ON THE (UN)IMPORTANCE OF THE CONTEXTUAL FACTORS IN HMM-BASED SPEECH SYNTHESIS AND CODING

Milos Cernak, Petr Motlicek, Philip N. Garner

Idiap Research Institute, Martigny, Switzerland

{Milos.Cernak, Petr.Motlicek, Phil.Garner}@idiap.ch

ABSTRACT

This paper presents an evaluation of the contextual factors of HMM-based speech synthesis and coding systems. Two experimental setups are proposed that are based on successive context addition from phonetic to full-context. The aim was to investigate the impact of the individual contextual factors on the speech quality. In that sense important and unimportant (i.e., not having significant impact on speech quality, also called *weak*) contextual factors were identified. The results imply that in speech coding the improvement in quality can be achieved just with reconstruction of syllable contexts. The sentence and utterance contexts are unimportant on the decoder side, and it is not necessary to deal with them. Although in speech coding the wider context was not necessary, in speech synthesis current syllable and utterance contexts are more important over others (previous and next word/phrase contexts).

Index Terms— HMM speech synthesis, very low bit-rate speech coding

1. INTRODUCTION

Current HMM-based speech synthesis systems (HTS) [1] use contextual factors (also called *prosodic context*) based on phonetic, syllabic, word, phrase and utterance levels. The number of factors is high, and that of their combination increases exponentially. It is not trivial to estimate the impact of the factors on the quality of synthesized speech. A standard approach is to use as many factors as possible, and decision-tree based context clustering is applied to use the most influential factors for modeling of the speech sounds. Moreover, the contextual factors are language-dependent, and this makes their evaluation a challenging task.

The contextual factors play an important role also in very low bit-rate (VLBR) HMM-based speech coding. In HMM-based speech recognition/synthesis techniques, the encoder transmits only symbols (e.g., phonetic labels) together with the pitch and duration information [2]. The task of the speech decoder is to reconstruct the contextual labels; however, it might be also challenging as the original text is not available (and the text analyzer cannot be used for it as in the case of the HTS system). It has already been shown that, for example, accentual context might be reconstructed from quantisized F0 symbol sequence [3].

In this work, we were interested in the impact of the contextual factors on the quality of synthesized speech. To the authors' knowledge, the quantitative relation between contextual factors and speech quality of HTS has not been studied before, and the findings of this work might be interesting for researchers in both speech synthesis and coding fields. The experimental study [4] investigated the usefulness of types of high-level features that are commonly used in corpus-based TTS, however it provided only background analysis without more concrete conclusions. Some authors, e.g., [5], specified the important contextual factors by looking into questions of trained decision trees, however without studying their impact on speech quality as we do in our work. While past work on HMM-based VLBR techniques focused on Japanese systems [2][6], we investigated an English VLBR system, and the necessary minimal contextual factors for high-quality speech coding that need to be reconstructed from the symbol sequence. VLBR speech coding in Japanese is an easier task as the language has only 26 phones and a syllabic structure. So the objective of this paper is to achieve a performance as high as the one in [2], but in English.

In this paper we present the evaluation of the contextual factors in two experimental setups: (i) for an English HTS system, and (ii) for a VLBR speech coding system. The evaluation is based on modification of contextual labels, and as it is not tractable to evaluate all the combinations, we approximate it on gradually increasing the context from phonetic to full-context. By measuring the speech quality, we identified the important and unimportant (weak) contextual factors. In addition, we investigated an upper speech quality limit of HMM-based speech coding, showing that it can target a similar quality as standardized LPC-10 coding scheme, however with an order of magnitude lower bit rates.

The structure of the paper is as follows: the next section describes contextual factors used in HTS systems, section 3 describes the two experimental setups followed by results in section 4 and conclusions in section 5.

2. THE CONTEXTUAL FACTORS

While speech recognition systems usually use triphones in modeling speech sounds, current HTS systems use sound pentaphone phonetic context followed by a number of additional contextual factors. In both training and synthesis parts of HTS, so called full-context labels are required that annotate each modeling sound with all the contextual factors.

