Syllable-based Pitch Encoding for Low Bit Rate Speech Coding with Recognition/Synthesis Architecture

Milos Cernak[1], *Xingyu Na*[1,2], *Philip N. Garner*[1]

[1]Idiap Reseach Institute
[2]Beijing Institute of Technology
{Milos.Cernak, Xingyu.Na, Phil.Garner}@idiap.ch

Abstract

Current HMM-based low bit rate speech coding systems work with phonetic vocoders. Pitch contour coding (on frame or phoneme level) is usually fairly orthogonal to other speech coding parameters. We make an assumption in our work that the speech signal contains supra-segmental cues. Hence, we present encoding of the pitch on the syllable level, used in the framework of a recognition/synthesis speech coder with phonetic vocoder. The results imply that high accuracy pitch contour reconstruction with negligible speech quality degradation is possible. The proposed pitch encoding technique operates on 30–35 bits per second.

Index Terms: speech coding, pitch analysis, speech synthesis

1. Introduction

The state of the art in audio coding is well represented by the MPEG (moving picture experts group) standards. In MPEG-1 the ubiquitous MP3 was introduced that makes use of perceptual limitations of the human ear to encode arbitrary signals. In MPEG-2, AAC (advanced audio coding) replaced MP3, bringing in the likes of Huffman coding and the MDCT (modified discrete cosine transform).

The above codecs make no assumptions about the content of the signal. This changed in MPEG-4 [1], which is more of a toolbox, where different codecs could be used for different purposes. In particular, if a signal is known a-priori to be speech, it is possible to use a speech codec. MPEG-4 includes two speech coders:

- Code-excited linear prediction (CELP) [2] and
- Harmonic Vector Excitation Coding (HVXC) [3],[4].

CELP operates at 4.0–16.0 kbit/s and HVXC constant bitrate on 2.0–4.0 kbit/s. Using a variable bit-rate technique, HVXC can also operate at lower bit-rates, typically 1.2–1.7 kbit/s. Very Low Bit Rate (VLBR) speech coding targets an order of magnitude lower bit rates, typically 100 – 150 bps. A VLBR system can be achieved by the integration of phoneme recognition (as an encoder) and speech synthesis (as a decoder), where a sequence of symbols, such as phonemes, is transmitted instead of a compressed audio signal. Additional information such as pitch and duration of the symbols is required to recover the original prosody. While corpus-based techniques have been applied in the past for VLBR speech coding systems (e.g., [5]), HMM-based speech synthesis systems (HTS) [6] are beneficial from an adaptation point of view, and the system footprint [7]. Here, a phonetic vocoder is usually used:

- The STRAIGHT vocoder [8]

Pitch (the fundamental frequency of the harmonic part of a signal) encoding differs with the type of the coding algorithm. While in the audio coders the pitch coding is fairly orthogonal to the coding of other parameters, this might generally not be true in speech coders, such as CELP. However, most previous research efforts on low bit rate speech coding were concentrated around independent pitch coding, based on quantizing pitch values on frame level [9], [10].

While CELP coders make assumptions about possible decomposition of the signal with a source-filter model of speech production, in our work we make the assumption that the speech signal contains supra-segmental cues. Hence, we present a phonetic/supra-segmental combination applied to pitch contour encoding. Instead of coding the pitch on frame or phoneme levels, we encode the pitch on the syllable level. The decoder directly reconstructs voiced segments, and using speech parameters generated from HMMs, a phonetic vocoder synthesises the speech.

In addition, we have already showed in our previous work [11] the importance of syllable context in a low bit rate recognition/synthesis speech coding system. Proposing also syllable-based pitch encoding together with syllable context HTS, we aim to unify the information transmission mechanisms on the same contextual level.

The structure of the paper is as follows: the next section describes pitch contour encoding algorithm, section 3 describes the experiments followed by conclusions in section 4.

2. Supra-segmental pitch encoding

In speech codecs, the goal of pitch encoding is to retain the original prosody through transmission. Previous research on frame-based codecs by Chen and Wang [12] has already introduced supra-segmental pitch encoding on a syllable span. The frame-based framework is incapable of aligning segmental and supra-segmental data streams. In the work by Chen and Wang [12], the contour parameters were quantized according to the tonality of Mandarin. Thus, it was a language dependent pitch encoder. In the recognition/synthesis framework, spectrum and pitch parameters are inherently aligned supra-segmentally, which allows supra-segmental modelling for separate parameters. To build a language-independent pitch encoder, the syllable is selected as the supra-segmental unit because the span covers the fundamental pitch variant for prosody event as suggested by the linguistic research by Xu et al. [13], [14]. The boundaries of a syllable are aligned with the phonemes it contains. Within a syllable, there is usually only one major pitch contour covering a vowel or a pseudo-vowel consonant. Given the syllable boundaries, we choose the beginning and the ending of the longest pitch contour within a syllable as the onset and the offset times for pitch encoding, and transmit the corresponding frame indices from the original pitch.

