PRONUNCIATION VARIANT SELECTION FOR SPONTANEOUS SPEECH SYNTHESIS – LISTENING EFFORT AS A QUALITY PARAMETER

Steffen Werner, Matthias Wolff and Rüdiger Hoffmann

Dresden University of Technology
Laboratory of Acoustics and Speech Communication, D-01062 Dresden, Germany
{ *steffen.werner* | *matthias.wolff* | *ruediger.hoffmann* } *@ias.et.tu-dresden.de*

ABSTRACT

In previous works (see for instance [1]) we introduced different duration control methods in speech synthesis. The most outstanding approach is to control the grapheme to phoneme conversion (and thus indirectly control the speaking rate) by selecting (reduced) pronunciation variants according to a pronunciation variant sequence model.

Listeners would only accept long synthesized utterances if the listening effort is nearly the same as the one when listening to natural speech. To evaluate the quality of the variant synthesis compared to the canonical one (as the state-of-the-art system), we performed a listening test with two different synthesis systems. The variant synthesis applying a pronunciation variant sequence model achieved a significant lower listening effort and a higher overall rate (MOS) compared to the canonical synthesis.

We also show that the listening effort can act as a quality parameter for a speech sample. The rating for the listening effort is correlated with the rating of the naturalness and intelligibility of synthesized speech sample.

1. INTRODUCTION AND MOTIVATION

Jurafski et al. showed in [2] and [3] that the local speaking rate of a word in an utterance is correlated with the language model probability of that word. Probable words are often pronounced faster and less accurately than less probable ones.

Based on these results a language model (LM) driven speaking rate control was integrated into our TTS system DRESS [4] as a first approach. In a second development stage we generated the target word durations predicted by the language model by selecting appropriate pronunciation variants with different degree of reduction from a variant lexicon. This was done since a greater speaking rate is rather produced by reduced pronunciations instead of a faster articulation of canonical ones. Thus, the speaking rate is still indirectly controlled by controlling the grapheme to phoneme conversion using the language model.

After we had identified the improper word boundary pronunciations as one reason for the occasionally bad listening impression, we involved knowledge about how well two subsequent variants fit together. Therefore we investigated a pronunciation variant sequence model (PVSM) for the selection of the pronunciation variants in a third development stage.

In different pair comparison tests all development stages brought an improvement of the naturalness of the synthesized utterances. The most outstanding approach is the indirect control of the speaking rate by selecting (reduced) pronunciation variants according to a PVSM. It showed that 73.7 % of the listeners rated the generated utterances as more colloquial and 54.4 % as more natural [1].

After having summarized the synthesis approach applying a PVSM in Section 2, we present the results (Section 3.2) of a new large listening test (Section 3.1) focused on measuring different parameters of synthesized canonical speech and of synthesized speech using pronunciation variants. We discuss the results (Section 3.2) and analyze the impact on different parameters (Section 3.3).

2. SPONTANEOUS SPEECH SYNTHESIS APPLYING A PRONUNCIATION VARIANT SEQUENCE MODEL

In this section we give a brief overview of the synthesis applying a pronunciation variant sequence model (PVSM). A more detailed description can be found in [1].

2.1. Database

We used three main databases: (I) a general language model, (II) a pronunciation lexicon, and (III) a pronunciation variant sequence model.

The general language model (I) as well as the PVSM (III) are interpolated n-multi-gram models with different interpolation weights and training sets. The probability $P(w)$ of a particular word w (e.g. a certain variant) in an utterance in relation to the preceding and following words is estimated by interpolating the n-grams of different orders [5].

For (I) the training data were selected from the German Verbmobil corpus and contain a total of 177,625 words with a vocabulary of 4,831 words. In our experiments the model order ranges from -3 to 3 (negative orders denote reverse n-grams).

The pronunciation lexicon (II) was automatically generated with the help of our pronunciation learning technique described in [6, 7] using the manually labeled read speech corpus German PhonDatII. It consists of a total of 7,310 words with a vocabulary size of 192. The resulting dictionary was manually optimized by removing obviously wrong pronunciations and unusual variants. The average number of pronunciation variants per word in the final lexicon is 2.8. This final pronunciation lexicon was re-aligned to the speech corpus. From the aligned variant sequence a variant n-multi-gram (the PVSM) (III) was built. The model order ranges from 0 to 2.

2.2. Algorithm

When pronunciation variants for synthesis are used, it is a major task to model the word boundary effects (elisions and assimilations). Former experiments with variant selection showed that the combination of improper pronunciation variant sequences yields a worse listening impression [1].

