

Polyglot Speech Prosody Control

Harald Romsdorfer

Speech Processing Group, ETH Zurich, Switzerland

romsdorf@tik.ee.ethz.ch

Abstract

Within a polyglot text-to-speech synthesis system, the generation of an adequate prosody for mixed-lingual texts, sentences, or even words, requires a polyglot prosody model that is able to seamlessly switch between languages and that applies the same voice for all languages. This paper presents the first polyglot prosody model that fulfills these requirements and that is constructed from independent monolingual prosody models. A perceptual evaluation showed that the synthetic polyglot prosody of about 82% of German and French mixed-lingual test sentences cannot be distinguished from natural polyglot prosody.

Index Terms: mixed-lingual, polyglot, speech synthesis, prosody control, neural networks, ensemble models

1. Introduction

All existing approaches for modeling prosody of multiple languages for speech synthesis have been concentrated so far on making the prosody models “language-independent”, as it was formulated by van Santen in [1]. A *multilingual prosody model* is able to generate the prosodic contour for multiple languages, but in general not by the same voice. Switching between languages is only possible at sentence boundaries and is usually accompanied by voice switching. Seamless language switching and correct prosody modeling of foreign word or word group inclusions is therefore not possible.

The limitations of multilingual prosody models restrict the usability of TTS synthesis systems to monolingual texts. The generation of an adequate prosody for mixed-lingual texts, sentences, or even words, requires a *polyglot prosody model* that is able to seamlessly switch between languages and that applies the same voice for all languages. Listening experiments verified this finding. E.g., [2] demonstrated the need of English prosody for the English inclusions in German sentences.

The requirements of a *polyglot prosody model* for polyglot TTS synthesis can be summarized as follows:

- First, for a prosody model to be *polyglot*,
 - the generation of prosodic contours must be done with *prosody models of the same speaker for all languages*, and
 - *seamless switching between languages* must be possible such that no rhythmic or melodic discontinuity is audible.
- And second, the model must be *language-independent*. E.g., it must be possible to extend a polyglot prosody model to cover an additional language without modifications of the model parameters for already supported languages.

2. Model Architecture

The polyglot prosody model consists of independent F_0 control and segment duration control modules that generate from the

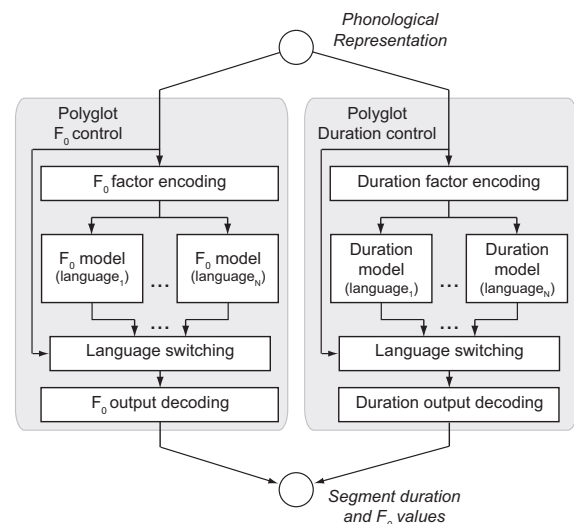


Figure 1: Schematic representation of the polyglot prosody model: independent F_0 control and segment duration control modules generate from the phonological representation of an utterance the corresponding F_0 and segment duration values.

phonological representation of an utterance the corresponding F_0 and segment duration values. Figure 1 displays a schematic overview of this model.

A *factor encoding* component converts the phonological representation into a language-independent input representation. An *output decoding* component converts the language-independent output representation into the actual acoustic parameter. These language-independent representations enable language switching between monolingual models and make it possible to add new models for others languages without requiring to modify the existing models. Language switching itself is triggered by the language tags of the phonological representation.

The polyglot F_0 control is described in Section 3 and the polyglot segment duration control in Section 4. Section 5 finally presents a perceptual evaluation experiment using the polyglot prosody model and a discussion of the results.

3. Fundamental Frequency Control

The polyglot F_0 control processes the phonological representation of a polyglot utterance as a sequence of syllable and boundary symbols. For each symbol, it generates a F_0 contour by applying the monolingual F_0 model that corresponds to the symbol’s language.

