

A Mixed-lingual Phonological Component which Drives the Statistical Prosody Control of a Polyglot TTS Synthesis System

Harald Romsdorfer, Beat Pfister, and René Beutler

Speech Processing Group
Computer Engineering and Networks Laboratory
ETH Zurich, Switzerland
{romsdorfer,pfister,beutler}@tik.ee.ethz.ch

Abstract. A polyglot text-to-speech synthesis system which is able to read aloud mixed-lingual text has first of all to derive the correct pronunciation. This is achieved with an accurate morpho-syntactic analyzer that works simultaneously as language detector, followed by a phonological component which performs various phonological transformations. The result of these symbol processing steps is a complete phonological description of the speech to be synthesized. The subsequent processing step, i.e. prosody control, has to generate numerical values for the physical prosodic parameters from this description, a task that is very different from the former ones. This article shows appropriate solutions to both types of tasks, namely a particular rule-based approach for the phonological component and a statistical or machine learning approach to prosody control.

1 Introduction

Following the approach of generative phonology in [CH68], text-to-speech (TTS) synthesis requires the underlying syntactic structure in order to derive the correct pronunciation, as has been shown, e.g., in [Tra95] and [Spr96]. This underlying structure is even more important in the case of polyglot TTS where mixed-lingual input text has to be processed. Such texts can contain various types of inclusions from other languages.¹

An appropriate morphological and syntactic analyzer for mixed-lingual text has been presented in [PR03]. Such an analyzer is most suitably realized with lexica and sets of grammar rules (i.e. the linguistic knowledge of the languages concerned), which are applied by means of a parser. In other words, it is obvious to use symbol processing methods here.

There are other subtasks of TTS synthesis, however, that demand for other types of solutions. Particularly, in prosody control, where linguistic knowledge is very scarce, we consider rule-based approaches inappropriate. We therefore prefer statistical models and machine learning approaches to solve the prosody control problem.

¹ Three major groups of foreign inclusions can be identified: mixed-lingual word forms with foreign stems, full foreign word forms that follow the foreign morphology, multi-word inclusions that are syntactically correct foreign constituents.

In the sequel we present two components of our TTS synthesis system, one using symbol processing methods and one based on a statistical model. The structure of the article is as follows: In Sect. 2 we sketch the architecture of our polyglot TTS system with its two main parts. Subsequently, we describe in Sect. 3 the phonological component which is based on a special type of rules. The output of this component is a minimal but complete information on each sentence to be synthesized. This information is called the phonological representation. Section 4 explains our approach to statistical prosody control and shows in more details how the fundamental frequency of a sentence is generated from its phonological representation.

2 Overview of the Polyglot TTS System polySVOX

The task of our TTS system, called polySVOX, is to produce high-quality synthetic speech from mixed-lingual text. Thereby a polyglot pronunciation is aimed at, which means that foreign inclusions are pronounced in accordance to the rules of the originating language. Assimilation to the base language is rather marginal and primarily happens close to language switching positions.

In contrast to so-called multilingual TTS systems that can be switched to operate in one of several languages modes, but that treat in general each language with an independent subsystem and synthesize it with a language-specific voice, our polyglot TTS system has to apply the knowledge of all languages concerned in parallel and obviously needs one and the same voice for all languages. These are vital prerequisites to seamlessly switch between languages within sentences and even within words.

Additionally, it is most desirable that the TTS system can easily be configured for an arbitrary set of languages. In order to achieve this, the language knowledge and the processing have strictly been separated.

Basically, polySVOX consists of two main parts, called transcription stage and phono-acoustical model (cf. Fig. 1).

The *transcription stage* is strictly voice-independent and implements the linguistic competence of the TTS process. It comprises a morphological and syntactic analyzer, realized as a bottom-up chart parser, plus a subsequent rule-based phonological component determining syllable stress levels, prosodic phrase boundaries and phone

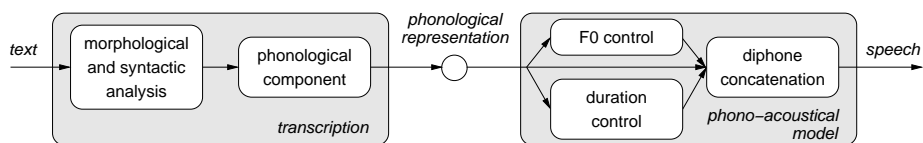


Fig. 1. Overview of the polyglot TTS system polySVOX: The voice-independent transcription stage transforms a sentence of the input text into the phonological representation of the corresponding utterance. The voice-dependent phono-acoustical model generates the sentence prosody from this phonological representation and produces the speech signal.

transformations. Due to the strict separation of linguistic data from the corresponding algorithms the transcription stage can easily be reconfigured for a new language. Provided that some basic rules are obeyed, a mixed-lingual transcription part can be constructed by loading a set of monolingual data sets, as shown in [PR03]. The output of the transcription stage is called phonological representation.

