TOWARDS SPONTANEOUS SPEECH SYNTHESIS - LM BASED SELECTION OF PRONUNCIATION VARIANTS

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

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

ABSTRACT

State of the art speech synthesis systems achieve a high overall quality. However, the synthesized speech still lacks naturalness. To make speech synthesis more natural and colloquial we are trying to integrate effects that are observable in spontaneous speech. In a previous paper 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 within the utterance. This idea is based on the observation that words that are very likely to occur in a given context are pronounced faster than improbable ones. Since probable words are not only pronounced faster but also less accurate we extend this approach by selecting appropriate pronunciation variants to realize the change in the local speaking rate.

1. INTRODUCTION

Jurafski et al. showed in [1] and [2], that the local speaking rate of a word in an utterance is correlated with the language model probability of that word. Probable words are frequently pronounced faster and less accurate than less probable words. The correlation between reduction and a particular n-gram depends on the word type [1]. While ordinary (left bound) bigram and trigram probabilities correlate with the reduction of function words, the reverse bigram correlates with a pronunciation reduction of content words.

Basing on the results of this study, we implemented a language model driven speaking rate control into our speech synthesizer [3]. In listening tests 58% of the synthesized utterances were rated “better” in terms of overall quality [4].

Even though these results were encouraging, they also showed that modifying the speaking rate only by shortening or lengthening syllables and phones is a too simple approach. In natural speech a greater speaking rate is rather produced by using reduced pronunciations instead of faster articulation of canonical ones. Slow speech does not necessarily mean to lengthen phones, but rather to pronounce more accurately (or canonically) and to insert more and longer pauses.

In this paper we describe a possible solution for making synthetic speech sound more natural and colloquial. In contrast to [4] we abandoned the direct modification of the syllable durations in favor of realizing the target word durations predicted by the language model by selecting appropriate (reduced) pronunciations from a variant lexicon. Thus we still indirectly control the speaking rate by directly controlling the grapheme to phoneme conversion using the language model.

In an informal listening experiment 74% of the generated utterances were rated “more colloquial” compared to utterances produced by our standard synthesis system.

The use of a language model and a variant lexicon integrates in our approach to a unified system for speech synthesis and recognition [5]. The idea is to design a system that uses the same algorithms and databases for both speech synthesis and speech recognition.

2. ALGORITHM

2.1. The Databases

As the proposed algorithm is entirely data-driven, we represent the required knowledge by databases. We use a multigram language model and a pronunciation lexicon along with a phone duration statistic.

2.1.1. The Language Model

Like in our previous work [4] we use an interpolated n-gram language model. The model order ranges from -3 to 3 (negative orders denote reverse n-grams). The language model was trained with texts from the German Verbmobil speech data base [6]. The training data were selected from

