# MODELING PRONUNCIATION VARIATION FOR SPONTANEOUS SPEECH SYNTHESIS

*Steffen Werner, Matthias Wolff, Matthias Eichner and Rüdiger Hoffmann*

Dresden University of Technology
Laboratory of Acoustics and Speech Communication
D-01062 Dresden, Germany

## ABSTRACT

Integration of pronunciation modeling into speech synthesis makes synthetic speech more natural and colloquial. Pronunciation variation as one observable effect in spontaneous speech is a step towards spontaneous speech synthesis.

In previous works [1, 2] we introduced different duration control methods in speech synthesis. These methods based on the observation that words, which are very likely to occur in a given context are pronounced faster and less accurate than improbable ones [3]. Therefore we use the probability of a word in its context either to control directly the local speaking rate or to select appropriate pronunciation variants to realize the change in the local speaking rate.

Extending these methods by a pronunciation sequence model, we involve knowledge about how well two subsequent variants fit together. With the here proposed algorithm we could further improve the natural and colloquial listening impression.

## 1. INTRODUCTION

In [1] we introduced a new approach for duration control in speech synthesis that uses the probability of a word in its context to control the local speaking rate. In listening tests 58 % of the synthesized utterances were rated better in terms of overall quality. However, in natural speech a greater speaking rate is rather produced by using reduced pronunciations instead of a faster articulation of canonical ones.

Therefore, we extend this approach in [2] by selecting appropriate pronunciation variants with different degree of reduction to consider the change in the local speaking rate. Although test participants rated the synthetic speech as more *natural* (54 %) and *colloquial* (74 %), the variant selection showed a strong tendency towards choice of unnatural variant sequences. We identified the combination of improper word boundary pronunciations as one reason for the occasionally bad listening impression.

To improve the naturalness of the synthesized variant sequences and to benefit from word boundary effects (elisions and assimilations) modeled by our pronunciation lexicon we investigated the use of a *pronunciation sequence model* for the selection of the pronunciation variants in combination with the word duration model introduced in [1, 2].

## 2. THE WORD PRONUNCIATION MODEL

We generate a variant lexicon automatically using the data driven pronunciation learning technique described in [4]. The canonical lexicon will be extended by phoneme hypotheses lattices which are generated by a phoneme recognizer for a transcribed training set. This method estimates probabilities for the pronunciation variants and is capable of extrapolating unseen variants. It also includes a post-processing step which removes statistically irrelevant and/or confusable variants from the lexicon.

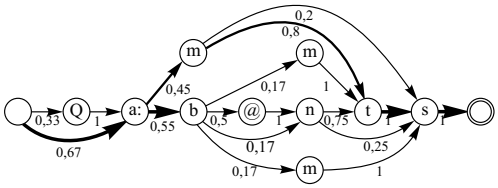


**Fig. 1**. Word pronunciation model for the German word "abends" (*in the evening*). This model was automatically learned from a read speech corpus with the method described in [4].

For the experiments described in this paper we selected a pronunciation lexicon with an average of 2.8 variants per word and performed a re-alignment of these variants to the speech signal of our training set. From the aligned variant sequences we built a variant bigram. So we obtained, in combination with the variant unigram probabilities contained in the lexicon, an interpolated zero-through-bigram pronunciation sequence model (see section 3.2).

## 3. ALGORITHM

The variant selection algorithm has two major stages. First, we calculate a target duration for each word of the utterance to be synthesized. Then we select an appropriate sequence of pronunciation forms. These pronunciation variants should on the one hand match the target durations well *and* form a probable sequence according to the pronunciation sequence model on the other hand. To find an optimal