The *phono-acoustical model* is the voice-dependent part of the TTS process that implements the linguistic performance. It generates the prosody from the phonological representation of an utterance by means of statistical models. For duration control a multivariate adaptive regression splines (MARS) model is used (see [Rie98]). Fundamental frequency (F_0) is generated by means of a recurrent neural network (RNN) which is presented in Sect. 4.2. Finally, the synthetic speech signal is generated by concatenation of diphones, as shown in [TP99].

From the sketch of these two main parts it is obvious that the two tasks are of quite a different nature. In our TTS system this is reflected by the different types of information processing methods: In the transcription stage, that transforms symbolic into symbolic information, knowledge- and rule-based methods are used. Here the main challenge to the polyglot system is the exploration of algorithms that allow a combined usage of linguistic competence of an arbitrary set of languages. The phono-acoustical model, however, maps symbolic information to physical parameters, mainly by means of statistical models (or machine learning approaches) and signal processing methods.

The fundamentally different solutions selected for the two parts are further detailed in the two following sections: Whereas in Sect. 3 the phonological component² is shown as an example for symbolic processing, Sect. 4 illustrates the prosody control of the polySVOX system, particularly the fundamental frequency modeling.

3 Phonological Component

In the polySVOX architecture, syllabification, stress assignment, phrasing and various phonological transformations are done in the so-called phonological component. This component processes the syntax tree from the morphological and syntactic analyzer and generates the phonological representation (cf. Fig. 2).

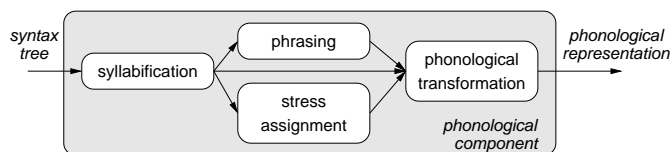


Fig. 2. Overview of the phonological component.

² The other component of the transcription stage, the morphological and syntactic analyzer for mixed-lingual text has been published in [PR03].

In the following, we demonstrate the functionality of the phonological component by means of the mixed-lingual example sentence: “Anciens Amis sind keine Amis anciens.” (“Ex-friends are no old friends.”). This German declarative sentence contains two incomplete French noun groups, i.e., the article is missing which corresponds to the German indefinite plural form. The syntax tree of this sentence is shown in Fig. 3. The phonological component processes it as follows:

- The phone sequence is split into phonological syllables. Such syllables may span across word boundaries, in contrast to orthographic ones. Note that the syllable structure may be changed by subsequent phonological transformations.
- Stress assignment and placement of prosodic phrase boundaries are described in [Tra95]. The conceptual ideas originate from [Kip66] and [Bie66], where algorithms for the generation of sentence stress patterns and phrase boundaries from the syntactic structures of German sentences are presented. Stress assignment and phrasing for English and French sentences are based on the same algorithms, requiring different, language specific patterns and rules, however.
- Phonological transformations for phenomena like liaison, elision, linking, aspiration, assimilation, etc. are expressed with so-called multi-context rules. More details about this rule formalism are given in Sect. 3.2.

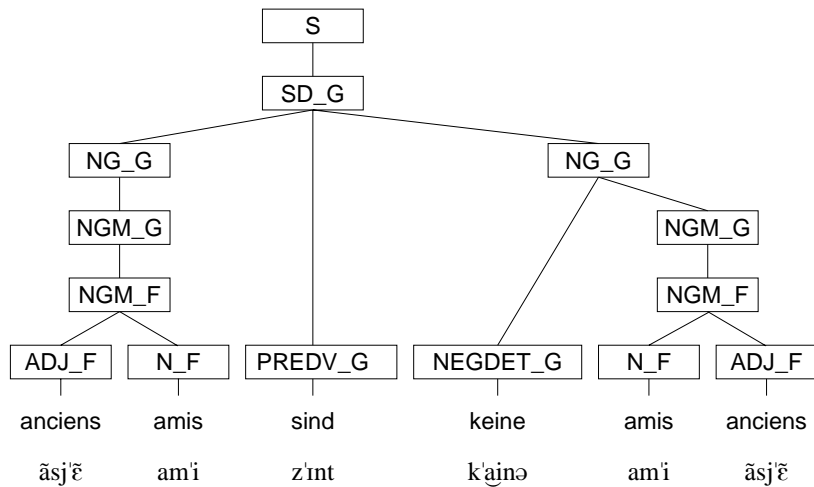


Fig. 3. Syntax tree of the sentence “Anciens Amis sind keine Amis anciens.”, including graphemic and phonological terminals. The phonetic symbols follow the IPA definition. The suffixes *_F* and *_G* of the constituent identifiers indicate the languages French and German, resp.

