Chapter 13
Exploiting Alternatives for Text-To-Speech Synthesis: From Machine to Human

Nicolas Obin, Christophe Veaux and Pierre Lanchantin

Abstract The absence of alternatives/variants is a dramatical limitation of text-to-speech (TTS) synthesis compared to the variety of human speech. This chapter introduces the use of speech alternatives/variants in order to improve TTS synthesis systems. Speech alternatives denote the variety of possibilities that a speaker has to pronounce a sentence—depending on linguistic constraints, specific strategies of the speaker, speaking style, and pragmatic constraints. During the training, symbolic and acoustic characteristics of a unit-selection speech synthesis system are statistically modelled with context-dependent parametric models (Gaussian mixture models (GMMs)/hidden Markov models (HMMs)). During the synthesis, symbolic and acoustic alternatives are exploited using a Generalized Viterbi Algorithm (GVA) to determine the sequence of speech units used for the synthesis. Objective and subjective evaluations support evidence that the use of speech alternatives significantly improves speech synthesis over conventional speech synthesis systems. Moreover, speech alternatives can also be used to vary the speech synthesis for a given text. The proposed method can easily be extended to HMM-based speech synthesis.

13.1 Introduction

Today, speech synthesis systems (unit selection (Hunt and Black 1996), HMM-based (Zen et al. 2009)) are able to produce natural synthetic speech from text. Over the last decade, research has mainly focused on the modelling of speech

N. Obin (✉)
IRCAM, UMR STMS IRCAM-CNRS-UPMC,
Paris, France
e-mail: Nicolas.Obin@ircam.fr

C. Veaux
Centre for Speech Technology Research, The University of Edinburgh, Edinburgh, UK
e-mail: cveaux@inf.ed.ac.uk

P. Lanchantin
Department of Engineering, Cambridge University, Cambridge, UK
e-mail: pkl27@cam.ac.uk

K. Hirose, J. Tao (eds.), *Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis,* Prosody, Phonology and Phonetics,
DOI 10.1007/978-3-662-45258-5_13

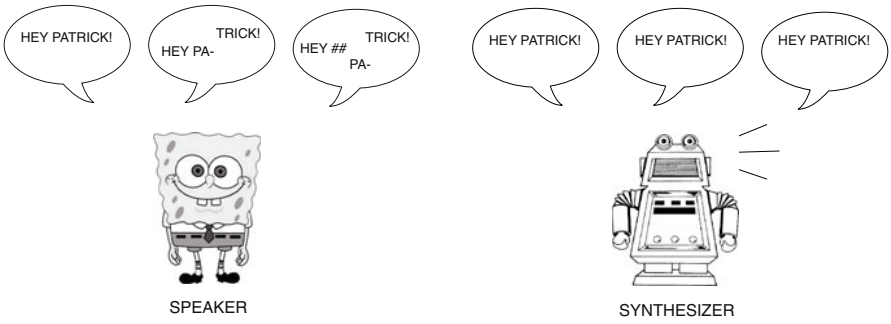

Fig. 13.1 Illustration of speech alternatives: human vs. machine

prosody—"the music of speech" (accent/phrasing, intonation/rhythm)—for text-to-speech (TTS) synthesis. Among them, GMMs/HMMs (Gaussian mixture models and hidden Markov models) are today the most popular methods used to model speech prosody. In particular, the modelling of speech prosody has gradually and durably moved from short-time representations ("frame-by-frame": Yoshimura et al. 1999; Zen et al. 2004; Tokuda et al. 2003; Toda and Tokuda 2007; Yan et al. 2009) to the use of large-time representations (Gao et al. 2008; Latorre and Akamine 2008; Qian et al. 2009; Obin et al. 2011b)). Also, recent researches tend to introduce deep architecture systems to model more efficiently the complexity of speech (deep neural networks (Zen et al. 2013)). However, current speech synthesis systems still suffer from a number of limitations, which consequence into the fact that the synthetic speech does not totally sound as "human". In particular, the absence of alternatives/variants in the synthetic speech is a dramatical limitation compared to the variety of human speech (see Fig. 13.1 for illustration): for a given text, the speech synthesis system will always produce exactly the same synthetic speech.

A human speaker can use a variety of alternatives/variants to pronounce a text. This variety may induce variations in the symbolic (prosodic event: accent, phrasing) and acoustic (prosody: prosodic contour; segmental: articulation, co-articulation) speech characteristics. These alternatives depend on linguistic constraints, specific strategies of the speaker, speaking style, and pragmatic constraints. Current speech synthesis systems do not exploit this variety during statistical modelling or synthesis. During the training, the symbolic and acoustic speech characteristics are usually estimated with a single normal distribution which is assumed to correspond with a single strategy of the speaker. During the synthesis, the sequence of symbolic and acoustic speech characteristics are entirely determined by the sequence of linguistic characteristics associated with the sentence—the *most-likely* sequence.