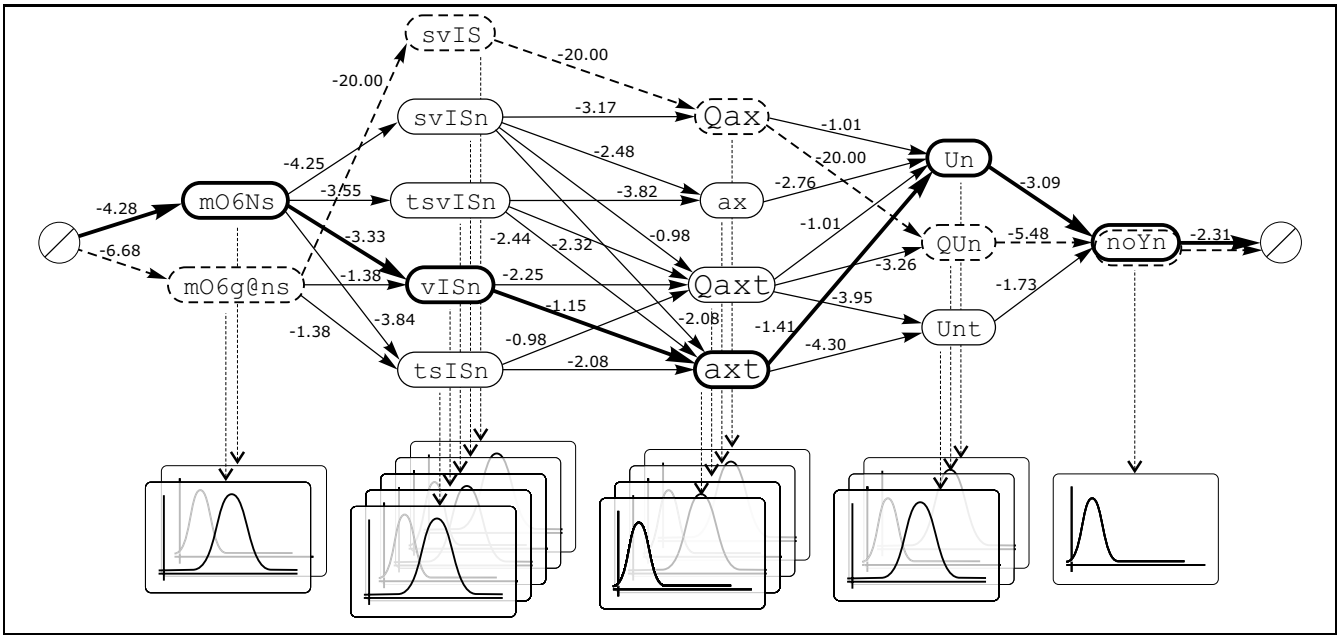


**Fig. 2**. Stochastic Markov Graph (SMG) representing a network of pronunciation variants for the German phrase "morgens zwischen acht und neun" (*between eight and nine in the morning*). Nodes represent word pronunciations, edges carry weights obtained from the pronunciation sequence model. The bold path denotes the pronunciations selected using the variant sequence model. The example shows correct sequencing of word boundary effects (e.g. elision of /t/ and assimilation of /s/ between the first two words). For comparison, the dashed edges show the path chosen considering target durations only [2].

pronunciation sequence we build a stochastic Markov graph for each utterance to be synthesized (Figure 2). Each node of that graph stands for a single pronunciation and links to an unidimensional Gaussian probability density function describing the duration of the variant. The edges of the graph carry transition weights taken from the sequence model.

### 3.1. The Word Duration Model

We use the word duration model introduced in [1] and [2]. This section gives just a very brief overview of the method. Please see the former papers for any details.

We start with calculating relative word durations $r(w)$ for each word of the utterance $U = \{w_1 \circ \cdots \circ w_N\}$ to be synthesized:

$$r(w) = \frac{\mathrm{sgn}(P(w)-\overline{P})+1}{2}\left[\frac{r_{min}-1}{1-\overline{P}}\left(P(w)-\overline{P}\right)+1\right] + \frac{\mathrm{sgn}(\overline{P}-P(w))+1}{2}\left[\alpha\frac{r_{min}-1}{1-\overline{P}}\left(P(w)-\overline{P}\right)+1\right] \quad (1)$$

with

$$\overline{P} = \frac{1}{|U|}\sum_{w\in U} P(w), r_{min} = 0.5, \alpha = 0.1 \quad (2)$$

$P(w)$ denotes the language model probability of the word $w \in U$ related to its left ($n$-gram) *and* its right (reverse $n$-gram) context [3]:

$$P(w) = \sum_{i=-l}^{i=l} f_i P(w_n | w_{n-i} \circ \cdots \circ w_{n-1}) \quad (3)$$

where $P(w_n|\cdot)$ denotes a single word $n$-gram (of order $i$) taken from the normal ($i < 0$) or reverse ($i > 0$) language model and $f_i$ denotes the multigram interpolation weights with $\sum f_i = 1$ and $f_{-1} = 0$.