3.1 Requirements of a Mixed-lingual Phonological Component

The requirements of a mixed-lingual phonological component are evidently determined by the kind of pronunciation wanted. For our TTS system we are aiming at a polyglot pronunciation of mixed-lingual text (cf. Sect. 2). Consequently, the phonological component of our polyglot TTS system must be able to cope with the different phonological phenomena of each language involved. These phenomena are language-specific and depend on various contexts, as illustrated by the following examples:

German aspiration: In word-initial position, the German unvoiced plosives [p], [t] and [k] preceding a vowel are aspirated, denoted as [p^h], [t^h] and [k^h], resp. They are also aspirated in word-final position before a break.

German terminal devoicing: All voiced plosives (obstruents) before a syllable or word boundary are devoiced.

French liaison: In French noun groups, liaison is forbidden between a singular noun and the consecutive adjective, e.g. “un bruit effroyable” [œ-brɥi-e-frwa-jabl]; between a plural noun and the following adjective it is optional, e.g. “les amis agréables” [le-za-mi-(z)a-ɡre-abl]; liaison is mandatory between the preceding adjective and a noun, e.g. “un bon ami” [œ-bɔ-na-mi].

Liaison is generally avoided between a singular noun and the following verb, e.g. “l’étudiant entend” [le-ty-djã-ã-tã]; it is optional between a plural noun and the following verb, e.g. “les étudiants entendons” [le-ze-ty-djã-(z)ã-tã-dɔ̃]; but liaison is mandatory between a clitic personal pronoun and the following verb, e.g. “on entend” [ɔ̃-nã-tã].

French liaison consonant realization: The phonetic liaison consonant can be directly derived from the corresponding graphemic consonant: “s”, “x” or “z” result in [z]; “c”, “q” or “g” in [k], etc.

English linking “r”: Word-final “r” is usually only pronounced, if the following word begins with a vowel, e.g. “four eggs” [fɔ:r-egz] but “four pounds” [fɔ:-paundz].

These examples show that phonological phenomena depend on various contexts: The German aspiration rule needs only phonetic context, whereas the English linking rule requires both phonetic and graphemic contexts. The French liaison rules need phonetic, graphemic and syntactic contexts. Furthermore, since all of these phonological phenomena are language-specific, language forms another context.

Additionally, cross-lingual assimilation phenomena must be considered. Even if foreign inclusions in German sentences virtually keep the pronunciation prescribed by the originating language and assimilation to the base language is very weak, it is clearly present and must be handled correctly.

In a word like e.g. “Dufourstrasse”, which is composed from the French proper name “Dufour” and the German noun “Strasse” (“street”), the French [ʁ] has to be replaced by the German [r]. It would sound rather affectedly to pronounce [dy-fur-ʃtra:-sə] instead of [dy-fur-ʃtra:-sə]. Such assimilations occur only near the language switching position, however, and only in short inclusions.

3.2 Phonological Transformation

The application of phonological transformations to a sentence produces the standard pronunciation of this sentence. For the example sentence of Fig. 3 this standard pronunciation is (to simplify matters, syllable stress and prosodic phrase information has been removed):

[ʔã̃s-jẽ-z a-mi- zɪnt- k^haj-nə- ʔa-mi-(z) ã̃s-jẽ].

Note that phonetic symbols in parentheses are optional. The following phonological transformations have been applied in this example:

The aspiration of [k^h] in the German word “keine” follows the German aspiration rule defined in Sect. 3.1. The plosive [t] in the word “sind” is not aspirated, however, because there is no break after this word.

The standard pronunciation of the French partial noun groups is obtained by applying the French rules for mandatory and optional liaison, resp. The first French inclusion “Anciens Amis” is pronounced [ã̃s-jẽ-za-mi]. Here the liaison consonant [z] was inserted. In the second incomplete French noun group “Amis anciens”, the liaison consonant [z] is optional (as defined in Sect. 3.1) which results in [a-mi-(z)ã̃s-jẽ]. The actual realization of this optional consonant depends on the style of pronunciation wanted.