In real-world speech synthesis applications (e.g. announcement, storytelling, or interactive speech systems), expressive speech is required (Obin et al. 2011a; Obin 2011). The use of speech alternatives in speech synthesis may substantially improve speech synthesis (Bulyko and Ostendorf 2001), and fill the gap of the machine to the human. First, alternatives can be used to provide a variety of speech candidates

that may be exploited to vary the speech synthesized for a given sentence. Second, alternatives can also be advantageously used as a relaxed constraint for the determination of the sequence of speech units to improve the quality of the synthesized speech. For instance, the use of a symbolic alternative (e.g. insertion/deletion of a pause) may conduct to a significantly improved sequence of speech units.

This chapter addresses the use of speech alternatives to improve the quality and the variety of speech synthesis. The proposed speech synthesis system (IRCAMTTS) is based on unit selection, and uses various context-dependent parametric models to represent the symbolic/acoustic characteristics of speech prosody (GMMs/HMMs). During the synthesis, symbolic and acoustic alternatives are exploited using a generalized Viterbi algorithm (GVA) (Hashimoto 1987). First, a GVA is used to determine a set of symbolic candidates, corresponding to the $K_{\text{symb.}}$ sequences of symbolic characteristics, in order to enrich the further selection of speech units. For each symbolic candidate, a GVA is then used to determine the $K_{\text{acou.}}$ sequences of speech units under the joint constraint of segmental and speech prosody characteristics. Finally, the optimal sequence of speech units is determined so as to maximize the cumulative symbolic/acoustic likelihood. Alternatively, the introduction of alternatives allows to vary the speech synthesis by selecting one of the K most likely speech sequences instead of the most likely one. The proposed method can easily be extended to HMM-based speech synthesis.

The speech synthesis system used for the study is presented in Sect. 13.2. The use of speech alternatives during the synthesis, and the GVA are introduced in Sect. 13.3. The proposed method is compared to various configurations of the speech synthesis system (modelling of speech prosody, use of speech alternatives), and validated with objective and subjective experiments in Sect. 13.4.

13.2 Speech Synthesis System

Unit-selection speech synthesis is based on the optimal selection of a sequence of speech units that corresponds to the sequence of linguistics characteristics derived from the text to synthesize. The optimal sequence of speech units is generally determined so as to minimize an objective function usually defined in terms of concatenation and target acoustic costs. Additional information (e.g. prosodic events—ToBI labels) can also be derived from the text to enrich the description used for unit selection.

The optimal sequence of speech units $\overline{\mathbf{u}}$ can be determined by jointly maximizing the symbolic/acoustic likelihood of the sequence of speech units $\mathbf{u} = [u_1, \ldots, u_N]$ conditionally to the sequence of linguistic characteristics $\mathbf{c} = [c_1, \ldots, c_N]$:

$$\overline{\mathbf{u}} = \underset{\mathbf{u}}{\text{argmax}}\ \text{p}(\mathbf{O}(\mathbf{u})|\mathbf{c}) \tag{13.1}$$

where $\mathbf{O}(\mathbf{u}) = [\mathbf{O}_{\text{symb.}}(\mathbf{u}),\ \mathbf{O}_{\text{acou.}}(\mathbf{u})]$ denotes the symbolic and acoustic characteristics associated with the sequence of speech units $\mathbf{u}$.

A suboptimal solution to this equation is usually obtained by factorizing the symbolic/acoustic characteristics:

$$\overline{\mathbf{u}}_{\text{symb.}} = \underset{\mathbf{u}_{\text{symb.}}}{\text{argmax}}\ \text{p}(\mathbf{O}_{\text{symb.}}(\mathbf{u}_{\text{symb.}})|\mathbf{c}) \tag{13.2}$$

$$\overline{\mathbf{u}}_{\text{acou.}} = \underset{\mathbf{u}}{\text{argmax}}\ \text{p}(\mathbf{O}_{\text{acou.}}(\mathbf{u}_{\text{acou.}})|\mathbf{c}, \overline{\mathbf{u}}_{\text{symb.}}) \tag{13.3}$$

where $\mathbf{u}_{\text{symb.}}$ is the symbolic sequence of speech units (typically, a sequence of prosodic events, e.g. accent and phrasing), and $\mathbf{u}_{\text{acou.}}$ is the acoustic sequence of speech units (i.e. a sequence of speech units for unit-selection and a sequence of speech parameters for HMM-based speech synthesis). This acoustic sequence of speech units represents the short- (source/filter) and long-term (prosody: F0, duration) variations of speech over various units (e.g. phone, syllable, and phrase).