Pitch contour encoding generally involves parameterization using curve fitting techniques. The *discrete (Legendre) orthogonal polynomial* (DLOP) has been proved capable of capturing speaker identity in speech synthesis [15] and speaker verification [16]. A segment of a normalized contour with the length of $N+1$, $f(i/N)$, is approximated using DLOP as

$$\hat{f}\left(\frac{i}{N}\right) = \sum_{j=0}^{J-1} a_j \cdot \phi_j\left(\frac{i}{N}\right), \quad 0 \leq i \leq N \qquad (1)$$

where the parameters are

$$a_j = \frac{1}{N+1}\sum_{i=0}^{N} f\left(\frac{i}{N}\right) \cdot \phi_j\left(\frac{i}{N}\right), \quad 0 \leq j \leq J-1 \qquad (2)$$

and J represents the order of approximation. The first four polynomials $\phi_j(i/N)$ are defined and used as introduced by Chen and Wang [12]. Using the transformation described by Eq. 2, a pitch contour is encoded into a parameter vector $(a_0, a_1, \ldots, a_{J-1})$.

At the decoder, the pitch contour between the onset and the offset frame indices are reconstructed using Eq. 1. Other frames are assumed to be unvoiced. Hence, the supra-segmental cues, i.e. the polynomial parameters and the onset/offset information, compose the data stream of the proposed pitch encoder.

3. Experiments

One challenge for the recognition/synthesis framework is the recognition error. Distorted phoneme sequences lead to different syllabification results. Therefore, the pitch encoding method should be robust against syllabification differences, so that the original prosody can be reconstructed from the phoneme sequence obtained from the recognizer in the decoding end. Hence, we evaluated three different syllabification methods:

1. Textual: syllable boundaries are extracted from syllable context labels generated by a speech synthesis front-end.
2. Phonetic: syllable boundaries are estimated from phonemes. As we used the true input labels, we extracted a stream of phonemes, and syllable boundaries were selected as transitions from a consonant to a vowel.
3. Manual: syllable boundaries are extracted from manually corrected syllable context labels, that were available with our test data.

We evaluated the impact of these syllabification methods on the pitch encoding quality. The most important method here is the phonetic one, as this is the only method that is available in a real recognition/synthesis speech coder.

As we selected syllables as the segmentation of voicing parts of the speech signal, we hypothesise, (i) that any syllabification method could be used for pitch encoding without significant performance degradation. Secondly, if the syllabification method is good enough, i.e., it segments the voiced speech at “right places”, we hypothesise (ii) that the lowest order approximation would be sufficient. To prove our hypotheses, we propose the experimental setup described in the next section. Results described in later sections 3.2.1 and 3.2.2 discuss both hypotheses (i) and (ii), respectively.

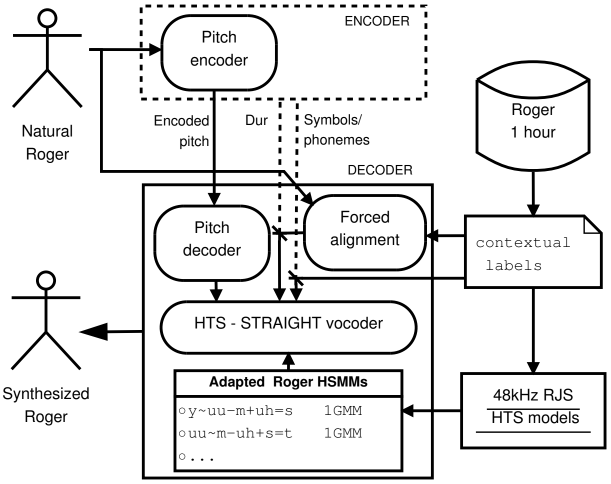

Figure 1: *VLBR speech coding experimental setup with recognition-synthesis architecture, abstracting the encoder (dotted lines) except for pitch encoding and decoding modules.*

3.1. Experimental setup

We used the Roger corpus [1] of 1 hour of speech data from the University of Edinburgh. Roger was used as a test speaker. We used an existing voice model trained from 4 hours of speech uttered by a British speaker RJS, and adapted it using the MAP-VTLN parameter estimation of [17] to the Roger voice. HTS models 59 dimensional mel-generalized cepstral features, pitch as $\log(f0)$, five band aperiodicity, their delta and delta-delta coefficients, and duration in the unified framework of hidden semi-Markov models (HSMMs). The STRAIGHT vocoder [8] was used to synthesize speech from the parameters generated using HTS.