Furthermore, a cross-lingual phenomenon has to be considered which arises from the German glottal stop rule. According to this rule, potentially every initial vowel of a word or a stem morpheme is preceded by a glottal stop. A similar rule applies also for foreign inclusions, i.e., a glottal stop has to be assigned to the initial vowel of the inclusion. “Anciens Amis” therefore has to be pronounced as [ʔã̃s-jẽ-za-mi] in our example sentence.

From this example it is obvious that only the application of both the German and the French phonological transformations is able to produce the desired standard pronunciation for a German sentence with French inclusions.

Multi-context Rules: A rule formalism that is flexible enough to describe all possible context restrictions of such phonological transformations was introduced in [RP04]. This so-called multi-context rule formalism allows to define phonological transformations which are restricted by specific syntactic, graphemic and/or phonological contexts. Formally a multi-context rule consists of a subtree pattern, the separation symbol ‘:’ and an associated phonological transformation:

SubtreePattern : *Transformation* ;

Transformations are specified in the form: $\sigma/\rho \Leftrightarrow L_R$. These context-dependent rewrite rules are similar to the well-known two-level rules (see e.g. [Kos83]), but have been extended here to operate on all types of symbols, i.e., graphemic and phonological ones at the same time.

The subtree pattern specifies the syntactic context and defines for each constituent whether the graphemic and/or phonological terminals are subject to the phonological transformation defined by the rule. These patterns may be specified using constituent

Table 1. Wild-card and special symbols used within multi-context rules

*	any sequence (0..n) of constituents including their (possibly empty) subtrees
?	any constituent (exactly one) including its (possibly empty) subtree
(...)	syntax hierarchy marker
<...>	feature specification
[]	phonological representation operator
{}	graphemic representation operator
%id	set identifier

symbols plus additional wild-card symbols as listed in Table 1. The application of the associated transformation gets triggered whenever the subtree pattern can be matched with a part of the syntax tree. Examples of such subtree patterns are shown in Fig. 4.

Implementation: In the polySVOX system, the subtree patterns of multi-context rules are represented as strings. This representation includes the special symbols listed in Table 1. In the two multi-context rules for French liaison of Fig. 5, the two left syntax patterns of Fig. 4 are used. Figure 6 shows a mixed-lingual rule with the syntax pattern of Fig. 4.

Applying these multi-context rules to the syntax tree of the example sentence (cf. Fig. 3) results in following transformations: The first multi-context rule for French liaison of Fig. 5 inserts an optional [(z)]. The phonological input sequence as selected

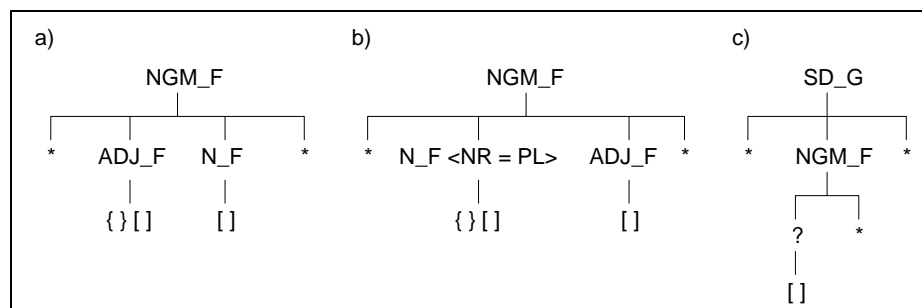


Fig. 4. Examples of subtree patterns: Pattern a) specifies the syntactic context of French mandatory liaison between a noun and a preceding adjective within a French noun group. The operator '[]' selects the phonological terminals of both constituents for application by the associated phonological rule. Analogously, the operator '{}' selects the graphemic terminals of the first constituent. Pattern b) specifies the syntactic context of French optional liaison between a plural noun and a subsequent adjective. The pattern specifies the noun with an additional feature-value pair <NR=PL>, i.e., to select only plural nouns. Pattern c) specifies a mixed-lingual syntactic context for glottal stop assignment within French noun groups as foreign inclusions in German declarative sentences.

```

%P set of all phone symbols
%V set of vowel symbols

NGM_F ( * N_F <NR=PL> { } [ ] ADJ_F [ ] * ) :
    @/'(z)' <=> 's' '}' ' [' { %P } %V _ ']' ' [' %V ;

NGM_F ( * ADJ_F { } [ ] N_F [ ] * ) :
    @/'z' <=> 's' '}' ' [' { %P } %V _ ']' ' [' %V ;