In other words, the symbolic sequence of speech units $\overline{\mathbf{u}}_{\text{symb.}}$ is first determined, and then used for the selection of acoustic speech units $\overline{\mathbf{u}}_{\text{acou.}}$. This conventional approach suffers from the following limitations:

1. Symbolic and acoustic modelling are processed separately during training and synthesis, which remain suboptimal and may degrade the quality of the synthesized speech.
2. A single sequence of speech units is determined during synthesis, while the use of alternatives enlarges the number of speech candidates available, and then improves the quality of the synthesized speech.

To overcome these limitations, the ideal solution is: the joint symbolic/acoustic modelling in order to determine the sequence of speech units that is globally optimal (Eq. 13.1); and the exploitation of speech alternatives in order to enrich the search for the optimal sequence of speech units. The present study only addresses the use of symbolic/acoustic alternatives for speech synthesis. In the present study, symbolic alternatives are used to determine a set of symbolic candidates $\overline{\mathbf{u}}_{\text{symb.}}$ so as to enrich the further selection of speech units (Eq. 13.2). For each symbolic candidate, the sequence of acoustic speech units $\overline{\mathbf{u}}_{\text{acou.}}$ is determined based on a relaxed-constraint search using acoustic alternatives (Eq. 13.3). Finally, the optimal sequence of speech units $\overline{\mathbf{u}}$ is determined so as to maximize the cumulative likelihood of the symbolic/acoustic sequences.

The use of symbolic/acoustic alternatives requires adequate statistical models that explicitly describe alternatives, and a dynamic selection algorithm that can manage these alternatives during speech synthesis. Symbolic and acoustic models used for this study are briefly introduced in Sects. 13.2.1 and 13.2.2. Then, the dynamic selection algorithm used for unit selection is described in Sect. 13.3.

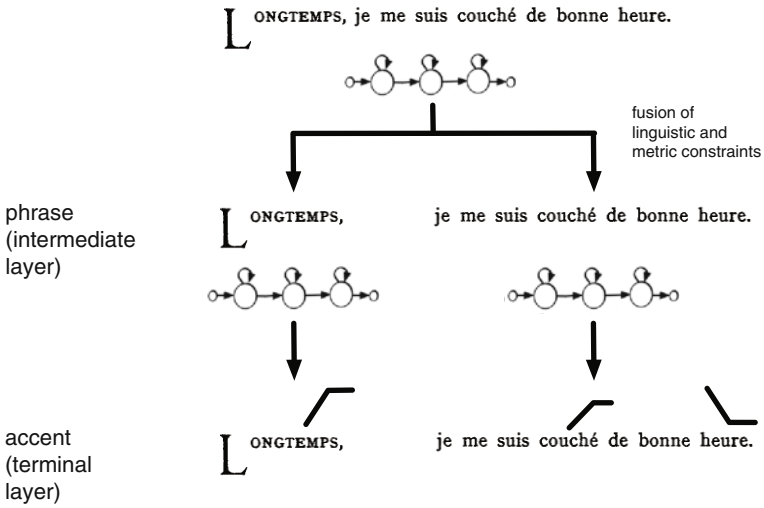

Fig. 13.2 Illustration of the HHMM symbolic modelling of speech prosody for the sentence: "*Longtemps, je me suis couché de bonne heure*" ("*For a long time I used to go to bed early*"). The *intermediate layer* illustrates the segmentation of a text into phrases. The *terminal layer* illustrates the assignment of accents

13.2.1 Symbolic Modelling

The prosodic events (accent and phrasing) are modelled by a statistical model based on HMMs (Black and Taylor 1994; Atterer and Klein 2002; Ingulfen et al. 2005; Obin et al. 2010a, 2010b; Parlikar and Black 2012; Parlikar and Black 2013). A hierarchical HMM (HHMM) is used to assign the prosodic structure of a text: the root layer represents the text, each intermediate layer a phrase (here, intermediate phrase and phrase), and the final layer the sequence of accents. For each intermediate layer, a segmental HMM and information fusion are used to combine the linguistic and metric constraints (length of a phrase) for the segmentation of a text into phrases (Ostendorf and Veilleux 1994; Schmid and Atterer 2004; Bell et al. 2006; Obin et al. 2011c). An illustration of the HHMM for the symbolic modelling of speech prosody is presented in Fig. 13.2.