The experimental setup is similar to the setup we used in the previous work [11], and it is depicted in Fig. 1. Focusing on pitch contour encoding, we abstracted the encoder side, and used the true input to the decoder, i.e., symbol sequence from the syllable context labels, and state durations from the forced alignment against natural speech.

The HTS system for speech synthesis uses by default the full-context labels that specify pentaphone phonetic context followed by the contextual factors and/or their combination (grouped by letters shown in Tab. 1). We have already showed that high speech quality recognition/synthesis VLBR speech coding system can be achieved just with reconstruction of syllable context [11]. The word, phrase and utterance contexts

Table 1: *Groups of the contextual factors, as defined by the HTS documentation.*

Context	Previous	Current	Next
Syllable	/A	/B	/C
Word	/D	/E	/F
Phrase	/G	/H	/I
Utterance		/J	

[1] Available through Blizzard Challenge 2010 at `http://www.cstr.ed.ac.uk/projects/blizzard/`

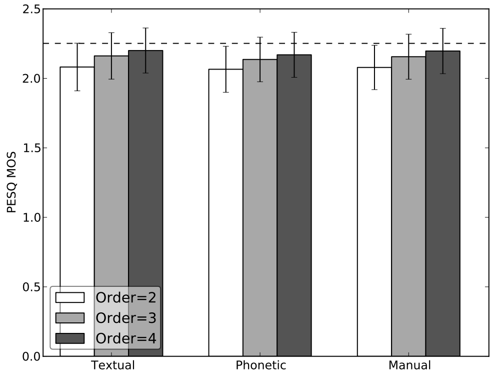

Figure 2: *Objective evaluation results for speech coding: three syllabification methods used textual, phonetic and manual labels, and three coding alternatives with different approximation orders. The dotted line represents reference with original pitch.*

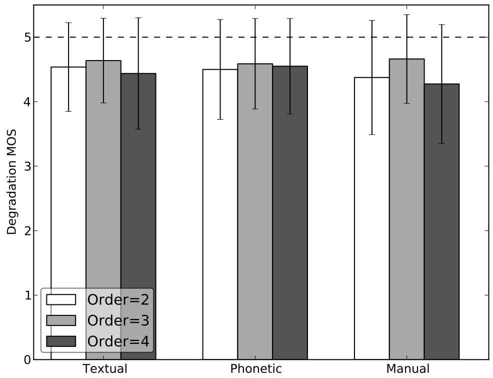

Figure 3: *Subjective evaluation results for speech coding: three syllabification methods used textual, phonetic and manual labels, and three coding alternatives with different approximation orders. The dotted line represents inaudible degradation.*

(the contextual groups /D – /J, basically guess part-of-speech of words and numbers of the syllables, words and phrases) are unimportant on the decoder side, and it is not necessary to deal with them. Therefore, in our current experiment we used only phoneme and syllable contexts (from phonetic to /C/ incl.).

We used a testing set of 100 recordings from the Roger database. Each utterance was synthesized with original pitch using the TEMPO method of [18], and pitch encoding variants with three different approximation orders: 2, 3 and 4. In addition, we investigated different syllabification methods.

3.2. Results

3.2.1. Naturalness

Naturalness of the proposed pitch encoding method was evaluated both objectively and subjectively. The aim was to capture speech quality variations based on different syllabification methods and approximation orders.

First, we performed an objective evaluation. We used the common industry standard ITU-T recommendation P.862.2 (11/2005): Perceptual Evaluation of Speech Quality (PESQ). The PESQ measure is one of the most complex to compute and is the one recommended by ITU-T for speech quality assessment of narrow-band speech codecs. We downsampled all testing examples to 8 kHz, and used the natural speech as reference.

Fig. 2 shows the objective evaluation PESQ-MOS results. A t-test confirms that the differences between all evaluated groups are statistically insignificant ($p > 0.05$).