```

Fig. 5. Multi-context rules for French liaisons: The first rule inserts an optional liaison [(z)] between French plural noun and subsequent French adjective within the nominal part of a French noun group. The second rule inserts a liaison [z] between preceding French adjective and French noun within the nominal part of a French noun group. Both rules have the same graphemic context (the grapheme “s” preceding a graphemic word boundary ’’) and the same phonetic context (two neighboring vowels in front of and after a phonetic word boundary ’]’ ’[’).

by the associated syntax pattern is shown on the left side and the corresponding output sequence is shown on the right:

$$\{\text{amis}\}[\text{ami}][\text{ãsjẽ}] \Rightarrow \{\text{amis}\}[\text{ami(z)}][\text{ãsjẽ}]$$

The second multi-context rule for French liaison inserts a mandatory liaison [z]:

$$\{\text{anciens}\}[\text{ãsjẽ}][\text{ami}] \Rightarrow \{\text{anciens}\}[\text{ãsjẽz}][\text{ami}]$$

The mixed-lingual rule of Fig. 6 which defines glottal stop insertion matches twice:

$$\begin{aligned} [\text{ãsjẽz}] &\Rightarrow [ʔ\text{ãsjẽz}] \\ [\text{ami(z)}] &\Rightarrow [ʔ\text{ami(z)}] \end{aligned}$$

```

%V set of vowel symbols

SD_G ( * NGM_F ( ? [ ] * ) * ) :
    @/'?' <=> ' [' _ %V ;

```

Fig. 6. Multi-context rule for glottal stop assignment to French inclusions in German sentences: This rule inserts a glottal stop at the beginning of the first subconstituent of French noun groups only if they are inclusions in a German declarative sentence.

3.3 Phonological Representation

The output of the transcription stage (cf. Fig. 1) is called phonological representation. This representation consists of the phonetic transcription of the words of an utterance

Table 2. Overview of special symbols used in the phonological representation

#{x}	phrase boundary where increasing x denotes decreasing boundary strength: #{0} sentence break before and after the sentence. ≥#{1} sentence-internal phrase boundaries.	
(X)	phrase type mark set at the beginning of a phrase: P progredient phrase Y question with rising pitch at the end T terminal phrase W question with falling pitch at the end S semi-terminal phrase LI, LM, LF enumeration: initial, middle and final	
\X\	switch to language X, where X is e.g. E, F, G, etc. for English, French, German, resp.	
[x]	syllable stress level x, where x is E emphatic stress 1 main phrase stress 2, 3, 4 decreasing stress levels 0 completely unstressed syllable (may be omitted)	
-	syllable boundary	

together with prosodic information such as strength and type of phrase boundaries and stress levels of syllables. In case of mixed-lingual input text, appropriately placed language tags mark the language switching positions. For the example sentence of Fig. 3, "Anciens Amis sind keine Amis anciens.", the following phonological representation is produced by the transcription stage of the polySVOX system:

#{0} (P) \F\[2]?äs-jē-[1]z a-mi- #{2} (T) \G\zint- [2]k^haj-nə- \F\[3]?a-mi-[1](z) äs-jē #{0}

Besides phonetic symbols (the system-internal phonetic symbols are replaced by the corresponding IPA symbols here), phonological representations may include various special symbols which are explained in Table 2.

4 Prosody Control

In TTS synthesis, prosody control means to generate information on the following physical parameters of the speech signal to be synthesized: fundamental frequency (F_0), signal intensity and durations of phones and pauses. There are quite a number of factors that affect these parameters. The most important ones are syllable stress level, phrase type, position and strength of phrase boundaries, speaking style, language, etc. How all these factors in all possible combinations influence the physical parameters is virtually unknown, however. It is therefore appealing to solve the prosody control problem by means of a machine learning approach, i.e., to learn the interrelation between the influencing factors and the physical parameters of prosody (so-called prosodic parameters) from examples of natural speech.

Assuming that mutual dependency between the prosodic parameters is negligible, it is possible to model each prosodic parameter individually. Accordingly, prosody control of the polySVOX system consists of independent statistical models for F_0 and for

duration control. Signal intensity is not actively controlled, since the intrinsic intensities of the diphones proved to be sufficient for the desired speaking style.

In the following we present one of the statistical prosody models, namely the F_0 model. But first it is shown how speech examples have to be prepared for the training of these independent models.