The variant selection algorithm selects an appropriate sequence of pronunciation forms. On the one hand it should match the target durations well *and* on the other hand it should form a probable sequence according to the pronunciation variant sequence model (PVSM).

To find an optimal pronunciation variant sequence we used the following algorithm:

- Calculate language model probability.
- Calculate the initial local speaking rate in terms of an initial relative word duration from the word probability estimated by the language model.
- Process accent information. Accented syllables serve as anchors in speech production. To preserve the accent property of a syllable and in this way the accent structure of the whole utterance, do not *shorten* accented syllables, even if the language model suggests it.
- Calculate target duration from the relative duration (a relative duration of 1 corresponds to the canonical variant).
- Build a stochastic Markov graph (SMG, [8]) for each utterance to be synthesized (see Figure 1). Each node of that graph stands for a single pronunciation and links to a unidimensional Gaussian probability density function describing the duration of the variant. The edges of the graph carry transition weights taken from the sequence model.
- Search the best path through that graph and select the respective variant lying on that way.

2.3. Selection of Pronunciation Variants

The selection of the pronunciation variants of a word sequence according to the variant sequence model can be expressed as a stochastic Markov graph. An n-gram of 2^{nd}-order implies an SMG of first order. Each node of that graph stands for a single pronunciation and links to a unidimensional Gaussian probability density function describing the duration of the variant. Each edge stands for the transition from the pronunciation variant A_s to the variant A_e. Figure 1 shows an example of a pronunciation SMG for the German phrase { morgens ∘ zwischen ∘ acht ∘ und ∘ neun }. The edges are weighted by an interpolation of zerograms, unigrams and bigrams of pronunciation variants:

$$\nu = \ln\left(f_2 P(A_e|A_s) + f_1 P(A_e) + f_0 P_0\right) \tag{1}$$

where f_n denotes the according n-gram weighting factor, and P_0 denotes the zerogram probability.

Of course, the usage of higher order SMGs is also possible. However, gathering statistically sound n-grams of pronunciation forms would require a huge database.

Given the desired absolute lengths $d_{tar}(w_i)$ for the word w_i in a sentence or phrase to be synthesized, the optimal sequence of pronunciation variants can be found by searching the best path through the SMG-model, which is given by:

$$\mathcal{A}^* = \arg\max_{\mathcal{A}\in\mathcal{G}} \sum_{A_i\in\mathcal{A}} \left[\nu_i(A_i|A_{i-1}) + \gamma \ln p(d_{tar}(w_i)|\mathcal{N}_i)\right] \tag{2}$$

where $\nu_i(A_i|A_{i-1})$ is the edge weight of the transition $A_{i-1}A_i$ and $p(d_{tar}(w_i)|\mathcal{N}_i)$ stands for the probability density of the desired word length $d_{tar}(w_i)$ in the duration statistics of a pronunciation variant A_i. By including the scaling factor γ it is possible 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 γ to 0.85. For more details please see [1].

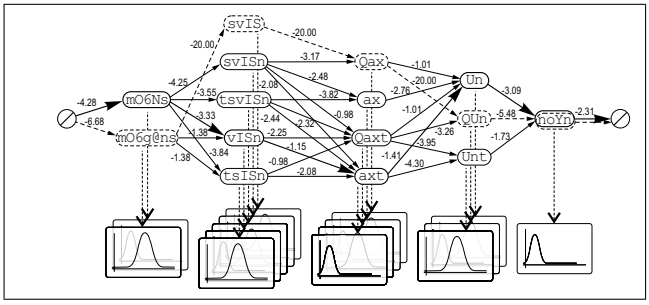

Fig. 1. *Stochastic Markov Graph (SMG) representing a network of pronunciation variants for the German phrase "morgens zwischen acht und neun" (*between eight and nine in the morning*). Nodes represent word pronunciations, edges carry weights obtained from the pronunciation variant sequence model. The bold path denotes the pronunciations selected using the variant sequence model. The example shows the correct consideration 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 only target durations [1].*

3. EXPERIMENTS FOR MEASURING THE LISTENING EFFORT

By measurement of the listening effort it is possible to yield a quality parameter for the rating of naturalness and intelligibility of synthesized speech.

3.1. Design and performance of the listening test

The listening effort of longer speech samples should be measured only because shorter samples (like the one often used in pair comparison tests) do not attract the attention of the listener long enough.