13.2.2 Acoustic Modelling

The acoustic (short- and long-term) models are based on context-dependent GMMs (cf. Veaux et al. 2010; Veaux and Rodet 2011, for a detailed description). Three different observation units (phone, syllable, and phrase) are considered, and separate GMMs are trained for each of these units. The model associated with the phone unit is merely a reformulation of the target and concatenation costs traditionally used in

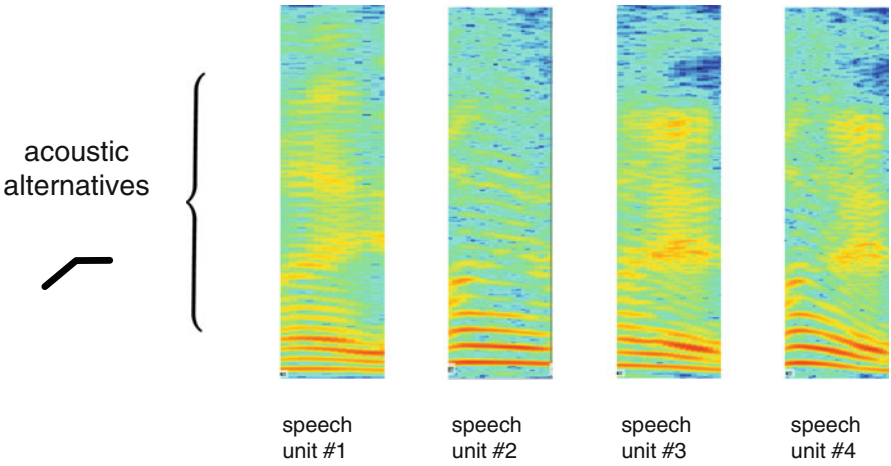

Fig. 13.3 Illustration of acoustic alternatives for a given symbolic unit

unit-selection speech synthesis (Hunt and Black 1996). The other models are used to represent the local variation of prosodic contours (*F0* and durations) over the syllables and the major prosodic phrases, respectively. The use of GMMs allows to capture prosodic alternatives associated with each of the considered units (Fig. 13.3).

13.3 Exploiting Alternatives

The main idea of the contribution is to exploit the symbolic/acoustic alternatives observed in human speech. Fig. 13.4 illustrates the integration of symbolic/acoustic alternatives for speech synthesis. The remainder of this section presents the details of the generalized Viterbi search to exploit symbolic/acoustic alternatives for TTS synthesis.

In a conventional synthesizer, the search for the optimal sequence of speech units (Eq. 13.1) is decomposed in two separate optimisation problems (Eqs. 13.2 and 13.3). These two equations are generally solved using the Viterbi algorithm. This algorithm defines a lattice whose states at each time t are the N candidate units. At each time t, the Viterbi algorithm considers N lists of competing paths, each list being associated to one of the N states. Then, for each list, only one survivor path is selected for further extension. Therefore the Viterbi algorithm can be described as a N-list 1-survivor (N,1) algorithm. The GVA (Hashimoto 1987) consists in a twofold relaxation of the path selection.

- First, more than one survivor path can be retained for each list.
- Second, a list of competing paths can encompass more than one state.

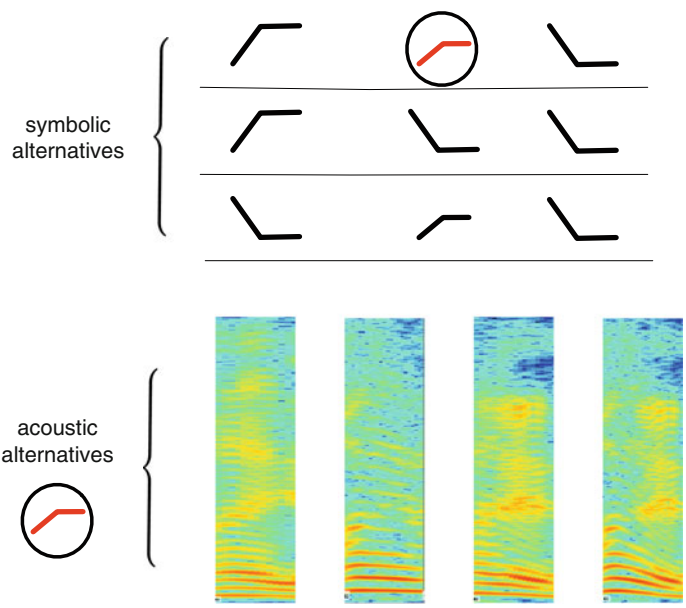

Fig. 13.4 Illustration of symbolic/acoustic alternatives for text-to-speech synthesis. The *top* of the figure presents three symbolic alternative sequences to a given input text. The *bottom* of the figure presents four acoustic alternatives to the symbolic event *circled* on *top*. Fundamentally, each text has symbolic alternative sequence, and each symbolic alternative sequence has acoustic alternative sequences

An illustration of this approach is given in Fig. 13.5, which shows that the GVA can retain survivor paths that would otherwise be merged by the classical Viterbi algorithm. Thus, the GVA can keep track of several symbolic/prosodic alternatives until the final decision is made.