4.1 Preparation of Speech Examples for Prosody Modeling

It goes without saying, that if a natural speech example has to be used for the training of F_0 and phone duration models, F_0 and phone boundaries have to be estimated. In order to achieve independent models, however, F_0 must not be described in function of time, but e.g. per syllable.

This is achieved as follows: The speech signal of a sentence is first segmented into phones by means of an HMM-based phone recognizer. Since the speech signal generally comes from a trained speaker who has read the corresponding text, this text can be used to generate the phonological representation (using the transcription stage of the TTS system) which includes also the sequence of phone symbols that describes the standard pronunciation. It is therefore possible to estimate the phone boundaries by forced alignment, which is more accurate than recognition, but still not accurate enough. Even trained speakers do not strictly speak in accordance to the standard pronunciation. In Fig. 7 it can be seen, that e.g. the words “Büros sollten” with the standard pronunciation [by-ro:s- zəl-tən] was spoken as [by-ro:s- əl-tən]. Hence we use speaker-specific pronunciation variation rules to produce alternative pronunciations that are put together

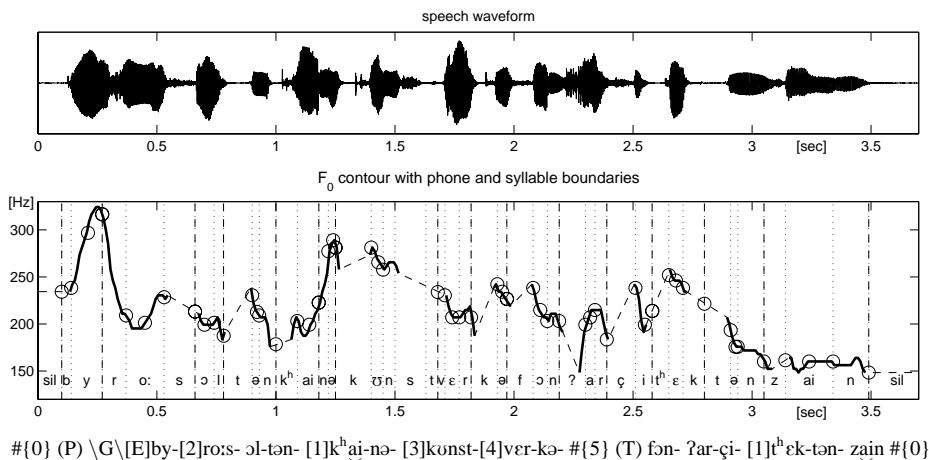


Fig. 7. The outputs of the prosodic analysis of the sentence “Büros sollten keine Kunstwerke von Architekten sein!” (“Office buildings shouldn’t be works of art of architects!”) are the phone boundaries (dotted lines), the syllable boundaries (dash-dotted lines) and the F_0 contour that has been interpolated across unvoiced sections (dashed parts). The desired time-independent representation of F_0 is attained by resampling the curve at five selected points per syllable, indicated by circles. For syllables with empty onset or coda the two first or last values are coincident.

in a recognition network and the Viterbi algorithm is used to determine the best matching path through the network (see [Rom04] for more details). Thus we achieve both an adapted phonetic transcription and an accurate phonetic segmentation of the signal. Additionally, we can decide by means of the phonological description which phone boundaries are also syllable boundaries.

The F_0 detection initially produces F_0 values in function of time. As can be seen in Fig. 7, the originally piecewise F_0 curve has been linearly interpolated across unvoiced sections to get a continuous curve.³ A syllable-wise and therefore time-independent representation of the F_0 contour can now be achieved by taking five F_0 values for each syllable, namely two at the syllable boundaries, two at the boundaries of the syllable nucleus, and the last one at the center of the nucleus. For syllables with empty onset or coda the two first or last values are coincident.

For each language concerned a collection of some 800 sentences, covering all different phrase types listed in Table 2, have been prepared as described above. All sentences were spoken by the same trained speaker. The resulting databases have been used for the construction of our F_0 and phone duration models.

4.2 Fundamental Frequency Control

Basically, the task of our F_0 control is to generate for each syllable five F_0 values, using the symbolic information of the phonological representation as input. Concatenating the F_0 values of all syllables yields the complete F_0 contour of the utterance to be synthesized.

