Pronunciation Variant Selection for Spontaneous Speech Synthesis — A Summary of Experimental Results

Steffen Werner and Rüdiger Hoffmann

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

Abstract

To make synthesized speech more natural and colloquial the regularity of synthesized speech has to be overcome and spontaneous speech effects have to be integrated into the synthesis process. In a first step towards spontaneous speech we introduced different duration control methods in speech synthesis.

In this paper we summarize the results of previous works (see for instance [1]) of changing the speaking rate indirectly by controlling the grapheme-to-phoneme conversion through different pronunciation variant selection algorithms. The presented results of listening experiments show a significant improvement in the category colloquial impression.

To evaluate the quality of the most outstanding variant selection approach compared to the canonical synthesis (as the state-of-the-art system), we performed a new listening test on longer speech samples. 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.

1. Introduction and Motivation

Synthesized speech still lacks spontaneity and naturalness. Keller describes the state-of-the-art as too regular and states that the variability in phonetic and prosodic parameters has to be increased [2]. But simply varying some parameters would only decrease the intelligibility. That's why, we have to look at human speech production process. To make speech synthesis more natural and colloquial we have to integrate various effects that are observable in spontaneous speech.

Jurafski et al. showed in [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. Conventional speech synthesis systems do not consider the probability of a word in its context.

In a first stage we change the speaking rate directly by shortening or lengthening the syllables of a word depending on the language model probability of that word. Since probable words are not only pronounced faster but also less accurately this approach was extended by selecting appropriate pronunciation variants of a word with different degree of reduction from a variant lexicon according to the language model probability. We called this duration controlled variant selection (DcVS). This second stage changes the local speaking rate indirectly by controlling the grapheme-phoneme conversion.

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, in a third stage we used a pronunciation variant sequence model (PVSM) to select the appropriate variants according to their sequence probability.

A description of the three different algorithms can be found in [1]. In this paper we pay attention to the two variant selection algorithms: (I) duration controlled variant selection (DcVS) and (II) the variant selection according to the pronunciation variant sequence model (PVSM). We summarize the results obtained by performing pair comparison tests (compare Section 2) and present the new results from a listening test, which measures the listening effort (compare Section 3).

For all experiments we used two pronunciation variant dictionaries, one with an average number of 3.7 (3.7–PVDict) and another one with an average number of 2.5 (2.5–PVDict) pronunciation variants per word.

2. Pair Comparison Listening Experiments

2.1. Experimental Setup

Our multilingual, diphone-based, time domain TTS system DRESS [4] was used for evaluation of the proposed algorithms. For intonation control an adaptation of the Fujisaki model [5] was applied, since this model yielded the best results in previous evaluations of the intonation of the DRESS modules [6].

Since the variant lexicon and variant sequence model were trained with the PhonDatII corpus 25 utterances were selected from that database. The test participants (20 for DcVS per PVDict and 35 for PVSM) were asked

to judge each pair of sentences in the three categories: intelligibility, naturalness and colloquial speech.

In order to check the overall impression (MOS) of the synthesized speech samples, an evaluation with absolute category rating (ACR) was conducted. Therefore 15 sentences were randomly chosen from the former experiment in order to reduce the listeners' effort. 50 persons were asked to rate these 15 sentences, which were synthesized with three different algorithms: canonical, DcVS and PVSM.

2.2. Experimental Results

The results of the pair comparison test are summarized in Table 1. The evaluation of both variant selection methods yielded a significant improvement of the synthesis quality in the category of colloquial speech and a slight improvement in the category of naturalness. Nevertheless, the evaluation of the category intelligibility showed clearly that most of the participants voted in favor of the canonical synthesis. That is not surprising, because the overarticulated canonical form can be expected to be more intelligible than any reduced form. Better results could be achieved in all three categories with the PVSM algorithm as compared to the DcVS with the same PV–Dict.

The MOS ratings of the variant selection algorithms are also displayed in Table 1 (canonical synthesis result is shown in brackets). The difference between the considered algorithms is not significant. The canonical synthesis was slightly preferred with an MOS score of 3.21 as opposed to the DcVS (2.85) and the PVSM (2.93).