In this study, the GVA is first used to determine a set of symbolic candidates corresponding to the $K_{\text{symb.}}$ most-likely sequences of symbolic characteristics, in order to enrich the further selection of speech units. For each symbolic candidate, a GVA is then used to determine the $K_{\text{acou.}}$ most-likely sequences of speech units under the joint constraint of segmental characteristics (phone model) and prosody (syllable and phrase models). Finally, the optimal sequence of speech units is determined so as to maximize the cumulative symbolic/acoustic likelihood.

13.4 Experiments

Objective and subjective experiments were conducted to address the use of speech alternatives in speech synthesis, with comparison to a BASELINE (no explicit modelling of speech prosody, no use of speech alternatives) and CONVENTIONAL (explicit

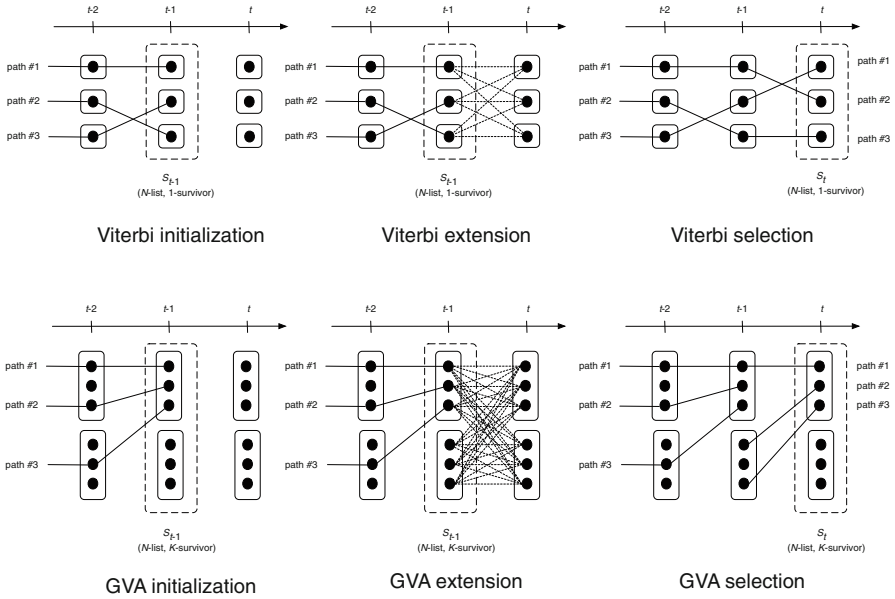

Fig. 13.5 Illustration of VITERBI SEARCH and GENERALIZED VITERBI SEARCH. The *boxes* represent the list of states among which the best S path is selected. For the VITERBI SEARCH, only one path is retained at all time, and only one survivor is retained during selection. For the GENERALIZED VITERBI SEARCH, K paths are retained at all time, and K survivors are retained during selection (alternative candidates, here $K = 3$). At all time, the GENERALIZED VITERBI SEARCH has a larger memory than the VITERBI SEARCH

Table 13.1 Description of TTS systems used for the evaluation. Parentheses denote the optional use of symbolic alternatives in the TTS system

	Symbolic	Acoustic	
	Alternatives	Prosody	Alternatives
BASELINE	(✓)	–	–
CONVENTIONAL	(✓)	Syllable/phrase	–
PROPOSED	(✓)	Syllable/phrase	✓

modelling of speech prosody, no use of speech alternatives) speech synthesis systems (Table 13.1). In addition, symbolic alternatives have been optionally used for each compared method to assess the relevancy of symbolic and acoustic alternatives separately.

13.4.1 Speech Material

The speech material used for the experiment is a 5-h French storytelling database interpreted by a professional actor, which was designed for expressive speech synthesis. The speech database comes with the following linguistic processing: orthographical transcription; surface syntactic parsing (POS and word class); manual speech segmentation into phonemes and syllables, and automatic labelling/segmentation of prosodic events/units (cf. Obin et al. 2010b for more details).

13.4.2 Objective Experiment

An objective experiment has been conducted to assess the relative contribution of speech prosody and symbolic/acoustic alternatives to the overall quality of the TTS system. In particular, a specific focus will be made on the use of symbolic/acoustic alternatives.