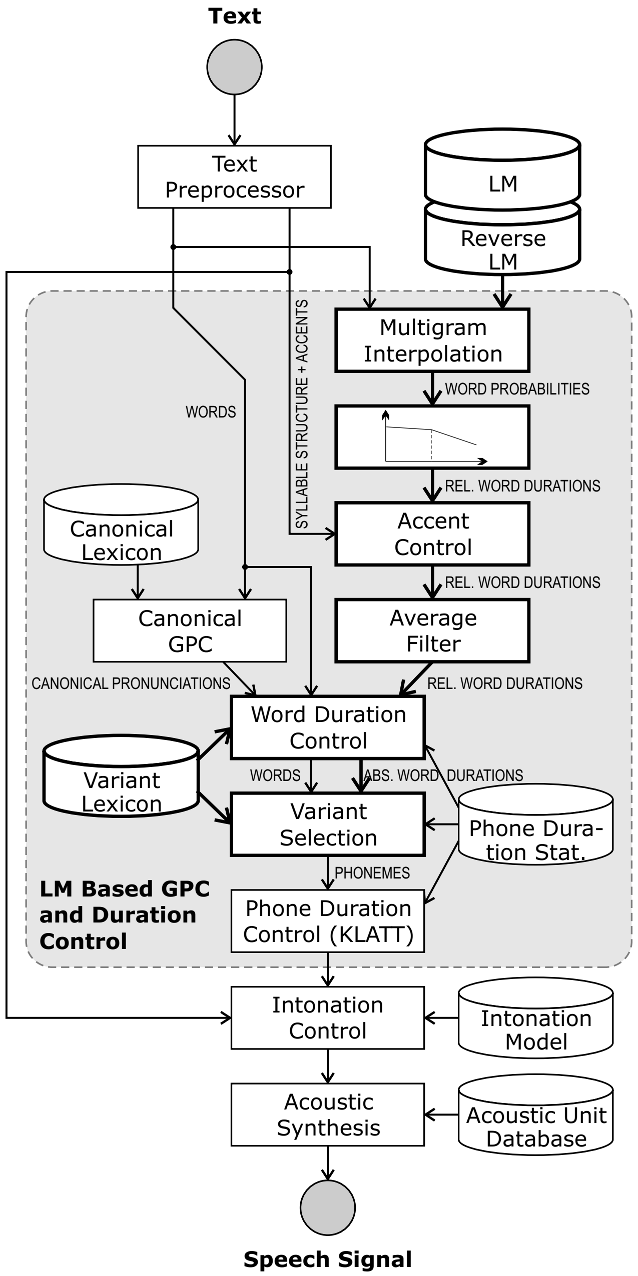

Figure 1: Integrated duration control and grapheme to phoneme conversion driven by language model

a limited domain and contain a total of 177,625 words with a vocabulary of 4831 words. All multigram interpolation weights were set to $^1/_{15}$ except for the reverse bigram ($^2/_3$) and the zerogram (0).

2.1.2. The Pronunciation Lexicon

We use a variant lexicon which was automatically generated with our pronunciation learning technique described in [7]. To achieve an optimal quality of the variants we used a manually labeled read speech corpus taken from the German PHONDAT II material as training database. It consists of a total of 7310 words with a vocabulary size of 192. This dictionary was hand optimized by removing obviously wrong pronunciations and unusual variants. The average number of pronunciation variants per word is 3.7 in the final lexicon.

Each pronunciation variant $A=\{a_1 \circ \ldots \circ a_N\}$ is annotated with an estimated duration $d(A)$.

$$d(A)=\sum_{a\in A}\bar{d}(a) \qquad (1)$$

where a denotes a phoneme and $\bar{d}(a)$ the average duration of the respective phone. The phone duration statistic was trained with phones from our synthesis unit database.

2.2. Word Duration Control

First we calculate a relative duration r for each word w in the utterance $U=\{w_1 \circ \ldots \circ w_N\}$ to be synthesized. As the calculation is nearly identical to the strategy described in [4], we will only give a short overview here. Please refer to the former paper for more details.

The language model probability is used to calculate an initial relative word duration $r_{LM}(w)$ for each word.

$$\begin{aligned} r_{LM}(w) = {} & \frac{\operatorname{sgn}(p(w)-\bar{p})+1}{2}\left[\frac{r_{min}-1}{1-\bar{p}}(p(w)-\bar{p})+1\right] + \\ & \frac{\operatorname{sgn}(\bar{p}-p(w))+1}{2}\left[a\frac{r_{min}-1}{1-\bar{p}}(p(w)-\bar{p})+1\right] \end{aligned} \qquad (2)$$

with:

$$\bar{p}=\frac{1}{|U|}\sum_{w\in U}p(w),\ r_{min}=0.5,\ a=0.1 \qquad (3)$$

where $p(w)$ denotes the language model probability of the word. $\bar{p}$ is the average of all $p(w)$ in the utterance.

In a subsequent accent control step we reset the relative duration to 1 for all accented words that are shortened by the language model. This is done in order to preserve the accent structure of the utterance. In contrast to [4] the accent control is done on word instead of syllable level here. An accent control on the syllable level is no longer necessary because we do not modify the word durations other than indirectly through pronunciation variants anymore.

$$r_{ACC}(w)=\begin{cases} w \text{ not accented} \vee r_{LM}(w)>1 & : r_{LM}(w) \\ w \text{ accented} \wedge r_{LM}(w)\le 1 & :1 \end{cases} \qquad (4)$$

After the accent control step the relative durations are smoothed by using an average filter.

$$r_F(w_n) = \frac{1}{2L+1} \sum_{i=n-L}^{n+L} F(i) \cdot r_{ACC}(w_i) \qquad (5)$$

with:

$$\sum_{i=-L}^{L} F(i) = 1 \qquad (6)$$

We use a filter length of L=1. The center weight is $^5/_7$, the other weights $^1/_7$.