Table 1: Results (in %) of the listeners' preference in the pair comparison tests

	(I) DcVS		(II) PVSM
PV–Dict:	3.7	2.5	2.5
intelligibility	20.8 %	15.8 %	22.3 %
naturalness	53.7 %	53.6 %	54.4 %
colloquial speech	72.4 %	66.4 %	73.7 %
MOS (canonical: 3.21)	–	2.85	2.93

2.3. Discussion

The results show that the presented approaches are capable of making synthetic speech sound more "spontaneous".

We suggest that some test participants relate the category naturalness to intelligibility and some to 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 sometimes to the category intelligibility.

Because of the different results of the pair comparison and the MOS evaluation we investigated the synthesized samples more deeply:

- Most diphone databases store only diphones according to the canonical pronunciation. That's why a lot of 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 (comp. Table 2). Short sentences 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 an additional parameter to consider when shortening or lengthening a word.

Table 2: Ratings depending on the length of an utterance.

synthesis with 2.5–PV–Dict	DcVS utterances shorter than 2 sec.	DcVS utterances longer than 3 sec.	PVSM utterances shorter than 2 sec.	PVSM utterances longer than 3 sec.
intelligibility	13.7 %	20.6 %	11.4 %	20.6 %
naturalness	50.7 %	52.2 %	46.1 %	57.1 %
colloquial speech	62.0 %	67.8 %	71.4 %	73.0 %
MOS variant sel.	2.80	2.84	2.75	2.89
MOS canonical	3.26	3.09	3.47	3.09

- 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 1, example 1. On the contrary, omitting phonemes in word transitions often results in a higher score (Figure 1, example 2).

Example 1:	ich	will	morgen	abend	nach	fankfurt
canonical:	QIC	vIl	mO6g@n	Qa:b@nt	na:x	fraNkfU6t
DcVS (3.7):	IC	vIl	mo:gN	Qa:b@n	na:	fraNfU6
DcVS (2.5):	IC	vIl	mO6gn	Qa:b@n	na:x	fraNfU6
PVSM (2.5):	QIC	vIl	mO6N	Qa:bm	na:x	fraNfU6t
Example 2:	geht	es	nicht	eher		
canonical:	ge:t	QEs	nICt	Qe:@6		
DcVS (3.7):	ge:d	Es	ICt	Qe:6		
DcVS (2.5):	ge:t	Es	ICt	Qe:6		
PVSM (2.5):	ge:t	s	nIC	Qe:6		

Figure 1: Examples of reduced utterances. The number behind the synthesis methods stands for the used PV–Dict. (Ex.1: I want to go tomorrow evening to Frankfurt*; Ex.2:* Isn't it possible earlier?*).*

- 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.

3. Measuring the listening effort

3.1. Design and performance of the listening test

To minimize the influence of system specific features on the listener by measuring the listening effort we used two different synthesis systems: (A) DRESS [4] and (B) MBROLA [7]. 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).

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

The test for measuring the listening effort was performed with 37 participants (7 were experienced listeners). In addition to the main parameters listening effort and overall performance (MOS), the three categories: intelligibility, naturalness and colloquial speech were 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. Here, we used a continuous (0.1 step size) bipolar scale from -5 to 5, whereby the opposite limits are situated at both ends.

3.2. Results and Discussion

Table 3 shows the results (as arithmetic means) of the listening test. It can be seen that in all categories the variant synthesis algorithm with PVSM yields better results than the pure canonical synthesis (or at least nearly the same). Especially the listening effort for both systems was reduced and the overall impression (MOS) was improved. Similar to the pair comparison test in Section 2, the PVSM-synthesis was rated as much more colloquial but only slightly more natural. In the category intelligibility the canonical synthesis was slightly preferred.