13.4.2.1 Procedure

The objective experiment has been conducted with 173 sentences of the fairy tale "*Le Petit Poucet*" ("*Tom Thumb*").
For this purpose, a *cumulative* log-likelihood has been defined as a weighted integration of the *partial* log-likelihoods (symbolic, acoustic). First, each partial log-likelihood is averaged over the utterance to be synthesized so as to normalize the variable number of observations used for the computation (e.g. phonemes, syllable, and prosodic phrase). Then, log-likehoods are normalized to ensure comparable contribution of each partial log-likelihood during the speech synthesis. Finally, the cumulative log-likelihood of a synthesized speech utterance is defined as follows:

$$LL = w_{\text{symbolic}} LL_{\text{symbolic}} + w_{\text{acoustic}} LL_{\text{acoustic}} \tag{13.4}$$

where LL_{symbolic} and LL_{acoustic} denote the partial log-likelihood associated with the sequence of symbolic and acoustic characteristics; and w_{symbolic} and w_{acoustic}, corresponding weights.
Finally, the optimal sequence of speech units is determined so as to maximize the cumulative log-likelihood of the symbolic/acoustic characteristics. In this study, weights were heuristically chosen as $w_{\text{symbolic}} = 1$, $w_{\text{phone}} = 1$, $w_{\text{syllable}} = 5$, and $w_{\text{phrase}} = 1$; 10 alternatives have been considered for the symbolic characteristics, and 50 alternatives for the selection of speech units.

13.4.2.2 Discussion

Cumulative likelihood obtained for the compared methods is presented in Fig. 13.6, with and without the use of symbolic alternatives. The PROPOSED method (modelling

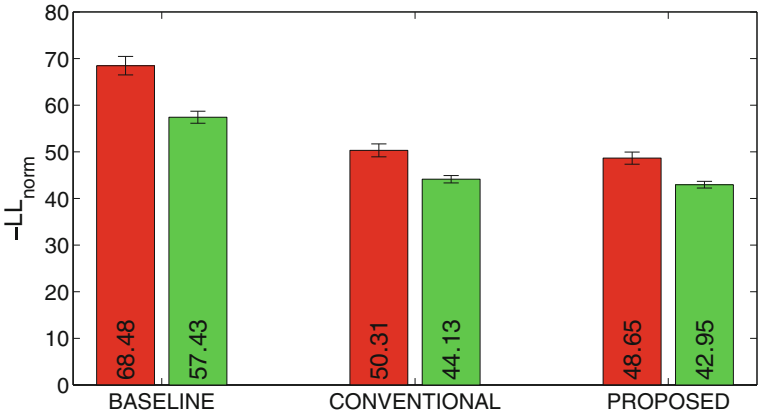

Fig. 13.6 Cumulative negative log-likelihood (mean and 95 % confidence interval) obtained for the compared TTS, without (*left*) and with (*right*) use of symbolic alternatives

of prosody, use of acoustic alternatives) moderately but significantly outperforms the CONVENTIONAL method (modelling of prosody, no use of acoustic alternatives); and dramatically outperforms the BASELINE method. In addition, the use of symbolic alternatives conducts to a significant improvement regardless of the method considered. Finally, the optimal synthesis is obtained for the combination of symbolic/acoustic alternatives with the modelling of speech prosody.

For further investigation, partial likelihoods obtained for the compared methods are presented in Fig. 13.7, with and without the use of symbolic alternatives. Not surprisingly, the modelling of speech prosody (syllable/phrase) successfully constraints the selection of speech units with adequate prosody, while this improvement comes with a slight degradation of the segmental characteristics (phone). The use of acoustic alternatives conducts to an improved speech prosody (significant over the syllable, not significant over the phrase) that comes with a slight degradation of the segmental characteristics (nonsignificant). This suggests that the phrase modelling (as described by Veaux and Rodet 2011) has partially failed to capture relevant variations, and that this model remains to be improved. Finally, symbolic alternatives are advantageously used to improve the prosody of the selected speech units, without a significant change in the segmental characteristics.

13.4.3 Subjective Experiment

A subjective experiment has been conducted to compare the quality of the BASELINE, CONVENTIONAL, and PROPOSED speech synthesis systems.