The sentences of the prosody databases showed that high stress levels and strong phrase boundaries generally affect F_0 over several neighboring syllables, whereas other phonological properties affect F_0 only locally. Therefore, two syllable-aligned streams of symbols are extracted from the phonological representation and fed into the network:

- In order to consider syllable stress and phrase boundaries sufficiently, a context of three preceding and six following syllables is used to generate the F_0 values of the current syllable. For each of these syllables a three bit input is generated: two bits describing syllable stress level and one denoting the presence of a phrase boundary directly before that syllable. For context syllables beyond the start or the end of the input stream a special filler code is used (i.e. all three bits are set to zero).
- For the current syllable the following phonological properties are encoded and fed to a 28 bit input: length (short / long) and intrinsic pitch (low / high) of the syllable nucleus; presence of a plosive in the onset or coda of the syllable; the syllable stress level and prosodic phrase type; first or last position of the syllable in word and phrase.

Figure 8 shows the complete architecture of the neural network used for F_0 control: there are 65 input nodes (55 for information of the phonological representation plus 10 for recurrent links), 20 nodes in the first hidden layer, 10 nodes in the second hidden

³ Such continuous F_0 contours have proved to be better suited for our modeling than standard ones, where unvoiced segments are generally set to zero.

layer that are reconnected to the input layer, and five output nodes. All nodes have a sigmoidal activation function. At the start of each utterance the feedbacks are set to zero, both in training and generation mode.

The outputs of the neural network are five continuous values between 0 and 1 denoting normalized F_{0N} values. These F_{0N} values are linearly mapped onto absolute F_0 values in a voice-dependent interval, e.g. for our female voice onto the interval 120 - 420 Hz. In accordance with the training data, the five output values are positioned at the syllable boundaries, at the nucleus boundaries and in the center of the nucleus (cf. Fig. 8).

The training of the RNN was done using the back-propagation through time algorithm. This algorithm is described in [Wer90]. The 800 sentences of the prosody database were partitioned into a training set of 600 sentences and a test set of 200

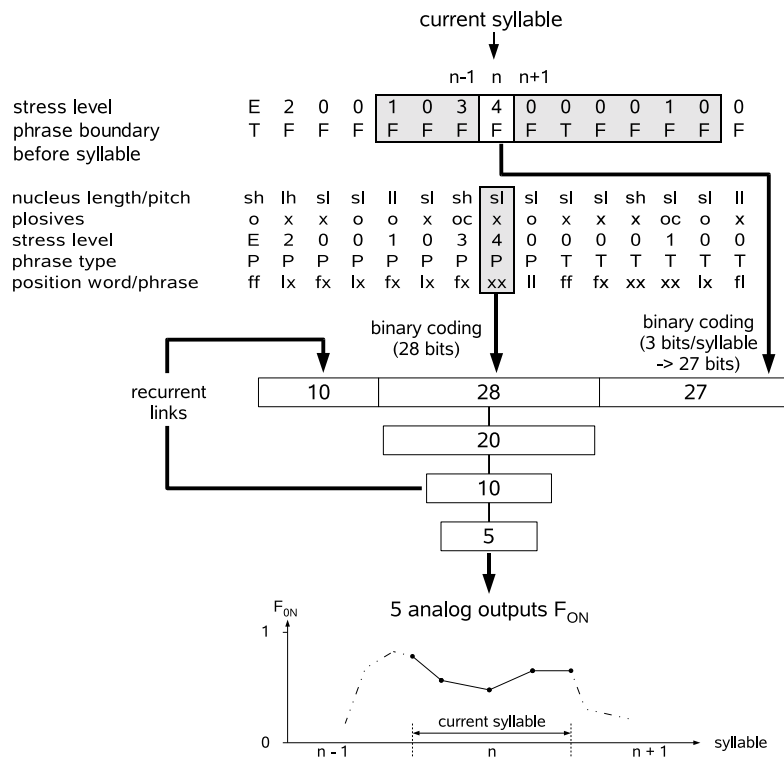


Fig. 8. Architecture plus input and output coding of the RNN used for F_0 generation: For each syllable five normalized F_{0N} values are generated using symbolic information derived from the phonological representation as input. The input information shown here has been derived from the phonological representation of the example of Fig. 7. Concatenating the F_{0N} values of all syllables yields the complete F_{0N} contour of the utterance to be synthesized.

sentences. As error measure we used the mean square error between the generated F_0 values and the corresponding natural F_0 values taken from the prosody database.

The mean square error proved to be no good measure for naturalness of synthetic F_0 contours. Therefore the quality of the F_0 generation model for German was finally evaluated in a listening test: 40 sentences which randomly carried either the natural or synthetic F_0 contour predicted by the RNN were presented to 12 listeners. Their task was to decide whether the F_0 contour of a heard sentence was natural or synthetic. The results showed that on average about 70 % of the decisions were correct. Note that the result of pure guessing would be 50 %. The test showed that listeners were largely unable to distinguish between synthetic and natural F_0 contours.