The contextual factors are based on phonetic, syllabic, word, phrase and utterance levels, and the most detailed description was introduced in [7]. These contextual factors are used in HTS for English, but a similar combination is used for Brasilian Portuguese and probably could be used for other Indo-European languages. The HTS for Japanese uses different factors, e.g., inflected forms of the words, the number of mora in accent phrases and introduces a breath group that contains accent phrases. In this study we focus on English and the Japanese contextual factors are beyond the scope of this work (however the experiments described in the next sections could be easily applied also to Japanese).

The factors investigated in this work are similar to those de-

scribed in [7]. The full-context labels use pentaphone phonetic context followed by the contextual factors and/or their combination (following HTS's label file format):

- @ current phoneme: forward and backward positions in current syllable
- /A/ previous syllable: stressed, accented and the number of phonemes
- /B/ current syllable: stressed, accented, forward/backward positions and number of stresses/accented syllables before/after in current phrase, and vowel within syllable
- /C/ next syllable: stressed, accented and the number of phonemes
- /D/ previous word: guess part-of-speech and number of syllables
- /E/ current word: guess part-of-speech, number of syllables, forward/backward positions and number of content words before/after in current phrase
- /F/ next word: guess part-of-speech and number of syllables
- /G/ previous phrase: number of syllables/words
- /H/ current phrase: number of syllables/words, forward/backward positions and TOBI endtone in utterance
- /I/ next phrase: number of syllables/words
- /J/ utterance: number of syllables, words and phrases

3. EXPERIMENTAL SETUP

In this section, we present two experimental setups for evaluating the importance of contextual factors. The first is an English HTS system described in section 3.1 and the second is a VLBR speech coder presented by section 3.2, based on recognition/synthesis technique with synthesis taken from the first experiment.

3.1. The English HTS system

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 MAP-VTLN parameter estimation of [8] to the Rogers 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 [9] was used to synthesize speech from the parameters generated using HTS.

Investigating the impact of contextual factors on speech quality, we used a testing set of 100 recordings from the Roger database. We synthesized the utterances using hand-corrected phone labels and hand-annotated full-context labels. Each utterance was synthesized with various contextual labels: starting with phonetic context only, and then gradually adding contextual factors from @ to /J/, evaluating the degradation of speech quality compared to the reference natural speech.

3.2. The VLBR speech coding

Speech coding on very low bit-rates can be achieved by 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

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

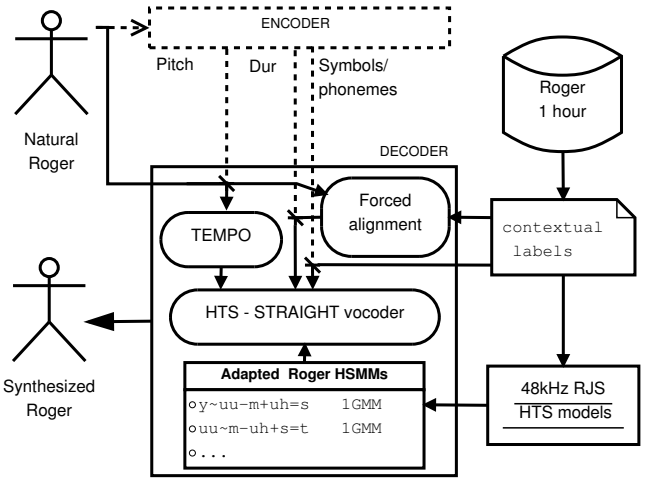

Fig. 1. *VLBR speech coding experimental setup with recognition-synthesis architecture, abstracting the encoder (dotted lines) and using the true inputs for the decoder. Symbol sequence is obtained from hand-labeled labels, state durations from the forced alignment, and the pitch from the TEMPO tool.*

such as pitch and duration of the symbols is required to recover the original prosody. Such a system can operate on 100 bps (bits per second). Our previous experiments showed that it is beneficial to use a different set of HMMs for phoneme recognition and synthesis - a high-quality HTS voice. That significantly improves the overall quality of the coding system. For this experiment, as we focused on contextual factors on the decoder side and not on the complete VLBR system, we abstracted the encoder side, and used the true input to the decoder, i.e., symbol sequence from hand-labeled labels, state durations from the forced alignment against natural speech, and the pitch of the natural speech using the TEMPO method of [10]. The experimental setup is depicted in Fig. 1. Using the true input to the decoder with variable contexts could reveal the speech quality limitation (the upper bound) of such a VLBR architecture.