The relative word durations $r(w)$ are filtered and smoothed in a post-processing step. From the smoothed relative word durations we derive absolute target durations $d(w)$:

$$d(w) = (r(w) - \beta) \cdot d(A_{can,w}) \quad (4)$$

with $d(A_{can,w})$: duration of the canonical pronunciation of $w$ (estimated from a phoneme duration statistic).

### 3.2. The Pronunciation Sequence Model

In the second stage we select an appropriate sequence of pronunciation variants for the word durations calculated in the first stage. As stated above we use a sequence model, more precisely an *interpolated $n$-multigram* [5, 6], which estimates the probability of a sequence $\tilde{A}$ of pronunciation forms $A_i$ from single pronunciation $n$-grams:

$$P(\tilde{A}) = \prod_{i=1}^{|\tilde{A}|}\left[\sum_{n=1}^{N} f_n P(A_i | A_{i-n+1} \circ \cdots \circ A_{i-1})\right] \quad (5)$$

with $\sum f_n = 1$. $F = \{f_0, \cdots f_N\}$ denotes the $n$-gram interpolation vector, $P(A_i|\cdot)$ denotes a single $n$-gram probability and $N$ represents the maximal $n$-gram order.

Let $U = \{w_1 \circ \cdots \circ w_N\}$ be a sequence of words to be synthesized and $\mathcal{A}(w_i)$ be the set of pronunciation variants of the word $w_i$. Then we can express the pronunciation

model $\mathcal{G}$ of this word sequence as a stochastic Markov graph (SMG, [7]):

$$\mathcal{G} = \left\{ V, E, \{\mathcal{N}\}, \nu^{(V)}, \pi^{(E)} \right\} \tag{6}$$

with the node set $V$, the edge set $E$, a set of unidimensional Gaussian distributions $\{\mathcal{N}\}$ describing the duration of the pronunciation variants and two maps $\nu^{(V)} : V \rightarrow \{\mathcal{N}\}$ assigning a Gaussian to each node and $\pi^{(E)} : E \rightarrow \mathbb{R}$ assigning a transition weight (see equation (9)) to each edge.

The node and edge sets are constructed as follows:

$$V = \bigcup_{w_i \in W} \mathcal{A}(w_i) \tag{7}$$

$$E = \bigcup_{w_i \in W} \mathcal{A}(w_i) \times \mathcal{A}(w_{i-1}) \tag{8}$$

Each edge stands for the transition from the pronunciation variant $A_s$ denoted by its initial node $v_s$ to the variant $A_e$ denoted by its terminal node $v_e$. Figure 2 shows an example of a pronunciation SMG for a German phrase ($W = \{$ morgens $\circ$ zwischen $\circ$ acht $\circ$ und $\circ$ neun $\}$). As $\mathcal{G}$ is of first order, the maximal $n$-gram order to be included into the edge weight is two. So we weight the edges by an interpolation of zerograms, unigrams and bigrams of pronunciation variants (5):

$$\begin{aligned} w(v_s, v_e) &= \ln\left(P(A_e|A_s)\right) \\ &= \ln\left(f_2 P(A_e|A_s) + f_1 P(A_e) + f_0 P_0\right) \end{aligned} \tag{9}$$

where $f_0$ through $f_2$ denote the zerogram, unigram and bigram weigthing factors and $P_0$ denotes the zerogram probability.

Of course, the usage of higher order SMG's is also possible. However, gathering statistically sound $n$-grams of pronunciation forms requires a huge database.