To minimize the influence of system specific features on the listener we used two different synthesis systems: (A) DRESS [4] and (B) MBROLA [9]. However, the word target duration was calculated independently from the synthesis system. In case of system (B) the accent structure was taken from the DRESS-System (A) and (B) was only used to convert the phonemes into sound.

A second reason for using two different systems was the different number of diphones used in the system databases. Normally the diphones are stored according to the canonical pronunciation. By applying variants to the synthesis process, phoneme combinations are observed, which would not appear in the canonical case. Missing a diphone in the synthesis process produces in most of the cases a bad listening impression.

For both synthesis systems (A) and (B) we generated a listening sample with canonical synthesis and another one by applying the pronunciation variant sequence model (PVSM). Since the PhonDatII corpus contains utterances from the domain "travel information" with a limited vocabulary size, mostly short sentences with a special information were combined with a 0.8 seconds pause in between. The final four synthesized speech samples were around 60 seconds long.

The listening test for measuring the listening effort was performed with 37 participants (7 were experienced listeners). In addition to the main parameter listening effort, the three categories: intelligible, natural and colloquial speech should also be rated at an equidistant scale from 0 (less) to 1 (more). Furthermore, the ratings of the speech parameters: sentence melody, speech rhythm, emphasis, speech rate, and pronunciation were asked for. Therefore, a continuous bipolar scale from -5 to 5 was used, whereby the opposite limits are situated at both ends.

3.2. Results and Discussion

Table 1 shows the results (as arithmetic means of the scores) of the listening test. It can be seen that in all categories the variant synthesis algorithm with PVSM yields better (or at least nearly the same) results than the pure canonical synthesis. Especially the listening effort for both systems could be reduced and the overall impression (MOS) could be improved (best at the MBROLA System). Similar to the pair comparison test in [1], the PVSM-synthesis was rated as much more colloquial but only slightly more natural. In the category intelligibility the canonical synthesis is slightly preferred.

Table 1. *Results (arithmetic means) for the listening tests to measure the listening effort using two different synthesis systems (A) and (B).*
In that table "can." stands for the standard synthesis algorithm and "PVSM" stands for the synthesis with variant selection with a pronunciation variant sequence model.

	(A) DRESS		(B) MBROLA	
	can.	PVSM	can.	PVSM
listening effort	2.35	2.32	2.43	2.30
intelligible	0.63	0.56	0.59	0.59
natural	0.39	0.42	0.42	0.44
colloquial	0.34	0.49	0.36	0.48
MOS	3.03	3.05	2.92	3.19
sentence melody	0.86	-0.05	0.51	0.36
speech rhythm	-0.31	0.48	-0.30	-0.04
emphasis	0.92	0.33	1.12	0.80
speech rate	-0.01	0.28	-0.14	0.24
pronunciation	0.53	0.53	0.17	0.16

The values are:

listening effort:	1 *not strenuous* . . . 5 *very strenuous;*
intelligible, natural, colloquial:	0 *less* . . . 1 *more;*
overall impression (MOS):	1 *bad* . . . 5 *very good;*
sentence melody:	-5 *not present* . . . 5 *too intrusive;*
speech rhythm:	-5 *stagnant* . . . 5 *fluent;*
emphasis:	-5 *monotonous* . . . 5 *wrong emphasized;*
speech rate:	-5 *too slow* . . . 5 *too fast;*
pronunciation:	-5 *very indistinct* . . . 5 *very distinct.*

The following conclusions based on the results should be pointed out:

- The results confirm the assumption from [1] that for the special domain "travel information" a canonical realization is more adequate. However, it could not be confirmed that the general rating decreases. In opposite, for these longer speech samples there is a rating in favor of the PVSM-syntheses.
- Similar to the pair comparison tests from [1] the category colloquial speech gets better rating in case of the PVSM-synthesis. The ratings in the categories naturalness and intelligibility are relatively balanced. The latter achieved worse results in the pair comparison test with PVSM-synthesis. The better ratings here could be due to a familiarization effect of the listeners.
- The other speech parameters also show a favorable rating in case of the PVSM-syntheses. For instance, the speech rhythm is more fluent and the speech rate is higher which make synthetic speech not so tedious. Both should have been achieved by introducing variants to synthesis.
- The rating of the speech parameter pronunciation is remarkable, because it yields nearly the same scores for

both synthesis algorithms. It seems that the introduction of variants (and less accurately pronounced, transformed, and deleted phonemes) is not bothering the listeners if the sentence is longer.

