Making Sense of Variations: Introducing Alternatives in Speech Synthesis

Nicolas Obin[1], *Christophe Veaux*[2,†], *Pierre Lanchantin*[3,†]

[1] IRCAM-CNRS-9912-STMS, Paris, France
[2] Centre for Speech Technology Research, Edinburgh, UK
[3] Cambridge University Engineering Department, Cambridge, UK
nobin@ircam.fr

Abstract

This paper addresses the use of speech alternatives to enrich speech synthesis systems. Speech alternatives denote the variety of strategies that a speaker can use to pronounce a sentence - depending on pragmatic constraints, speaking style, and specific strategies of the speaker. During the training, symbolic and acoustic characteristics of a unit-selection speech synthesis system are statistically modelled with context-dependent parametric models (GMMs/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 supports evidence that the use of speech alternatives significantly improves speech synthesis over conventional speech synthesis systems.
Index Terms : speech synthesis, speech prosody, speech alternatives.

1. Introduction

A speaker has a variety of alternatives that may be likely used to pronounce a sentence. These alternatives depend on the speaking style, specific strategies of the speaker, pragmatic constraints, and eventually arbitrary choice of the speaker. This variability can be observed either in terms of symbolic (prosodic prominence, prosodic break) or acoustic (prosodic contour) speech characteristics. Current speech synthesis systems [1, 2] 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, story-telling, or interactive speech systems), expressive speech is required. However, current speech synthesis systems are often perceived as poorly natural due to the presence of speech artefacts and the absence of variety in the synthesized speech. The use of speech alternatives in speech synthesis may significantly improve both the *variety* and the *quality* of the synthesized speech. Firstly, alternatives can be used to provide a variety of speech candidates that may be exploited to vary the speech synthesized for a given sentence. Secondly, 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.

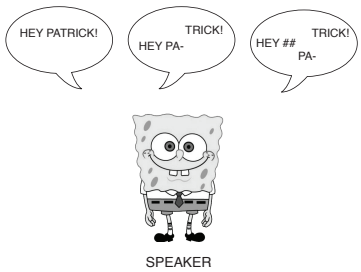


FIGURE 1 – Illustration of alternatives in speech prosody.

This paper 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) ([3]). First, a GVA is used to determine a set of symbolic candidates - corresponding to the K 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 optimal sequence 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 cumulated symbolic/acoustic likelihood.

The speech synthesis system used for the study is presented in section 2. The use of speech alternatives during the synthesis, and the GENERALIZED VITERBI ALGORITHM are introduced in section 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 evaluations in section 4.

2. Speech Synthesis System

Unit selection speech synthesis is based on the optimal selection of a sequence of speech units that corresponds to the

†. The present study has been conducted during the stay of authors in the sound analysis and synthesis department at IRCAM.

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 linguistic information (e.g., prosodic events -ToBI labels) can also be derived from the text to enrich the linguistic description used for unit selection.

Idealistically, 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 $\mathbf{c} = [c_1, \ldots, c_N]$:

$$\overline{\mathbf{u}} = \underset{\mathbf{u}}{\operatorname{argmax}}\ \mathrm{p}(\mathbf{O}(\mathbf{u})|\mathbf{c}) \quad (1)$$

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

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

$$\overline{\mathbf{u}}_{\mathrm{symb.}} = \underset{\mathbf{u}_{\mathrm{symb.}}}{\operatorname{argmax}}\ \mathrm{p}(\mathbf{O}_{\mathrm{symb.}}(\mathbf{u}_{\mathrm{symb.}})|\mathbf{c}) \quad (2)$$

$$\overline{\mathbf{u}}_{\mathrm{acou.}} = \underset{\mathbf{u}}{\operatorname{argmax}}\ \mathrm{p}(\mathbf{O}_{\mathrm{acou.}}(\mathbf{u}_{\mathrm{acou.}})|\mathbf{c}, \overline{\mathbf{u}}_{\mathrm{symb.}}) \quad (3)$$

In other words, the symbolic sequence of speech units (e.g., prosodic events) is first determined, and then used for the selection of acoustic speech units.

This conventional approach presents two main inconsistencies :

1. symbolic and acoustic modelling are processed separately during training and synthesis, which remain sub-optimal and may degrade the quality of the speech synthesized.
2. a single sequence of speech characteristics is determined for unit selection, while the use of symbolic/acoustic alternatives may improve the quality and the variety of the speech synthesized.

The optimal solution would consists of a joint symbolic/acoustic unit selection system combined with the integration of speech alternatives. For clarity, the present study will focus only on the use of symbolic/acoustic alternatives in unit selection speech synthesis. In the present study, symbolic alternatives are used to determine a set of symbolic candidates $\overline{\mathbf{u}}_{\mathrm{symb.}}$ so as to enrich the further selection of speech units (eq. (2)). For each symbolic candidate, the sequence of acoustic speech units $\overline{\mathbf{u}}_{\mathrm{acou.}}$ is determined based on a relaxed-constraint search using acoustic alternatives (eq. (3)). Finally, the optimal sequence of speech units $\overline{\mathbf{u}}$ is determined so as to maximize the cumulated symbolic/acoustic likelihood.