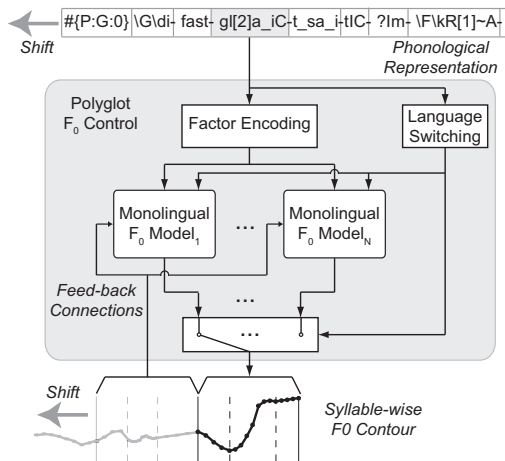


Figure 2: Schematic representation of the polyglot F_0 model: for each syllable and for each boundary symbol of the phonological representation, a set of language independent input factors is extracted and encoded. For language switching, the F_0 output values of the last syllable are fed back to the input of the monolingual models.

3.1. Model Architecture

The polyglot F_0 control consists of a language-independent input factor representation, that is described in Section 3.3, a language and time independent F_0 output representation, which is presented in Section 3.4, and an independent, monolingual F_0 model for each individual language. In order to provide language switching between the individual monolingual models, the F_0 outputs of the preceding syllable are fed back to the inputs of the monolingual models. Figure 2 gives a schematic overview of the polyglot F_0 model.

Each monolingual F_0 model is a weighted ensemble of recurrent neural networks (RNNs) that is constructed using the procedure presented in [3]. Each RNN has its own input factor selection that chooses the optimal set of input factors for this network. The basic RNN structure is similar to the RNN-based F_0 model presented in [4]. The network setup of the RNN ensemble members of the German and of the French F_0 models is given in Table 1.

| German F_0 ensemble | | | | | | | | | |
|-----------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Network Nr. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | |
| Factors | 170 | 380 | 350 | 352 | 160 | 270 | 290 | 370 | |
| Layer 1 | 28 | 15 | 14 | 17 | 24 | 22 | 21 | 15 | |
| Layer 2 | 27 | 22 | | | 22 | | | 27 | |
| French F_0 ensemble | | | | | | | | | |
| Network Nr. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Factors | 455 | 342 | 616 | 560 | 447 | 547 | 453 | 547 | 447 |
| Layer 1 | 11 | 18 | 10 | 11 | 14 | 12 | 14 | 12 | 14 |
| Layer 2 | | | | | | | | | |

Table 1: Network structure of each RNN member of the best ensemble for German and for French F_0 control. For each ensemble member, the number of input factors and the number of nodes of the first and of an optional second hidden layer is given.

3.2. Language Switching

The feed-back connections from the last hidden layer in the RNN-based F_0 model of Traber mainly serve to control the general level of F_0 whereas the more local phenomena are controlled by the direct input to the network, cf. [4]. In order to enable language switching without audible melodic discontinuities, the feed-back connections of the RNN model of the preceding language can be used to initialize the recurrent input of the RNN model of the new language. Thus, the model for the new language continues at the same general level of F_0 as defined by the model of the preceding language.

The ensemble models for F_0 modeling therefore apply feed-back connections from the F_0 outputs. Each RNN ensemble member uses its own F_0 outputs as feed-back. These feed-back connections can be set to some external F_0 values: either to zero, in order to initialize the RNN at the start of an utterance, or to the F_0 values of the preceding syllable, in order to set the general F_0 level for language switching.

3.3. Input Representation

The phonological representation of an utterance is processed as a sequence of syllable and boundary symbols. Each input symbol is represented by a vector of input factors. All elements of this vector are set to zero by default. The values of ordinal factors are directly set in the vector. For categorical factors, a 1-out-of-n encoding is applied such, that each categorical factor is represented by n binary factors.

It is generally acknowledged, that the F_0 contour of a syllable depends on a relatively wide phonological context as far as accentuation and phrasing information is concerned, whereas the influence of segmental properties on the F_0 contour of a syllable is much more local. However, the correct size of these contexts for the different factors is unknown and depends on the prosodic phenomena to be modeled. The author therefore applied a context of 3 preceding and 6 subsequent symbols for accentuation and phrasing information (equal to the context used in [4]), and a context of 2 preceding and 2 subsequent symbols for segmental properties. In total, 910 input factors are derived for F_0 control. They consist of 190 accentuation and phrasing factors, 715 syllable structure and segmental factors, and 5 sentence length and syllable position factors.