- Only the sentence intonation (as a measure of the variation in the fundamental frequency) and the parameter emphasis (as a measure of the stressing of syllables and words) were rated in favor of the canonical synthesis. Applying variants to the synthesis, the sentence intonation is softened and the emphasis is too monotonous. Both are probably due to a wrong prosodic structure in the case of the PVSM-synthesis. Even though the prosody matches the canonical synthesis well, the one-to-one adaptation to the PVSM-synthesis yields a speech sample with an incorrect prosodic structure. Therefore, not only the grapheme to phoneme conversion has to be improved, but also the prosodic generation, which has not been considered up to now.

3.3. Impact factors on the listening effort

In order to measure the impact on the single ratings, a simple correlation analysis was performed. Therefore, the Spearman rank order correlation coefficient between the measured listening effort respectively MOS and the rated categories (intelligibility, naturalness, colloquial speech) as well as the speech parameters (sentence intonation, speech rhythm, emphasis, speech rate, pronunciation) were calculated.

The results of the correlation analysis are shown in Table 2. As it can be seen, the listening effort and the MOS depend strongly on the ratings in the categories intelligibility and naturalness. The dependencies show that a better intelligibility or naturalness lead to a smaller listening effort or a higher MOS. It should be mentioned that for the PVSM-synthesis the rating in the category colloquial speech has also a slight impact. At a more colloquial speech the listening effort decreases and the MOS score increases. Therefore, the listening effort can be seen as a parameter which ranks the overall quality of a speech sample as MOS does.

A strong dependency can also be identified between the listening effort respectively MOS and the speech parameters: pronunciation and speech rhythm, and a slight dependency on the speech rate. A worse pronunciation, a stagnant speech rhythm or a too slow speech rate lead to a higher listening effort or a smaller MOS.

4. CONCLUSION

The results show that the synthesis applying a variant selection with a pronunciation variant sequence model is capable of making synthetic speech sound more "spontaneous" and reduces the listening effort for the longer utterances. The proposed algorithm selects a variant for a given word by considering the variant selection for the surrounding words.

The measured parameter listing effort is suitable for evaluation of the overall performance of a (longer) synthesized utterance.

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

Table 2. *Correlation coefficients for the results of the listening test for measuring the listening effort.*
The abbreviations stand for: can.: standard canonical synthesis and PVSM: Synthesis with variant selection through pronunciation variant sequence model

Parameter:	listening effort		MOS	
	can.	AVFM	can.	AVFM
intelligible	−0.63	−0.80	0.61	0.63
natural	−0.22	−0.42	0.66	0.70
colloquial	0.06	−0.16	0.13	0.29
pronunciation	−0.53	−0.58	0.43	0.45
emphasis	0.02	0.00	−0.28	−0.21
sentence intonation	−0.12	−0.01	−0.13	0.01
speech rhythm	−0.36	−0.62	0.52	0.62
speech rate	0.09	0.12	−0.13	−0.18

5. REFERENCES

[1] S. Werner, M. Eichner, M. Wolff, and R. Hoffmann, "Towards Spontaneous Speech Synthesis - Utilizing Language Model Information in TTS," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 4, pp. 436445, July 2004.

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

[3] A. Bell, M. Gregory, M.L.Brenier, D. Jurafsky, A. Ikeno, and C. Girand, "Which Predictability Measures Affect Content Word Durations?," in *Proc. PMLA*, Estes Park, USA, 2002.

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

[5] F. Bimbot, et al, "Variable length sequence modeling: theoretical foundation and evaluation of multigrams," *IEEE Signal Processing Letters*, , no. 2(6), 1995.

[6] C.M. Westendorf and J. Jelitto, "Learning Pronunciation Dictionary from Speech Data," in *Int. Conference on Spoken Language Processing (ICSLP)*, 1996, pp. 1045–1048.

[7] M. Eichner and M. Wolff, "Data driven generation of pronunciation dictionaries in the German Verbmobil project - discussion of the experimental results," in *Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Istanbul, Turkey, June 2000, pp. 1687–1690.

[8] F. Wolfertstetter and G. Ruske, "Structured Markov Models for Speech Recognition," Detroit, Michigan, USA, 1995, pp. 544–547.

[9] T. Dutoit, V. Pagel, N. Pierret, F. Bataille, and O. van der Vrecken, "The MBROLA project: towards a set of high quality speech synthesizers free of use for non commercial purposes," in *Int. Conference on Spoken Language Processing (ICSLP)*, Oct. 1996, vol. 3, pp. 1393–1396.