length N.

[3] Recall that the frequency $f(k)$ is the time derivative of the phase $\phi(k)$.

for simulation and testing, along with their pitch labelling and phonetic segmentation. Besides, 36 TTS test sentences were defined that are synthesised by concatenation of short segments from the original donor database.

5.1. Simulations

Quality of the model Simulations indicated that a 4^{th} or 5^{th} order DCT model is capable of accurately representing the amplitude and frequency variations that are present within one single phoneme. This was illustrated by the fact that the goodness-of-fit of the harmonic model (measured in terms of the signal-to-noise ratio) remains sufficiently high (mostly between 10 and 15 dB for vowels), even for long realisations of phonemes (e.g. above 150 ms). As such, for vowels and voiced consonants without strong dynamics, the number parameters is (almost) independent on the length of the phoneme realisation (see below). These findings confirm that signal variabilities are indeed limited within a phoneme, and that our approach to model phonemes in their integrity is valuable. Even though the use of a fixed order for the AM and FM will yield accurate results in most cases, it can be interesting to adapt the value of M_{FM} and K to the segment characteristics (length, phoneme identity, pitch contour,...). See also [3] for related work.

Numerical stability Even though the objective function is strongly non-linear with lots of local minima, the Least Squares optimisation in the FM HSM converges rapidly (3 to 4 iterations are sufficient in most cases), which proves that the initialisation based on the available pitch files is very accurate. The reason for the accurate pitch estimation is that it was performed off-line, which makes it possible to include a DP approach with a sufficiently large look-ahead to avoid halving and doubling errors. No examples of divergence were encountered.

Number of model parameters For vowels and voiced consonants for which a DCT-based gain contour can be used, we have on average the following number of model parameters : phoneme length (1), VCO contour (4), pitch contour (5), phase offsets (10...40)[4], noise gain contour (6), harmonic gain contour (6), and TFAR (90). This yields a total number of parameters in the range 120 ... 150.

For unvoiced phonemes we have : phoneme length (1), energy contour (6 for DCT-based contour *or* 1 parameter every 2 ms in case of strong dynamics[5]), and TFAR (90). This yields a total number in the range 100 ... 130. If we use a lower order for the spectral envelope (e.g. 10), this number can be reduced by 40.

We now make an estimate of the number of parameters per second of speech. The average phoneme lengths from the database are as follows : 74 ms for a vowel, 63 ms for a voiced consonant, 100 ms for an unvoiced consonant, 42 ms for a voiced plosive, 53 ms for an unvoiced plosive. Based on the relative occurrences of these phoneme classes, we end up with approximately 15 phonemes per second of stored speech, from which 1700 to 2150 model parameters per second have to be extracted, and subsequently coded. Given the excellent coding algorithms that exist for AR parameters, and the low sensitivity of DCT coefficients to coding errors, we are confident that it will be possible to obtain natural speech quality at low bit rates based on our DCT-based harmonic-plus-noise model.

5.2. Listening experiments

Extensive listening tests were performed in which ten subjects (the authors included) participated.

Resynthesis of donor files In these experiments, we applied HNM analysis-resynthesis to the speech files from the donor database. Listening tests revealed that the synthesised speech is of very high quality and *almost indistinguishable from natural speech*.

[4]This number is dependent on the voicing and on the pitch.

[5]based on an average length of 50 ms for this kind of phonemes

TTS synthesis In this experiment, we used WSOLA-based concatenation for the concatenation of phonemes after HNM analysis-resynthesis. In this way, 36 different TTS sentences were produced. Listening tests have shown that the TTS sentences are of good quality. The artefacts - e.g. discontinuities in prosody - are mainly due to the compromises that have to be made in the TTS segment selection, and they are inherent to the corpus-based TTS paradigm.

We are currently replacing the WSOLA-concatenation by an interpolation-based synthesis and concatenation. We believe that this will lead to smoother transitions at the segment boundaries and to an increase in the perceptual quality of the synthesised speech.

6. CONCLUSIONS

We have shown that phonetic segmentation and annotation information from a TTS system can be successfully exploited to model speech phonemes in their integrity with a harmonic-plus-noise model that incorporates a DCT-based expansion of the pitch and amplitude contours. Listening tests have shown that synthesised speech is very natural, despite the limited number of parameters in the speech model. The good coding properties of our HNM parameters make it feasible to develop a (very) low bit rate coder, which is especially useful for embedded applications in TTS. In our future work, we will investigate coding strategies and the possibilities of this representation for time and pitch scaling during synthesis, since this could reduce the number of speech segments that are needed in the donor database of the TTS system.

7. REFERENCES

[1] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Trans. on SAP*, vol. 9, no. 1, pp. 21–29, Jan. 2001.

[2] L. Girin, M. Firouzmand, and S. Marchand, "Long term modeling of phase trajectories within the speech sinusoidal model framework," in *Proc. ICSLP*, Jeju Island, Korea, Oct. 2004.

[3] M. Firouzmand and L. Girin, "Perceptually weighted long term modeling of sinusoidal speech amplitude trajectories," in *Proc. ICASSP*, Philadelphia, PA, USA, Mar. 2005, vol. I, pp. 369–372.

[4] R.J. McAulay and T.F. Quatieri, "Sinusoidal coding," in *Speech Coding and Synthesis*, W.B. Kleijn and K.K. Paliwal, Eds., pp. 121–173. Elsevier, 1995.

[5] K. Hermus, H. Van hamme, and S. Irhimeh, "Estimation of the voicing cut-off frequency contour based on a cumulative harmonicity score," *IEEE SP Letters*, 2007, To appear.

[6] H. Van hamme, "Robust speech recognition using cepstral domain missing data techniques and noisy masks," in *Proc. ICASSP*, Montreal, Canada, May 2004, vol. I, pp. 213–216.

[7] M. Jachan, G. Matz, and F. Hlawatsch, "Time-frequency-autoregressive random processes: modeling and fast parameter estimation," in *Proc. ICASSP*, Hong Kong, P.R.C., Apr. 2003, vol. VI, pp. 125–128.

[8] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," in *Proc. ICASSP*, Minneapolis, U.S.A., Apr. 1993, vol. 2, pp. 554–557.

[9] R.J. McAulay and T.F. Quatieri, "Speech analysis-synthesis based on a sinusoidal representation," *IEEE Trans. on ASSP*, vol. 34, no. 4, pp. 744–754, Aug. 1986.