2.3. Variant selection

From the relative word duration we derive an absolute target duration $d_0(w)$. The calculation assumes that a relative word duration of 1 corresponds to the canonical pronunciation of the word. The duration of the canonical form $d(A_{can,w})$ can be estimated according to equation (1).

$$d_0(w) = [r_F(w) - \alpha] \cdot d(A_{can,w}) \qquad (7)$$

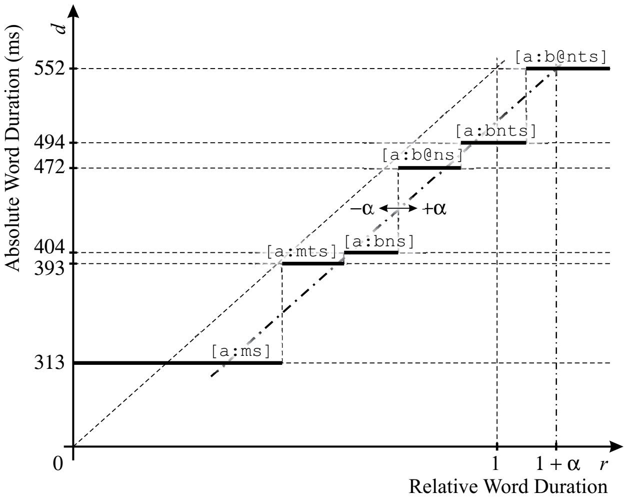

Figure 2: Example of variant selection for the word "abends" (German: "in the evening") with 6 pronunciations

The parameter α controls the preference of shorter or longer pronunciations. In our experiments we set $\alpha = 0.3$ which means that we slightly prefer short variants and choose the canonical one (typically the longest) only when the language model probability of the word is significantly less than the average $\overline{p}$ (see equation 3).

The actual selection of a pronunciation variant takes place by minimizing the distance between the target duration and the duration of the variants:

$$A^* = \underset{A \in \underline{\boldsymbol{A}}(w)}{\arg\min} \left| d(A) - d_0(w) \right| \qquad (8)$$

where A^* denotes the chosen pronunciation variant and $\underline{\boldsymbol{A}}(w)$ represents the set of variants for word w.

3. EXPERIMENTS

We used our multilingual, diphone based, time domain synthesis system DreSS [3] for evaluation of the proposed algorithm. Several approaches for intonation control in the system are available. We used an adaptation of the Fujisaki model for the experiments, since this model yielded the best results in previous evaluations.

3.1. Evaluation

To evaluate the performance of the proposed method we performed a perceptive pair comparison test and asked for the preference in the three categories: intelligibility, naturalness and colloquial speech.

For this purpose 25 sentences were selected from the PHONDAT II data base. They were synthesized with and without the proposed selection algorithm for pronunciation variants. 20 persons were asked to judge each pair of sentences in the mentioned three categories. Six participants in this test work in the field of speech processing and are experienced listeners; the remaining ones took part as naive listeners.

The evaluation yielded a slight improvement of the synthesis quality in the category of naturalness and a considerable improvement in the category of colloquial speech. 74% of the samples generated using the proposed algorithm where rated as more colloquial (only 54% as more natural) compared to samples generated with the conventional, canonical form based system. However, most of the listeners (79%) decided that the synthesis using canonical pronunciation is more intelligible than the one using different pronunciation variants. That is not surprising, because we can expect the over-articulated canonical form more intelligible than any reduced form.

Although the number of participants in the evaluation test is not sufficient for a statistically proved prediction, the experiment shows a tendency and can be interpreted as a preliminary result.

3.2. Discussion

The results show that even a quite simple approach to include pronunciation variants into synthetic speech is capable of making it sound much more "spontaneous". However, we were not able to enhance the "naturalness" to the same extent.