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 section 2.1 and 2.2. Then, the dynamic selection algorithm used for unit selection is described in section 3.

2.1. Symbolic Modelling

The symbolic modelling of prosodic events is a statistical model in which linguistic and metric constraints are combined - based on HMMs [4] and explicit modelling of the metric constraint (length of a prosodic unit) (cf. [5] for a detailed description). Additionally, information fusion is used for the optimal combination of linguistic and metric constraints. The prosodic events used cover accent and boundaries, associated with intermediate prosodic phrase and prosodic phrase. Prosodic phrases refer to speech segments that end with a prosodic prominence followed by a long pause ; intermediate prosodic phrases refer to syntactic chunks that end with a prosodic prominence.

2.2. Acoustic Modelling

In order to capture the natural speech prosody of a speaker, the acoustic and prosodic models are based on context-dependent GMMs (cf. [6] 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 unit-selection speech synthesis [2]. 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.

3. Exploiting Alternatives

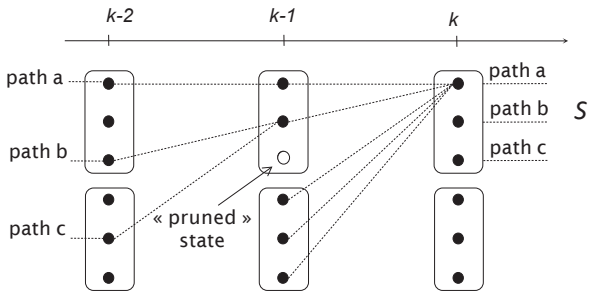


FIGURE 2 – Illustration of the GENERALIZED VITERBI SEARCH. The boxes represent the list of states among which the best *S* path are selected. Some survivor path (alternative candidates) can share the same previous states whereas some unlikely states may be pruned in order to limit the overall complexity of the search.

In a conventional synthesizer, the search for the optimal sequence of speech units (eq. (1)) is decomposed in two separate optimisation problems (eq. (2) and (3)). These two equations are generally solved using the Viterbi algorithm. This algorithm defines a trellis 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 GENERALIZED VITERBI ALGORITHM [3] 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.

An illustration of this approach is given in figure 2, 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 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 optimal sequence 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 cumulated symbolic/acoustic likelihood.

4. Evaluation

Objective and subjective evaluations 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 a CONVENTIONAL (explicit modelling of speech prosody, no use of speech alternatives) speech synthesis systems (table 1).

	symbolic	acoustic	
	alternatives	prosody	alternatives
BASELINE	(✓)	-	-
CONVENTIONAL	(✓)	syllable/phrase	-
PROPOSED	(✓)	syllable/phrase	✓

TABLE 1 – Description of TTS systems used for the evaluation.

Additionally, symbolic alternatives have been optionally used for each compared method to assess the relevancy of symbolic and acoustic alternatives separately.

4.1. Speech Material

The speech material used for the evaluation is a 5 hours French story-telling database interpreted by a professional actor, that 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. [4] for more details).

4.2. Objective evaluation

An objective evaluation 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.

4.2.1. Procedure

The objective evaluation has been conducted with the 173 sentences of the fairy tale "*Le Petit Poucet*" ("*Tom Thumb*").

For this purpose, a *cumulated* log-likelihood has been defined as a weighted integration of the *partial* log-likelihoods (symbolic, acoustic). First, each partial log-likelihood have been averaged over the utterance to be synthesized so as to normalize the variable number of observations used for the computation (e.g., phonemes, syllable, prosodic phrase). Then, log-likehoods have been normalized to ensure comparable contribution of each partial log-likelihood during the speech synthesis. Finally, the cumulated log-likelihood of a synthesized speech utterance has been defined as follows :

$$LL = w_{\text{symbolic}} LL_{\text{symbolic}} + w_{\text{acoustic}} LL_{\text{acoustic}} \quad (4)$$

where LL_{symbolic} and LL_{acoustic} denote the partial log-likelihood associated with the sequence of symbolic and acoustic characteristics ; and w_{symbolic}, w_{acoustic} corresponding weights.

Finally, the optimal sequence of speech units is determined so as to maximize the cumulated log-likelihood of the symbolic/acoustic characteristics. In this study, 10 alternatives have been considered for the symbolic characteristics, and 50 alternatives for the selection of speech units, and weights have been heuristically chosen as $w_{\text{symb}} = 1, w_{\text{phone}} = 1, w_{\text{syllab}} = 5$, and $w_{\text{phrase}} = 1$.

4.2.2. Discussion

Cumulated likelihood obtained for the compared methods is presented in figure 3, without and with the use of symbolic alternatives. The PROPOSED method (modelling 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. Additionally, the use of symbolic alternatives conducts to a significant improvement regardless to the method considered. Finally, the optimal synthesis is obtained for the combination of symbolic/acoustic alternatives with the modelling of speech prosody.