For polyglot F_0 control, this input representation must be language independent. This means that no language specific segment types or phrase types can be used, but the language-independent description of manner and place of articulation of phones of the IPA and a basic, language-independent set of phrase types. Also, information about syllable language or language switching position may not be part of the factor set. Language information is only used to switch between the monolingual F_0 models.

3.4. Output Representation

In order to make F_0 control independent from duration control, a time-independent representation of the F_0 contour is necessary. This can be achieved by applying a linear approximation of the original, linearized F_0 contour using a constant number of equidistant F_0 samples for each syllable.

Empirical findings concerning the timing of F_0 peaks within syllables due to segmental constraints [5, 6] or semantic constraints [7] show that certain anchor points for positioning F_0 peaks within a syllable are necessary. A manual inspection of the F_0 contours of the syllables of the prosody corpora

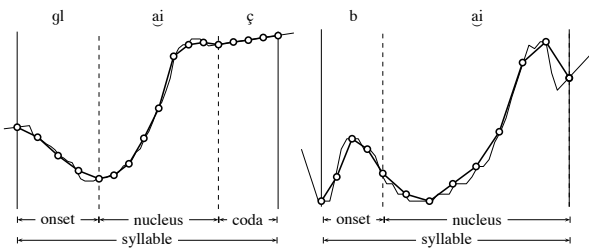


Figure 3: Modeling of the original, linearized F_0 contour (thin line) of each of the syllables [glajç] (left contour) and [baj] (right contour) by 17 F_0 values (indicated as circles).

revealed for identical vowels roughly similar patterns in the nucleus part of the F_0 contours. Figure 3 shows the F_0 contours of the syllables [glajç] and [baj] of the German prosody corpus. While the overall F_0 contours of these two syllables look rather different, the nucleus parts of both syllable have more similar F_0 patterns. To incorporate these findings into the F_0 model, the author introduced a “sub-syllabic” representation of the F_0 contour.

This *sub-syllabic representation* bases on a segmentation of each syllable into onset, nucleus, and coda. Onset and coda parts of the F_0 contour are each linearly approximated using 5 equidistant F_0 samples, the nucleus part of the F_0 contour is modeled by 9 equidistant F_0 samples. F_0 samples at onset-nucleus and nucleus-coda boundaries are identical. Thus, this representation uses 17 F_0 samples in total. In case of an absent onset or coda, the respective 5 F_0 samples have the same value and lie upon each other. Figure 3 displays the application of this F_0 contour modeling on two accented syllables of the German prosody corpus.

This sub-syllabic representation also conforms very well to the requirements concerning the timing of F_0 peaks within syllables due to segmental constraints [5, 6] and semantic constraints [7].

4. Segment Duration Control

The polyglot segment duration control generates for each phone and for each pause of the phonological representation of a polyglot utterance the corresponding duration value. For each phone, it applies the appropriate monolingual duration model that corresponds to the language of the phone.

4.1. Model Architecture

Figure 4 shows a schematic overview of the polyglot segment duration control: it consists of a *factor encoding* module, that generates for each phone or pause of the phonological representation of a polyglot utterance a language-independent input factor representation, that is described in Section 4.3. A *language switching* component selects the appropriate model from a set of independent, monolingual segment duration models and sets the appropriate speech rate. The selected monolingual duration model finally generates the segment duration values.

Each monolingual duration model is a weighted ANN ensemble that is constructed using the procedure presented in [3]. Each ANN has its own input factor selection that chooses the optimal set of input factors for this network. The network setup of the ANN ensemble members of the German and of the French duration models is given in Table 2.

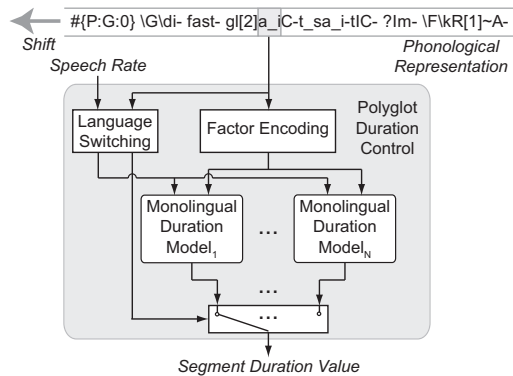


Figure 4: Schematic representation of the polyglot duration model: for each phone or pause of the phonological representation, the corresponding input factors are extracted and encoded.