### 3.3. Selection of Pronunciation Variants

Given the desired absolute lengths $d_i = d(w_i)$ (4) for the word $w_i$ in a sentence or phrase to be synthesized, the optimal sequence of pronunciation variants is:

$$\mathcal{A}^* = \arg\max_{\mathcal{A} \in \mathcal{G}} \sum_{A_i \in \mathcal{A}} \left[ w(A_i|A_{i-1}) + \gamma \ln p(d_i|\mathcal{N}_i) \right] \tag{10}$$

where $w(A_i|A_{i-1})$ is the edge weight of the transition $A_{i-1}A_i$ (9) and $p(d_i|\mathcal{N}_i)$ for the probability density of the desired word length $d_i$ in the duration statistic of a pronunciation variant $A_i$:

$$\begin{aligned} p(d_i|\mathcal{N}_i) &= p(d_i, \mathcal{N}(,)) \\ &= \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2}\frac{(d_i - \mu_i)^2}{\sigma_i^2}} . \end{aligned} \tag{11}$$

where $\mu_i$ is the value of the measured length of a pronunciation variant and $\sigma_i$ represents the standard deviation. The scaling factor $\gamma$ is used to adjust the preference for an exact match of the required word durations ($\gamma < 1$) or for the selection of probable variant sequences ($\gamma > 1$). In our experiments we set $\gamma$ to $0.85$.

## 4. EXPERIMENTS

We used our multilingual, diphone based, time domain synthesis system Dress [8] for evaluation of the proposed variant selection algorithm, which uses a variant bigram.

### 4.1. Evaluation

For the evaluation we held the same perceptive pair comparison test as in [2]. 30 persons had to judge each pair of sentences in three categories: intelligibility, naturalness and colloquial speech. Five participants in the test work in the field of speech processing and are experienced listeners; the remaining ones took part as naive listeners.

To obtain results comparable to the results to the experiments in [2] we used the same 25 sentences from the PHONDAT II data base for evaluation. All sentences were synthesized with and without the variant selection algorithm. The results of the evaluation are shown in Table 1.

The evaluation of the category intelligibility showed clearly that most of the participants (76.4 %) voted in favor of the canonical synthesis. 73.6 % of the sentences synthesized by the proposed algorithm were rated more colloqial, and 54.5 % more natural. Due to the new algorithm with a variant bigram we achieved better results in all three categories as compared to the pure variant selection according to [2].

In order to check the overall impression of the synthesized speech samples, we conducted an evaluation with absolute category rating (ACR). Therefore we chose randomly 15 sentences from the former experiment. The number of the utterances was reduced in order to hold the effort for the listeners at an acceptable level. We asked 30 persons to rate these 15 sentences, which were synthesized in three different manners: canonical, with variant selection according to [2] and with variant selection by using variant bigram probabilities according to the proposed algorithm. The test participants could rate the sentences on an MOS scale between 1 (worst) and 5 (best) points.

The difference between the considered algorithms lies in a variance of 0.3 points (compare Table 1). The canonical synthesis was slightly preferred with an MOS score of

**Table 1**. Results (in %) of the listeners' preference in the pair comparison test for the synthesis with I) pronunciation sequence modeling and II) pure variant selection (according to [2]) instead of the canonical synthesis. The MOS rating of the canonical synthesis is shown in brackets.

| | I) pronunciation sequence modeling | II) pure variant selection [2] |
|---|---|---|
| intelligibility | 23.6 % | 20.8 % |
| naturalness | 54.5 % | 53.7 % |
| colloquial speech | 73.6 % | 72.4 % |
| MOS rating | 2.84 (can. 3.08) | 2.80 (can. 3.08) |

**Table 2**. Ratings depending on the length of an utterance for synthesis with pronunciation sequence modeling

| | utterances shorter than 2 seconds | utterances longer than 3 seconds |
|---|---|---|
| intelligibility | 12.4 % | 22.2 % |
| naturalness | 46.2 % | 57.0 % |
| colloquial speech | 70.5 % | 72.2 % |
| MOS variant / (canonical) | 2.63 / (3.34) | 2.80 / (2.98) |

3.08 as opposed to the variant selection with bigram model (MOS: 2.84) and the variant selection according to [2] (MOS: 2.80).

### 4.2. Discussion

Some test participants relate the category naturalness to the intelligibility and some to the colloquial speech. Hence, the MOS rating in the ACR test corresponds sometimes to the rating in the category naturalness of the pair comparison test and other times to the category intelligibility.