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

- The results confirm that for the special domain "travel information" a canonical realization is more adequate, since intelligibility is very important. However, in opposite to the pair comparison test in Section 2 the general rating does not decrease. For these longer speech samples there is a rating in favor of the PVSM-syntheses.
- Similar to the pair comparison tests from Section 2 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 tediousness of synthetic speech is reduced by a more fluent speech rhythm and a higher speech rate. 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 but also the prosodic generation has to be improved, which has not been considered up to now.

Table 3: Results (arithmetic means) for the listening tests to measure the listening effort using two different synthesis systems (A) and (B).

	(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.*

Table 4: Correlation coefficients between all ratings

		MOS	PR	EM	SM	SR	Ra	Col	Nat	Int
listening effort	LE	−0.58	−0.55	0.03	−0.05	−0.50	0.10	−0.06	−0.34	−0.71
overall	MOS	1.00	0.44	−0.28	−0.07	0.59	−0.14	0.25	0.66	0.60
pronunciation	PR	–	1.00	0.01	0.07	0.43	−0.18	−0.11	0.22	0.63
emphasis	EM	–	–	1.00	0.39	−0.07	0.14	−0.24	−0.28	0.02
sentence melody	SM	–	–	–	1.00	−0.12	0.05	−0.23	−0.18	0.13
speech rhythm	SR	–	–	–	–	1.00	0.16	0.28	0.45	0.45
speech rate	Ra	–	–	–	–	–	1.00	0.06	−0.11	−0.23
colloquial	Col	–	–	–	–	–	–	1.00	0.33	0.01
natural	Nat	–	–	–	–	–	–	–	1.00	0.43
intelligibility	Int	–	–	–	–	–	–	–	–	1.00

3.3. Impact factors on the listening effort

The ratings for the measured categories and speech parameters also influence each other. Therefore we performed a simple correlation analysis and calculated the Spearman rank order correlation coefficient between all rated categories and speech parameters. The results are shown in Table 4.

As it can be seen, the listening effort and the MOS depend strongly on each other. Therefore, the listening effort can be seen as a parameter which ranks the overall quality of a speech sample as MOS does. It is also shown, that 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 lower listening effort or a higher MOS.

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

In the categories naturalness and colloquial speech the ratings depend mainly on the parameters pronunciation and speech rhythm. The category colloquial speech depends slightly on emphasis, sentence melody, and speech rhythm.

The impression from the pair comparison tests that the category naturalness is associated sometimes with the category intelligibility and sometimes with colloquial speech can be confirmed by the correlation coefficients of 0.43 and 0.33 respectively.

The dependencies between the speech parameters are due to the fact, that they have some features in common. E.g. the change in the fundamental frequency is a feature for the sentence melody at sentence level and the emphasis at word level.

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 (longer) synthesized utterances and can be considered as a quality parameter for a speech sample.

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

5. References

[1] Steffen Werner, Matthias Eichner, Matthias Wolff, and Rüdiger Hoffmann, "Towards spontaneous speech synthesis - utilizing language model information in tts," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 4, pp. 436–445, July 2004.

[2] Eric Keller, "A phonetican's view of signal generation for speech synthesis," in *Electronic Speech Signal Processing (ESSP), Tagungsband: Studientexte zur Sprachkommunikation*, Prague, Czech Republic, Sept. 2005, vol. 36.

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

[4] R. Hoffmann, "A multilingual text-to-speech system," *The Phonetician*, vol. 80, no. II, pp. 5–10, 1999.

[5] H. Fujisaki, S. Ohno, and C. Wang, "A command-response model for F0 contour generation in multilingual speech synthesis," in *Proc. ESCA Workshop on Speech Synthesis*, Jenolan, Australia, 1998.

[6] R. Hoffmann, D. Hirschfeld, O. Jokisch, U. Kordon, H. Mixdorff, and D. Mehnert, "Evaluation of a multilingual TTS system with respect to the prosodic quality," in *Proc. Int. Congr. of Phonetic Sciences*, San Francisco, California, USA, 1999, pp. 2307–2310.

[7] 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 *Proc. Intl. Conference on Spoken Language Processing (ICSLP)*, Oct. 1996, vol. 3, pp. 1393–1396.