Similar to the previous experiment, we investigated the impact of contextual factors on speech coding qualities using the same testing set from the Roger database. Original recordings were the reference speech and, we used various contextual labels: starting with phonetic context only, and then gradually adding contextual factors from @ to /J/, evaluating the degradation of speech quality compared to the reference. Speech quality achieved with full-context synthesis presents an oracle measurement.

In addition, we performed reference low bit rate 8kHz speech coding using the LPC-10 [11] at 2.4 kbps. LPC-10 is a telephony speech encoding standard developed by the United States Department of Defense and later by NATO. We stress that LPC-10 is no longer representative of the state of the art, and works with a significantly higher bit-rate than our HTS system. The intention is merely to compare with a widely available codec with well documented characteristics and familiar to the speech coding community.

4. RESULTS

Both experimental setups used the voice described in Section 3.1. First, we present subjective evaluation of speech synthesis in Section 4.1. In speech synthesis both durations and F0 were synthesized

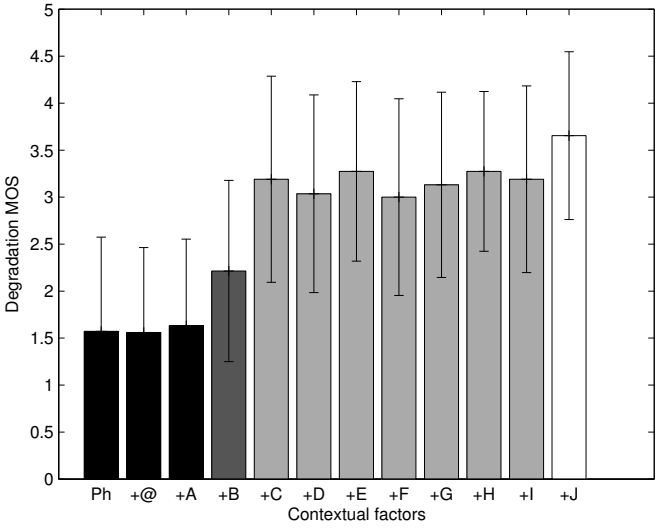

Fig. 2. *Subjective evaluation results for HTS synthesis with various contexts: phonetic context only Ph, and gradually added contexts from $+@$ to $+/J/$ (full-context).*

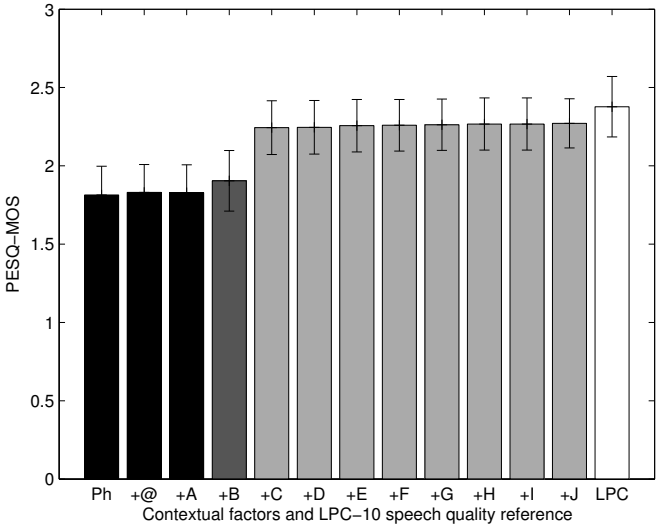

Fig. 3. *Objective evaluation results for speech coding: HTS synthesis with various contexts: phonetic context only Ph, gradually added contexts from $+@$ to $+/J/$, and LPC-10 system.*

from the trained HSMMs . For speech coding we present standard objective evaluation results in Section 4.2. Here, forced state alignment and original F0 were used instead of duration and F0 HSMMs.