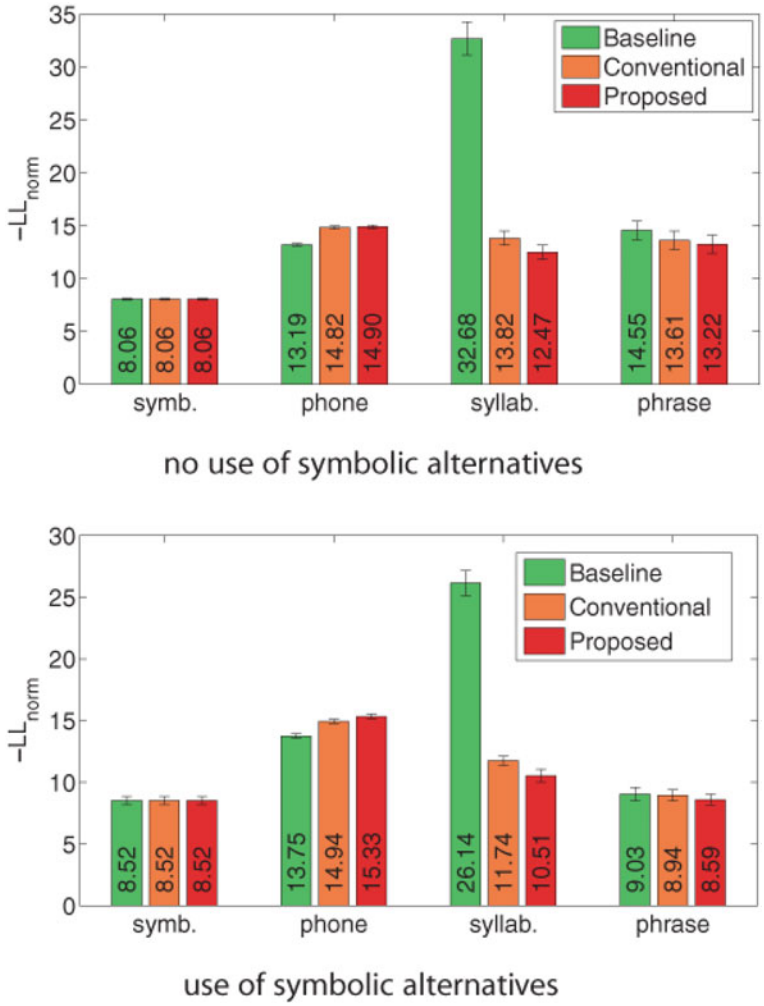

Fig. 13.7 Partial negative log-likelihoods (mean and 95 % confidence intervals) for the compared methods, with and without use of symbolic alternatives

13.4.3.1 Procedure

For this purpose, 11 sentences have been randomly selected from the fairy tale, and used to synthesize speech utterances with respect to the considered systems. Fifteen native French speakers have participated in the experiment. The experiment has been conducted according to a *crowdsourcing* technique using social networks. Pairs of synthesized speech utterances were randomly presented to the participants who were asked to attribute a preference score according to the *naturalness* of the

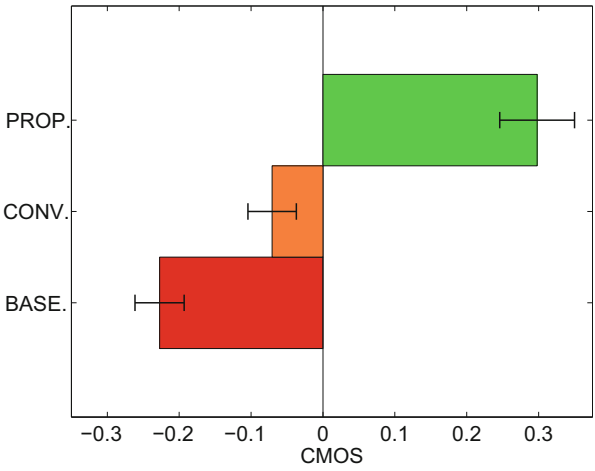

Fig. 13.8 CMOS (mean and 95 % confidence interval) obtained for the compared methods

speech utterances on the comparison mean opinion score (CMOS) scale. Participants were encouraged to use headphones.

13.4.3.2 Discussion

Figure 13.8 presents the CMOS obtained for the compared methods. The PROPOSED method is substantially preferred to other methods, which indicates that the use of symbolic/acoustic alternatives conducts to a qualitative improvement of the speech synthesized over all other systems. Then, CONVENTIONAL method is fairly preferred to the BASELINE method, which confirms that the integration of speech prosody also improves the quality of speech synthesis over the BASELINE system (cf. observation partially reported in Veaux and Rodet 2011).

13.5 Conclusion

In this chapter, the use of speech alternatives/variants in the unit-selection speech synthesis has been introduced. Objective and subjective experiments support the evidence that the use of speech alternatives qualitatively improves speech synthesis over conventional speech synthesis systems. The proposed method can easily be extended to HMM-based speech synthesis. In further studies, the use of speech alternatives will be integrated into a joint modelling of symbolic/acoustic characteristics so as to improve the consistency of the selected symbolic/acoustic sequence of speech units. Moreover, speech alternatives will further be used to vary the speech synthesis for a given text.

References

Atterer, M., and E. Klein. 2002. Integrating linguistic and performance-based constraints for assigning phrase breaks. In *International Conference on Computational Linguistics*, Taipei, Taiwan, 995–998.

Bell, P., T. Burrows, and P. Taylor. 2006. Adaptation of prosodic phrasing models. In *Speech Prosody*, Dresden, Germany.

Black, A., and P. Taylor. 1994. Assigning intonation elements and prosodic phrasing for English speech synthesis from high level linguistic input. In *International Conference on Spoken Language Processing*, Yokohama, Japan, 715–718.