4.2. Language Switching

Language switching within polyglot utterances must not result in audible rhythmic discontinuities. This requires that the general speech rates of both language specific duration models are similar. This could be achieved by recording prosody corpora of each language with similar, relatively constant speech rates having small variances. The speech rate of the German male prosody corpus displayed in Figure 5, e.g., exhibits such a “constant” speech rate with small variance (at least for longer utterances). However, as visible in Figure 5, the variances of speech rate of the German and of the French female prosody corpora are considerable and much larger than of the male corpus. In first experiments, switching between a German and a French duration model trained on these two corpora resulted therefore most of the time in an audible change of speech rate.

In order to cope with the large variances in speech rate, the speech rate and the number of syllables of a sentence were provided as additional input factors to the ANNs. The additional speech rate input made it possible to smoothly switch between the individual, monolingual duration models, simply by setting the same speech rate value as input for both duration models.

4.3. Input Representation

From the phonological representation of an utterance, a sequence of phone and pause segments is extracted. The hold (preplosive pause) and the burst part of plosives are hereby treated as two separate segments. For plosives after a speech pause, no preplosive pause is extracted. Diphthongs, triph-

| German duration ensemble | | | | | | | | | | |
|--------------------------|----|----|----|----|----|----|----|----|-----|-----|
| Network Nr. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Factors | 55 | 60 | 60 | 61 | 55 | 61 | 67 | 66 | 108 | 114 |
| Layer 1 | 30 | 25 | 30 | 25 | 30 | 23 | 50 | 50 | 50 | 50 |
| Layer 2 | | | | | | 5 | | | | |
| French duration ensemble | | | | | | | | | | |
| Network Nr. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | | |
| Factors | 80 | 70 | 60 | 60 | 58 | 56 | 68 | 60 | | |
| Layer 1 | 20 | 15 | 16 | 20 | 16 | 20 | 14 | 17 | | |
| Layer 2 | | 15 | 20 | 10 | 20 | 10 | 20 | 15 | | |

Table 2: Network structure of each ANN member of the best ensemble for German and for French duration control. For each ensemble member, the number of input factors and the number of nodes of the first and of an optional second hidden layer is given.

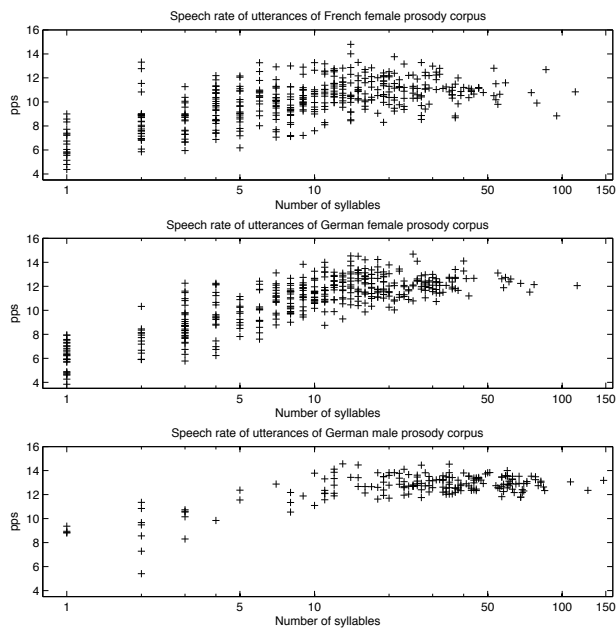


Figure 5: Speech rate in terms of phones per second (pps) of the utterances of the French (top) and the German (center) female prosody corpora, and of the German male (bottom) prosody corpus as a function of the number of syllables of an utterance.

thongs, and affricates are each treated as one segment. Each of these segments is represented by a vector of input factors similar to F_0 control.

Segment duration depends on a relatively local segmental context as far as segment type information is concerned, as shown, e.g., in [8]. The influence of accentuation and phrasing information, however, is wider. Similar to F_0 modeling, the correct size of the contexts for the different factors is unknown and depends on the prosodic phenomena to be modeled. Therefore a context of 2 preceding and 2 subsequent segments is applied for segmental information. For accentuation and phrasing information, a context of 2 preceding and 2 subsequent syllables is used. In total, 349 input factors are derived for duration control. They consist of 200 segmental factors, 95 accentuation, phrasing, and syllable length factors, 5 syllable level factors, 26 foot level factors, 7 phrase level factors, and 16 sentence level factors.