To confirm objective evaluation results, we performed a subjective evaluation using the Degradation Category Rating (DCR) procedure [19] quantifying the Degradation Mean Opinion Score (DMOS). Ten listeners, members of Idiap speech group, were asked to rate the degradation of synthetic signals (the second of each pair) compared with reference signals (the first of each pair) based on their overall perception. We asked listeners to focus on naturalness, especially on naturalness of the pitch contours, rather than on voice quality degradations. The synthesized speech with original pitch contour was selected as a reference signal in the test. Listeners had to describe degradation within the following five categories of intonation degradation: 1. very annoying, 2. annoying, 3. slightly annoying, 4. audible but not annoying, and 5. inaudible. The test corpus consisted of 8 sentences, randomly chosen from the Roger database, of at least 3 seconds duration. Listeners rated 9 versions of each sentence (three syllabification methods and 3 approximation orders).

Fig. 3 shows the subjective evaluation DMOS results. A t-test also confirms that the differences between all evaluated groups are statistically insignificant ($p > 0.05$). That indicates that even encoding with weak syllable boundaries (estimated with a simple phonetic syllabification method) and the lowest approximation order, performs almost equally well as manual syllabification and the highest approximation order. Results demonstrate our hypothesis that selection of a syllabification method does not impact the naturalness of generated speech with encoded pitch contours.

Fig. 4 shows an illustrative sample using third-order polynomials [2]. Although the syllable boundaries, and even the number of syllables vary, the generated pitch contours are visually almost identical. Based on the contour onset and offset detection, several voiced frames are decoded as unvoiced frames, for example the first few frames within the syllable at 1–1.5s in Fig. 4. The underlying assumption is that these frames are aperiodical endings of the previous phoneme because the syllable boundaries are given by the phonetic recognizer in a real speech coder.

3.2.2. Bit rates

Table 2: *Estimated bits per second (bps) with different syllabification methods and approximation orders.*

Order	Textual [bps]	Phonetic [bps]	Manual [bps]
2	34.6	32.2	35.1
3	43.3	40.3	43.9
4	51.9	48.3	52.7

The unvoiced syllables do not require pitch encoding. For the voiced syllables, which have more than 3 voiced segments, we need to transfer the pitch contour parameters and the contour onset information. Contour parameters consist of polynomial coefficients, each of which are stored using 2 bytes. Onset

[2] Voice samples are available at `http://www.idiap.ch/project/recod/demo/pitch-encoding`.

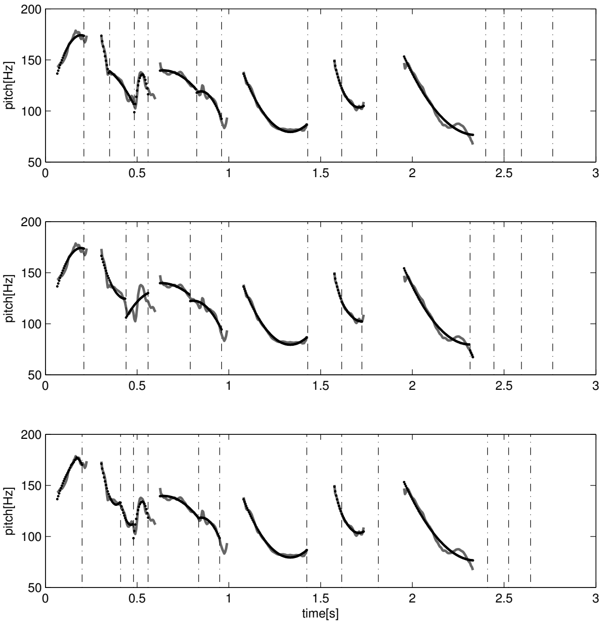

Figure 4: *Illustrative sample of different syllabifications. The grey line and dotted line respectively denote the original pitch and the synthesized pitch using syllable-based third-order polynomial. From top to bottom: syllabification obtained from text, phoneme and manually corrected label.*

information consists of the onset and offset time of the pitch contour in term of frame index, each of which are stored using 2 bytes. For instance, the pitch contour of a voiced syllable encoded by third order polynomials are transmitted using 10 bytes. Hence, the average bit rate is estimated as the total number of bits of the sentence divided by the length of it. Table 2 shows the estimated average pitch transmissions bps. Among them, the phonetic syllabification, as theoretically the closest form to a real speech coder, operates at the lowest bit rate.

Results confirm our second hypothesis that the pitch encoding with the lowest order approximation (linear) is sufficient.

4. Conclusions

We presented a recognition/synthesis very low bit rate speech coder with a combination of HMM-based phonetic vocoder, and a syllable-based pitch encoding technique. Pitch contours were modelled using the discrete (Legendre) orthogonal polynomials with different approximation orders.