Bulyko, I., and M. Ostendorf. 2001. Joint prosody prediction and unit selection for concatenative speech synthesis. In *International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, USA, 781–784.

Gao, B., Y. Qian, Z. Wu, and F. Soong. 2008. Duration refinement by jointly optimizing state and longer unit likelihood. In *Interspeech*, Brisbane, Australia, 2266–2269.

Hashimoto, T. 1987. A list-type reduced-constraint generalization of the Viterbi algorithm. *IEEE Transactions on Information Theory* 33 (6): 866–876.

Hunt, A., and A. Black. 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In *International Conference on Audio, Speech, and Signal Processing*, 373–376.

Ingulfen, T., T. Burrows, and S. Buchholz. 2005. Influence of syntax on prosodic boundary prediction. In *Interspeech*, Lisboa, Portugal, 1817–1820.

Latorre, J., and M. Akamine. 2008. Multilevel parametric-base F0 model for speech synthesis. In *Interspeech*, Brisbane, Australia, 2274–2277.

Obin, N. 2011. MeLos: Analysis and modelling of speech prosody and speaking style. PhD Thesis, Ircam - UPMC.

Obin, N., P. Lanchantin, A. Lacheret, and X. Rodet. 2010a. Towards improved HMM-based speech synthesis using high-level syntactical features. In *Speech Prosody*, Chicago, USA

Obin, N., A. Lacheret, and X. Rodet. 2010b. HMM-based prosodic structure model using rich linguistic context. In *Interspeech*, Makuhari, Japan, 1133–1136.

Obin, N., P. Lanchantin, A. Lacheret, and X. Rodet. 2011a. Discrete/continuous modelling of speaking style in HMM-based speech synthesis: Design and evaluation. In *Interspeech*, Florence, Italy, 2785–2788.

Obin, N., A. Lacheret, and X. Rodet. 2011b. Stylization and trajectory modelling of short and long term speech prosody variations. In *Interspeech*, Florence, Italy, 2029–2032.

Obin, N., P. Lanchantin, A. Lacheret, and X. Rodet. 2011c. Reformulating prosodic break model into segmental HMMs and information fusion. In *Interspeech*, Florence, Italy, 1829–1832.

Ostendorf, M., and N. Veilleux. 1994. A hierarchical stochastic model for automatic prediction of prosodic boundary location. *Journal of Computational Linguistics* 20 (1): 27–54.

Parlikar, A., and A. W. Black. 2012. Modeling pause-duration for style-specific speech synthesis. In *Interspeech*, Portland, Oregon, USA, 446–449.

Parlikar, A., and A. W. Black. 2013. Minimum error rate training for phrasing in speech synthesis. In *Speech Synthesis Workshop (SSW)*, Barcelona, Spain, 13–17.

Qian, Y., Z. Wu, and F. K. Soong. 2009. Improved prosody generation by maximizing joint likelihood of state and longer units. In *International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan, 3781–3784.

Schmid, H., and M. Atterer. 2004. New statistical methods for phrase break prediction. In *International Conference on Computational Linguistics*, Geneva, Switzerland, 659–665.

Toda, T., and K. Tokuda. 2007. A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. *IEICE Transactions on Information and Systems* 90 (5): 816–824.

Tokuda, K., H. Zen, and T. Kitamura. 2003. Trajectory modeling based on HMMs with the explicit relationship between static and dynamic features. In *European Conference on Speech Communication and Technology*, Geneva, Switzerland, 865–868.

Veaux, C., and X. Rodet. 2011. Prosodic control of unit-selection speech synthesis: A probabilistic approach. In *International Conference on Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, 5360–5363.

Veaux, C., P. Lanchantin, and X. Rodet. 2010. Joint prosodic and segmental unit selection for expressive speech synthesis. In *Speech Synthesis Workshop (SSW7)*, Kyoto, Japan, 323–327.

Yan, Z.-J., Y. Qian, and F. K. Soong. 2009. Rich context modeling for high quality HMM-based TTS. In *Interspeech*, Brighton, UK, 4025–4028.

Yoshimura, T., K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. 1999. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *European Conference on Speech Communication and Technology*, Budapest, Hungary, 2347–2350.

Zen, H., K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. 2004. Hidden semi-Markov model based speech synthesis. In *International Conference on Spoken Language Processing*, Jeju Island, Korea, 1397–1400.

Zen, H., K. Tokuda, and A. Black. 2009. Statistical parametric speech synthesis. *Speech Communication* 51 (11): 1039–1064.

Zen, A., A. Senior, and M. Schuster. 2013. Statistical parametric speech synthesis using deep neural networks. In *International Conference on Acoustics, Speech, and Signal Processing*, Vancouver, Canada, 7962–7966.