4.1. Speech Synthesis

Speech quality in the synthesis experiment was evaluated subjectively using the Degradation Category Rating (DCR) procedure [12] quantifying the Degradation Mean Opinion Score (DMOS). This method provides a quality scale of a high resolution, due to comparison of a distorted (synthesized) signal with a reference (natural) signal. The aim was to capture speech quality variations based on the contextual factors which are used.

Fourteen listeners 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. According to the DCR procedure, it is not fair to build a pair associating two synthetic signals since it would have implied that the first synthetic signal outclasses perception of the second one. Therefore natural speech was selected as a reference signal in the test. Listeners had to describe degradation within the following five categories: 1. very annoying, 2. annoying, 3. slightly annoying, 4. audible but not annoying, and 5. inaudible. The test corpus consisted of 6 sentences, randomly chosen from the Roger's database, of at least 2 seconds durations. Listeners rated 12 versions of each sentence that was synthesized with different contextual factors.

Fig. 2 shows the subjective evaluation DMOS results. A t-test confirms that the differences between each group of the contextual factors are significant ($p < 0.05$). The factors in the figure are grouped with the same color according to this significance. We notice successive improvements of quality in current syllable context (/B/), word and phrase contexts (/C/-/I/), and utterance context (/J/).

A detailed analysis of relative contributions of contextual factors to overall speech quality is shown in Fig. 4 (a). Phonetic context /Ph/ contributes to overall quality by 43%, and adding phoneme @ and previous syllable /A/ contexts has no impact on quality. A relative improvement of 17% is obtained by current syllable context

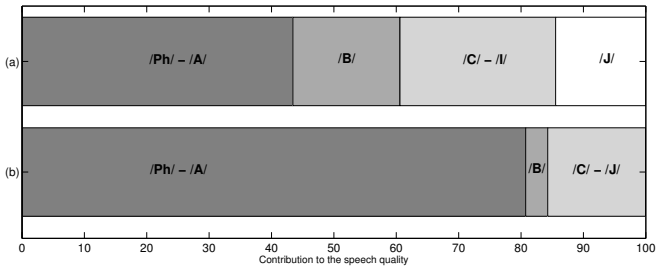

Fig. 4. *The relative contribution of the contextual factors to the speech quality for (a) HMM speech synthesis and (b) HMM speech coding.*

/B/. Adding next syllable /C/ context improves the quality by 25%, but further extending the context on word and phrase levels does not change the quality. Next, a final improvement 15% is achieved by adding the utterance context /J/. Overall, it seems that current contexts are more important than previous and forward contexts. It is seen from the /C/-/I/ group of the Fig. 2, where current word /E/ and current phrase /H/ contexts produce higher speech quality, although this impact is insignificant.

4.2. Speech Coding

We used an objective measure for evaluating speech quality in this coding experiment. As LPC-10 is the telephony standard, all the encoded examples were thus downsampled to 8kHz, and evaluated with 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 narrowband speech codecs.

Fig. 3 shows the objective evaluation results of PESQ-MOS, predicted subjective mean opinion scores (MOS) , calculated for the reference natural speech, and the evaluated coding schemes. As the VLBR decoder used the HTS voice from the synthesis experiment,

speech quality improvements shape the subjective evaluation already presented in the previous section. The lower scale of speech coding quality might be partly caused by different bandwidths used: listeners evaluated 48 kHz speech, while PESQ measured 8 kHz speech. A t-test confirms that the differences between each group of contextual factors are significant ($p < 0.05$). The factors in the figure are also grouped according to the significance. Unlike Fig. 2, the grouping of the contextual factors is different. While in speech synthesis the wider context was important, in speech coding the context wider than next syllable seems to be unimportant (i.e., having no significant impact on the quality of encoded speech). The most important factors are current /B/ and next /C/ syllables.

A detailed analysis of relative contributions of contextual factors to overall speech quality is shown in Fig. 4 (b). Phonetic context /Ph/ contributes to overall quality by 80%, and a relative improvement of 4% is obtained by adding current syllable context /B/. Adding next syllable /C/ context further improves relatively the quality by 16%. The results show that the improvement in speech quality can be achieved just with reconstruction of syllable context. The sentence and utterance contexts are unimportant on the decoder side, and it is not necessary to deal with them.