There are some possible explanations for the difference between the perception of natural and colloquial speech:

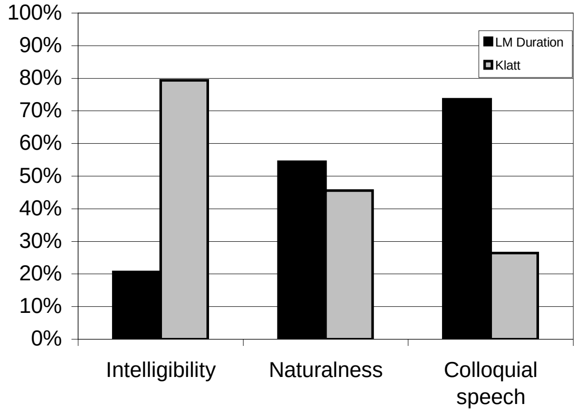

Figure 3: Result of the pair comparison test. LM duration stands for the proposed algorithm, Klatt is the standard duration control.

- We select pronunciation variants only by their duration. There is no knowledge involved about how well two subsequent variants *fit together*. Further not every variant can be chosen at any position of an utterance, e.g. shorter pronunciation variants should not be used in the first syllable of a sentence or between two improbable content words. These problems could be solved by using a variant bigram.
- The reduction of a word depends on different n-grams. We take this into account by using a multigram interpolation technique. However, we use the same n-gram weighting vector for all word types which is, according to [1], not quite correct. A different set of weights at least for function and content words should result in a more accurate prediction of the pronunciation reduction.
- As mentioned above, the usage of reduced pronunciations makes the synthetic speech less intelligible. Especially in very short sentences some reduced pronunciations were actually hard to understand. This effect seems to be stronger in synthetic than in natural speech and may compromise the naturalness.
- There is a strong interaction between duration and intonation which is not modeled very well by our current synthesis system. The usage of an adapted F0 contour should improve the naturalness as well as the intelligibility of the synthetic speech. A solution lies in extending our integrated prosody model [8] by the language model driven duration control.

4. CONCLUSION

The use of pronunciation variants in speech synthesis is a first step towards spontaneous synthesis systems. We showed that the use of variants makes the synthetic speech definitely more colloquial. The intelligibility decreased compared to the canonical synthesis. This effect is not surprising and can be observed in natural speech as well.

The proposed method selects the variant for a given word without considering the variants selected for the surrounding words. Although the chosen variants are valid variants, a human speaker would often not use them in a row. Obviously there is a relationship between pronunciation variants. That means that a certain variant implies the use of another variant. Our further work will focus on modeling those dependencies.

The use of pronunciation variants is just one observable effect in spontaneous speech. To make synthetic speech really spontaneous other effects like hesitations have to be modeled too.

5. REFERENCES

[1] Bell, A., Gregory, M. L., Brenier, M. L., Jurafsky, D., Ikeno, A., and Girand, C. “Which Predictability Measures Affect Content Word Durations?” *Proc. PMLA*, Estes Park (USA), 2002.

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

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

[4] Eichner, M., Wolff, M. and Hoffmann, R. “Improved duration control for speech synthesis using a multigram language model”, *Proc. ICASSP*, Orlando, Florida (USA), pp. 417 – 420, 2002.

[5] Eichner, M., Wolff, M., Hoffmann, R., “A unified approach for speech synthesis and speech recognition using Stochastic Markov Graphs”, *Proc. ICSLP*, Beijing (China), vol. 1, pp. 701-704, 2000.

[6] Wahlster, W. (Ed.), *Verbmobil: Foundations of Speech-to-Speech Translation*, Berlin etc.: Springer, 2000.

[7] Eichner, M., and Wolff, M., “Data driven generation of pronunciation dictionaries in the German Verbmobil project – discussion of experimental results”, *Proc. ICASSP*, Istanbul (Turkey), pp. 1687 – 1690, 2000.

[8] Mixdorf, H., Jokisch, O., “Building An Integrated Prosodic Model of German“, *Proc. EUROSPEECH*, Aalborg (Denmark), vol. 2, pp. 947-950, 2001.

[9] Klatt, D. H., “Review of text-to-speech conversion for English“, *J. Acoustic. Soc. Am., 88*, pp. 737-793, 1987.

[10] Deligne, S., Yvon, F.; Bimbot, F., “Introducing Statistical Dependencies and Structural Constraints in Variable-Length Sequence Models.”, *ENST – Dept. Signal & Dept. Informatique*, Paris, 1996.

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