The results show that high accuracy pitch contour reconstruction with negligible speech quality degradation is possible with the assumption that supra-segmental cues are present and can be extracted by the speech encoder, i.e. the phonetic recognizer. The proposed pitch encoding technique with phonetic syllabification and second order approximation operates on 32 bps, which allows it to be used for VLBR speech coding.

5. Acknowledgements

This research was partly supported under the RECOD project by armasuisse, the Procurement and Technology Center of the Swiss Federal Department of Defence, Civil Protection and Sport.

The work of Xingyu Na was partly supported by National Nature Science Foundation of China (NSFC): Grant No. 11161140319, “Ultra-realistic (acoustic) communication on next-generation internet”.

6. References

[1] I. 14496-3, *Information technology Coding of audio-visual objects, Part 3: Audio*. Geneva, Switzerland: ISO/IEC, 4th edition 2009.

[2] M. Schroeder and B. Atal, “Code-excited linear prediction(CELP): High-quality speech at very low bit rates,” in *Proc. of ICASSP*, vol. 10. IEEE, Apr. 1985, pp. 937–940.

[3] L. Nishiguchi, K. Iijima, and J. Matsumoto, “Harmonic vector excitation coding of speech at 2.0 kbps,” in *Speech Coding For Telecommunications Proceeding, 1997, 1997 IEEE Workshop on*. IEEE, Sep. 1997, pp. 39–40.

[4] M. Nishiguchi, “Harmonic vector excitation coding of speech,” *Acoustical Science and Technology*, vol. 27, no. 6, pp. 375–383, 2006. [Online]. Available: http://dx.doi.org/10.1250/ast.27.375

[5] G. V. Baudoin and F. El Chami, “Corpus based very low bit rate speech coding,” in *Proc. of ICASSP*, vol. 1. IEEE, Apr. 2003, pp. I–792–I–795 vol.1.

[6] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, “The HMM-based Speech Synthesis System Version 2.0,” in *Proc. of ISCA SSW6*, 2007.

[7] K. Tokuda, T. Masuko, J. Hiroi, T. Kobayashi, and T. Kitamura, “A very low bit rate speech coder using HMM-based speech recognition/synthesis techniques,” in *Proc. of ICASSP*, vol. 2. IEEE, May 1998, pp. 609–612 vol.2.

[8] H. Kawahara, “STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds,” *Acoust. Science and Technology*, vol. 27, no. 6, 2006.

[9] T. Eriksson and H.-G. Kang, “Pitch quantization in low bit-rate speech coding,” in *Proc. of ICASSP*, vol. 1. IEEE, Mar. 1999, pp. 489–492 vol.1.

[10] T. Hoshiya, S. Sako, H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitantura, “Improving the performance of HMM-based very low bit rate speech coding,” in *Proc. of ICASSP*, vol. 1. IEEE, Apr. 2003, pp. I–800–I–803 vol.1.

[11] M. Cernak, P. Motlicek, and P. N. Garner, “On the (Un)Importance of the Contextual Factors in HMM-based Speech Synthesis And Coding,” in *Proc. of ICASSP*, 2013.

[12] S.-H. Chen and Y.-R. Wang, “Vector quantization of pitch information in Mandarin speech,” *IEEE Trans. on Communications*, vol. 38, no. 9, pp. 1317–1320, September 1990.

[13] Y. Xu and Q. E. Wang, “Pitch targets and their realization: Evidence from Mandarin Chinese,” *Speech Communication*, vol. 33, no. 4, pp. 319–337, March 2001.

[14] Y. Xu and A. Wallace, “Multiple effects of consonant manner of articulation and intonation type on F_0 in English,” *J. Acoust. Soc. Am.*, vol. 115, no. 5, p. 2397, 2004.

[15] C.-C. Hsia, C.-H. Wu, and J.-Y. Wu, “Exploiting prosody hierarchy and dynamic features for pitch modeling and generation in HMM-based speech synthesis,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 1994–2003, 2010.

[16] N. Dehak, P. Dumouchel, and P. Kenny, “Modeling prosodic features with joint factor analysis for speaker verification,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2095–2103, September 2007.

[17] L. Saheer, J. Dines, and P. Garner, “Vocal tract length normalization for statistical parametric speech synthesis,” *IEEE Trans. on Audio, Speech, and Language Processing*, 2012.

[18] H. Kawahara, H. Katayose, A. de Cheveigné, and R. D. Patterson, “Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity,” in *Proc. of Eurospeech*, Budapest, Hungary, 1999.

[19] ITU-T Rec. P.800, “Methods for subjective determination of transmission quality,” (Geneva, Switzerland) 1996.