Another interesting result is a comparison of the the upper bound quality with the referenced system. The upper bound of the speech quality of the proposed VLBR coding scheme can achieve similar objective quality as the LPC-10 system, however with estimated bit rates around only 100 bps. While LPC-10 operates at 2.4 kpbs, its VQ version [13] with similar quality operates at 800 bps; we still target significant decrease in bit rates. The English VLBR coding system with syllable context reconstruction can achieve the same high-quality speech coding as published for Japanese[2].

5. CONCLUSIONS

We presented two experiments with successive context addition for HMM-based speech synthesis and coding. In the VLBR speech coding we identified the minimal context to be reconstructed at the decoder site, maintaining the speech quality.We also showed that the upper bound of recognition-synthesis based VLBR speech coding can approach the quality of industry standard LPC-10 and a similar system for Japanese. Further work has to be done on speaker-independent multilingual coding systems.

Previous work on contextual factors in HMM-based speech synthesis usually determined the factors for new languages, mainly aiming to improve prosody. In our work we presented a detailed analysis of relative speech quality improvements by each contextual factor usually used for all Indo-European languages. The work on Castilian Spanish [5] reported that accentual group of phonemes that incorporates syllable influence together with syllable start flags were presented in each trained decision tree. We confirm that the syllable context (/B/ and /C/) belongs to the most important contextual factors. We also quantitatively confirm the unimportance (or a weak importance) of the word-level contextual factors, as reported by [14].

We speculate that our other finding, that current factors play a more important role than others, could imply that further improvement can be achieved by additional improvement (e.g., using the adaptive context training) of the current syllable, word, phrase and utterance contexts. In the future we plan to investigate if removing unimportant contextual factors could help to alleviate the over-smoothing effect introduced by decision tree clustering.

[2]Running PESQ for the Japanese examples available from `http://www.sp.nitech.ac.jp/~tokuda/HTS_demo/hmm_vocoder/index.html` [2], proposed method 3 (146 bit/s) has 2.18 MOS while our average MOS prediction with +/C/ context achieves 2.24.

6. ACKNOWLEDGEMENTS

The authors would like to thank Junichi Yamagishi and Lakshmi Saheer for providing us with the HTS models. 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.

7. REFERENCES

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

[2] 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*. May 1998, vol. 2, pp. 609–612 vol.2, IEEE.

[3] T. Nose and T. Kobayashi, "Very low bit-rate F0 coding for phonetic vocoder using MSD-HMM with quantized F0 context," in *Proc. of ICASSP*. May 2011, pp. 5236–5239, IEEE.

[4] Oliver Watts, Junichi Yamagishi, and Simon King, "The role of higher-level linguistic features in HMM-based speech synthesis," in *Proc. of Interspeech*, 2010, pp. 841–844.

[5] X. Gonzalvo, J.C. Socoró, I. Iriondo, C. Monzo, and E. Martínez, "Linguistic and Mixed Excitation Improvements on a HMM-based speech synthesis for Castilian Spanish," in *6th ISCA Workshop on Speech Synthesis (SSW-6)*, 2007.

[6] 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*. Apr. 2003, vol. 1, pp. I–800–I–803 vol.1, IEEE.

[7] K. Tokuda, Heiga Zen, and A. W. Black, "An HMM-based speech synthesis system applied to English," in *Speech Synthesis, 2002. Proceedings of 2002 IEEE Workshop on*. 2002, pp. 227–230, IEEE.

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

[9] 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.

[10] 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.

[11] T.E. Tremain, "The Government Standard Linear Predictive Coding Algorithm: LPC-10," *Speech Technology*, pp. 40–49, Apr. 1982.

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

[13] X. Wang and C.-C. Jay Kuo, "An 800 bps VQbased LPC voice coder," *J. Acoust. Soc. Am.*, vol. 103, no. 5, pp. 2778, 1998.

[14] K. Yu, F. Mairesse, and S. Young, "Word-level emphasis modelling in HMM-based speech synthesis," in *Proc. of ICASSP*, 2010, pp. 4238–4241.