Because of the different results of the pair comparison and the MOS evaluation we investigated the synthesized samples more deeply and we conceived several ideas for further research:

- The most diphone databases were developed by storing the diphones according to the canonical pronunciation. This is a reason why most diphones match the canonical form better than the pronunciation variants. There should exist many more variants of differently pronounced diphones.
- Most of the participants rate long utterances better than short ones. Table 2 compares the ratings depending on the length of an utterance. In contrast to long sentences, short ones often received a lower score in the MOS and in the categories intelligibility and naturalness. This is obvious, since a longer context provides more information for correctly understanding speech. The length of an utterance should be used as an additional parameter to consider when shortening or lengthening a word.
- The listeners often rate utterances with slightly reduced variants as more natural than those containing strongly reduced forms. This is even more noticeable if the reduction is done by omitting phonemes mainly in the middle of a word as shown in Figure 3, example 2. On the contrary, omitting phonemes in word transitions often results in a higher score (Figure 3, example 1).
- The content of the utterance is very important for the acceptance of variants in the synthesized sentence. Our test corpus included mostly utterances from the field "travel information". We suppose that for such a sphere a canonical realization is more adequate. Investigations on content to speech concepts could confirm that content important words should not be reduced. Measuring the listening effort with an ACR test would show how content included utterances influence the rating.

| | | | | | | |
|---|---|---|---|---|---|---|
| Example 1: | geht | es | nicht | eher | | |
| varaints: | ge:t | s | nIC | Qe:6 | | |
| canonical: | ge:t | QEs | nICt | Qe:6 | | |
| Example 2: | ich | will | morgen | abend | nach | fankfurt |
| variants: | QIC | vIl | mO6N | Qa:bm | na:x | fraNfU6t |
| canonical: | QIC | vIl | mO6g@n | Qa:b@nt | na:x | fraNkfU6t |

**Fig. 3**. Examples of reduced utterances with thefollowing ratings in the category naturalness: Ex.1: 86.7 %; Ex.2: 23.3 % (Ex.1: *Is there nothing earlier*; Ex.2: *I want to go tomorrow evening to Frankfurt*)

## 5. CONCLUSION

The use of pronunciation variants in combination with a variant sequence model in speech synthesis improves the spontaneous listening impression of the synthesized utterances. The proposed algorithm selects a variant for a given word by considering the variant selection for the surrounding words.

Test participants accept synthesized speech better if the word boundaries are explicitly modeled by pronunciation sequences. But a too strong reduction and a reduction of content words have a bad influence on the listening impression. A model of these facts could contribute to an additional improvement in spontaneous speech synthesis.

## 6. REFERENCES


[1] M. Eichner, M. Wolff, and R. Hoffmann, "Improved duration control for speech synthesis using a multigram language model," in *Proc. ICASSP*, 2002, vol. 1, pp. 417–420, Orlando, FL, USA.

[2] M. Eichner, S. Werner, M. Wolff, and R. Hoffmann, "Towards spontaneous speech synthesis – LM based selection of pronunciation variants," in *Proc. ICASSP*, Apr. 2003, Hong Kong, PR China.

[3] D. Jurafsky, A. Bell, M. Gregory, and W. D. Raymond, "The effect of language model probability on pronunciation reduction," in *Proc. ICASSP*, 2001, vol. 2, pp. 801–804, Salt Lake City (USA).

[4] M. Eichner and M. Wolff, "Data driven generation of pronunciation dictionaries in the German Verbmobil project - discussion of the experimental results," in *Proc. ICASSP*, Istanbul, Turkey, June 2000, pp. 1687–1690.

[5] F. Bimbot et al., "Variable length sequence modeling: theoretical foundation and evaluation of multigrams," *IEEE Signal Processing Letters*, vol. 2(6), 1995.

[6] E. Schukat-Talamazzini, *Automatische Spracherkennung*, p. 221f, Vieweg Verlag, Braunschweig/Wiesbaden, 1995.

[7] F. Wolfertstetter and G. Ruske, "Structured Markov models for speech recognition," in *Proc. ICASSP*, 1995, pp. 544–547, Detroit.

[8] R. Hoffmann, "A multilingual text-to-speech system," *The Phonetician 80, (1999/II)*, pp. 5–10, 1999.