4.3 Prosody Control for Polyglot TTS

The approach to F_0 generation described above has shown to perform very well for TTS systems that have to synthesize speech from monolingual sentences. Our intention is to use it also for polyglot TTS synthesis, where prosody of several languages has to be considered within a single sentence.

Since prosody models are trained with an appropriately prepared prosodic database (see Sect. 4.1), this could be achieved simply by putting also a set of mixed-lingual sentences into this database. For flexibility reasons, we have decided to go another way. Our target is a polyglot TTS system that can be configured for an arbitrary set of languages. To have a prosody model for each of these possible sets would require quite a number of such models. Instead of specific models for each set of languages, we intend to combine the monolingual prosody models of those languages in some suitable manner.

Such a combination has to account for peculiarities of mixed-lingual sentences, of course. Very important is, e.g. that not only the language of a foreign inclusion, but also its length strongly influences the prosodic parameters.

5 Discussion and Conclusions

The aim of our research towards a polyglot TTS synthesis is to develop a system architecture which is as language-independent as possible and may be configured to virtually any set of languages. The configuration must completely be defined by the lingware (i.e. the language-dependent data set) that includes lexica, grammars, phonological rules, F_0 and duration model parameters, diphones, etc.

The lingware of the polyglot TTS synthesizer for a certain set of languages should primarily consist of the lingware of the individual languages and would ideally need no or only very few cross-lingual knowledge or data.

For the transcription stage of our polyglot TTS system this aim has been widely reached: the monolingual lexica, grammars and other rule sets can be combined arbitrarily and need only very small cross-lingual extensions (e.g. the inclusion grammars, see [PR03]). And given the diphones of several languages are from the same speaker, they can directly be used in the polyglot TTS synthesizer.

In the context of prosody control, however, there is still an number of remaining issues that will be subject to further research.

6 Acknowledgment

This work was partly supported by the Swiss National Science Foundation in the framework of NCCR IM2 and by the Swiss Commission for Technology and Innovation (CTI).

References

- [Bie66] M. Bierwisch. Regeln für die Intonation deutscher Sätze. *Studia Grammatica*, VII:99–201, 1966.
- [CH68] N. Chomsky and M. Halle. *The Sound Pattern of English*. Harper and Row, New York, 1968.
- [Kip66] P. Kiparsky. Über den deutschen Akzent. *Studia Grammatica*, VII:69–98, 1966.
- [Kos83] K. Koskeniemi. *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*. PhD thesis, Department of General Linguistics, University of Helsinki, Finland, 1983.
- [PR03] B. Pfister and H. Romsdorfer. Mixed-lingual text analysis for polyglot TTS synthesis. In *Proceedings of the Eurospeech 2003*, pages 2037–2040, Geneva, September 2003.
- [Rie98] M. Riedi. *Controlling Segmental Duration in Speech Synthesis Systems*. PhD thesis, No. 12487, Computer Engineering and Networks Laboratory, ETH Zurich (TIK-Schriftenreihe Nr. 26, ISBN 3-906469-05-0), February 1998.
- [Rom04] H. Romsdorfer. *An Approach to an Improved Segmentation of Speech Signals for the Training of Statistical Prosody Models*. Technischer Bericht Nr. 2 zum KTI-Projekt Nr. 6233.1 SUS-ET. Institut TIK, ETH Zürich, May 2004.
- [RP04] H. Romsdorfer and B. Pfister. Multi-context rules for phonological processing in polyglot TTS synthesis. In *Proceedings of the Interspeech 2004 - ICSLP*, Jeju Island (Korea), October 2004.
- [Spr96] R. Sproat. Multilingual text analysis for text-to-speech synthesis. In *Proceedings of the ICSLP'96*, Philadelphia, October 1996.
- [TP99] C. Traber, B. Pfister, et al. From multilingual to polyglot speech synthesis. In *Proceedings of the Eurospeech*, pages 835–838, September 1999.
- [Tra95] C. Traber. *SVOX: The Implementation of a Text-to-Speech System for German*. PhD thesis, No. 11064, Computer Engineering and Networks Laboratory, ETH Zurich (TIK-Schriftenreihe Nr. 7, ISBN 3 7281 2239 4), March 1995.
- [Wer90] P. J. Werbos. Backpropagation through time: What it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990.