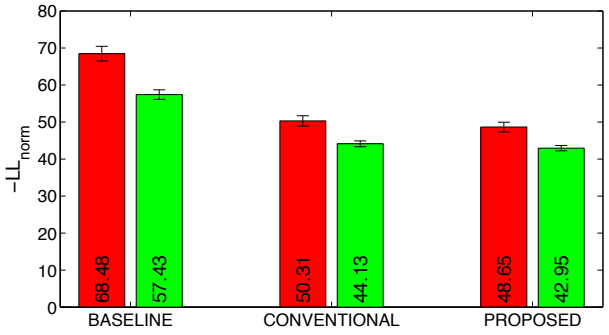
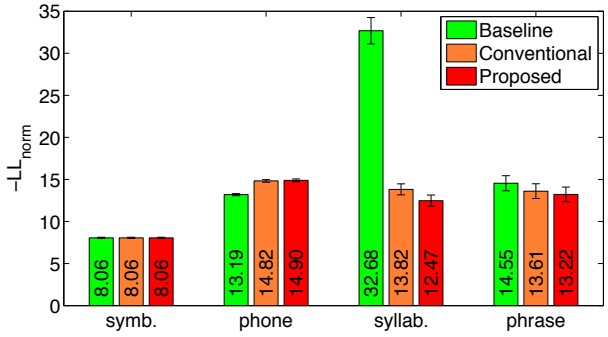


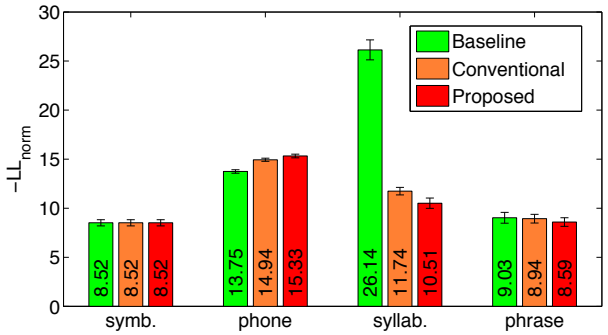
FIGURE 3 – Cumulated likelihood (mean and 95% confidence interval) obtained for the compared TTS, without (left) and with (right) use of symbolic alternatives.

For further investigation, partial likelihoods obtained for the compared methods are presented in figure 4, without and with 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 (non significant). This suggests that the phrase modelling (as described in [6]) 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.



(a) no use of symbolic alternatives



(b) use of symbolic alternatives

FIGURE 4 – Partial log-likelihoods (mean and 95% confidence intervals) for the compared methods, without and with use of symbolic alternatives.

4.3. Subjective evaluation

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

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. 15 native French speakers have participated in the evaluation. The evaluation has been conducted according to a *crowd-sourcing* technique using social networks. Pairs of synthesized speech utterances were randomly presented to the participants who have been asked to attribute a preference score according to the *naturalness* of the speech utterances on the comparison mean opinion score (CMOS) scale. Participants have been encouraged to use headphones.

4.3.2. Discussion

Figure 5 presents the CMOS obtained for the compared methods. The PROPOSED method is substantially preferred to the 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 [6]).

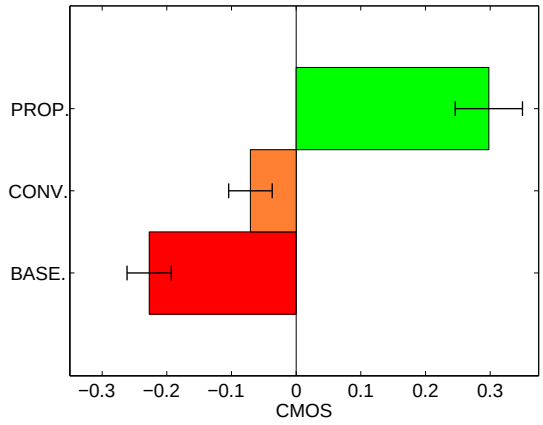


FIGURE 5 – CMOS (mean and 95% confidence interval) obtained for the compared methods.

5. Conclusion

In this paper, the use of speech alternatives in unit-selection speech synthesis have been introduced. Speech alternatives may be advantageously used either to improve the quality and the variety of the speech synthesis. Objective and subjective evaluations supports evidence that the use of speech alternatives qualitatively improves speech synthesis over conventional speech synthesis systems. 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 sequence of speech units.

6. References

[1] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

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

[3] T. Hashimoto, "A list-type reduced-constraint generalization of the Viterbi algorithm," *IEEE Transactions on Information Theory*, vol. 33, no. 6, pp. 866–876, 1987.

[4] N. Obin, A. Lacheret, and X. Rodet, "HMM-based Prosodic Structure Model Using Rich Linguistic Context," in *Interspeech*, Makuhari, Japan, 2010, pp. 1133–1136.

[5] N. Obin, P. Lanchantin, A. Lacheret, and X. Rodet, "Reformulating Prosodic Break Model into Segmental HMMs and Information Fusion," in *Interspeech*, Florence, Italy, 2011, pp. 1829–1832.

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