For polyglot duration control, this input representation must be language-independent. Thus, no language specific segment types or phrase types are used, but the language-independent description of manner and place of articulation of phones of the IPA and a basic, language-independent set of phrase types. Also, information about syllable language or about language switching position may not be included in the factor set. Language information is only used to switch between the individual monolingual duration models.

5. Evaluation

In order to evaluate the quality of the complete polyglot prosody control, a listening experiment was conducted with 7 subjects, cf. [3] for details. The subjects were presented a total of 160 sentences. These sentences consisted of 40 German and 40 French sentences, each one with its natural prosody and with the synthetic prosody predicted by the polyglot prosody control.

The sentences were presented in random order. 20 of the German sentences and 8 of the French sentences were taken from the polyglot test set. These mixed-lingual sentences contained either English, French or German foreign inclusions. The other sentences were taken from the German and from the French monolingual test sets.

The perceptual evaluation showed that about 87.5% of all synthetic German prosody contours and about 92.5% of all synthetic French prosody contours cannot be distinguished from natural prosody contours. Considering only the mixed-lingual sentences, this is the case for about 80% of the mixed-lingual German sentences and for about 87.5% of the mixed-lingual French sentences.

6. Conclusions

This new approach to prosody control achieves impressive improvements even when compared to the monolingual SVOX system that was regarded as one of the best TTS systems for German, cf. [4]. For duration control, an improvement of the prediction error of about 12% compared to the best MARS-based duration model of [8] was achieved, and for F_0 control, a prediction error improvement of about 24% compared to the best RNN-based F_0 model of [4] was reached. These improvements made it possible, that in a perceptual evaluation about 90% of 80 different monolingual and mixed-lingual test sentences having synthetic prosody were judged indistinguishable from the corresponding original recordings with human prosody. These results also show that it is possible to switch between monolingual prosody models at language boundaries without audible rhythmic or melodic discontinuities

7. Acknowledgements

This work was supported by ETH Zurich and by AAP-COMET.

8. References

- [1] J. P. H. van Santen, C. S. Shih, B. Möbius, E. Tzoukermann, and M. Tanenblatt, "Multi-lingual duration modeling," in *Proceedings of Eurospeech'97*, Rhodes, Greece, September 1997, pp. 2651–2654.
- [2] P. Olaszi, T. Burrows, and K. Knill, "Investigating prosodic modifications for polyglot text-to-speech synthesis," in *ISCA Tutorial and Research Workshop on Multilingual Speech and Language Processing (MultiLing 2006)*, Stellenbosch, South Africa, April 2006.
- [3] H. Romsdorfer, "Polyglot text-to-speech synthesis. Text analysis and prosody control," Ph.D. dissertation, No. 18210, Computer Engineering and Networks Laboratory, ETH Zurich (TIK-Schriftenreihe Nr. 101), January 2009.
- [4] C. Traber, "SVOX: The implementation of a text-to-speech system for German," Ph.D. dissertation, No. 11064, Computer Engineering and Networks Laboratory, ETH Zurich (TIK-Schriftenreihe Nr. 7, ISBN 3 7281 2239 4), March 1995.
- [5] J. P. H. van Santen and J. Hirschberg, "Segmental effects on timing and height of pitch contours," in *Proceedings of ICSLP'94*, Yokohama, Japan, September 1994, pp. 719–722.
- [6] J. P. H. van Santen, "Quantitative modeling of pitch accent alignment," in *Proceedings of Speech Prosody 2002*, Aix-en-Provence, France, April 2002, pp. 107–112.
- [7] K. J. Kohler, "Neglected categories in the modelling of prosody – pitch timing and non-pitch accents," in *Proceedings of 15th ICPHS*, Barcelona, Spain, 2003, pp. 2925–2928.
- [8] M. Riedi, "Controlling segmental duration in speech synthesis systems," Ph.D. dissertation, No. 12487, Computer Engineering and Networks Laboratory, ETH Zurich (TIK-Schriftenreihe Nr. 26, ISBN 3-906469-05-0), February 1998.