Modelling Word Stress for Use in Speech Synthesis

Daniel Tihelka, Jindřich Matoušek, Martin Vlach

University of West Bohemia in Pilsen, Department of Cybernetics,
Univerzitní 22, 306 14 Pilsen, Czech Republic
dtihelka@kky.zcu.cz, jmatouse@kky.zcu.cz

ABSTRACT

This paper deals with word stress modelling for use in text-to-speech systems. It describes one of the ways of stress modelling. Several experiments were carried out for different phonemes containing stress and other prosodic features. Listening tests with specially designed words were set up for the comparison of speech generated from each experiment. The results show that the examined way of stress modelling can increase intelligibility and naturalness of the synthetic speech.

1. INTRODUCTION

Stress is the basic feature for word delimitation in continuous spoken speech, emphasising different parts of words. The emphasised part of the word depends on the language; in Czech it is usually the first syllable. Stress is mostly the only feature that can delimit words in spoken speech without other information on the context and meaning of a sentence, e.g. "*topivo*" (material for burning) and "*to pivo*" (this beer). Therefore, in TTS systems stress can rapidly increase the intelligibility and naturalness of the speech generated by them.

Stress consists of all three prosodic features, i.e. *duration* of phonemes (also known as tempo of speech), *intonation* (frequency F_0 of voice base tone) and *intensity* (loudness). Each of these features is emphasised differently in individual languages. Intensity is the most important feature for Czech, followed by intonation and duration [3].

2. STRESS MODELLING

In our previous research, a text-to-speech system ARTIC (ARtificial Talker in Czech) was built. This system is based on unit concatenation in the time domain [2] and was used for experiments with stressed units presented in this paper.

A new speech corpus recorded by a female speaker, comprising several hours of speech, was used for building the speech segment database (SSD) for this system. A fully automatic process based on hidden Markov models (HMM) was used in a segmentation process, and the speech corpus was segmented into *fenemes* (each feneme corresponds to one state of HMM; each HMM is used for the modelling of a triphone, see Figure 1) [2].

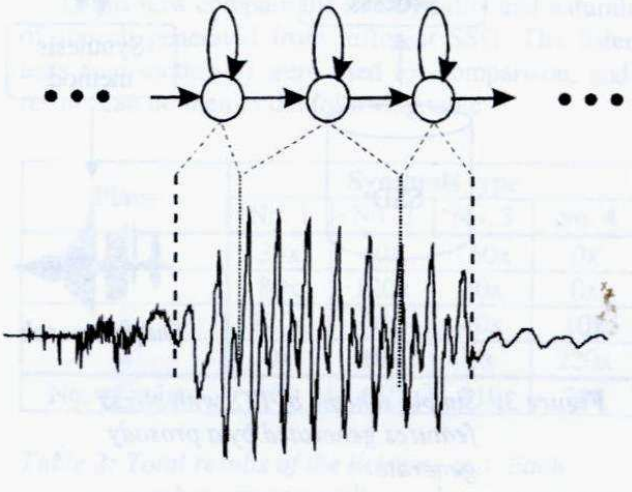

Figure 1: ***The correspondence of three-state HMM, a triphone and fenemes. The dashed lines show the boundary of a triphone, the dotted lines show the boundaries of the fenemes of the vowel "a"***

Representative segments were stored in the speech segment database after the segmentation process. A very simple scheme of this process can bee seen in Figure 2.

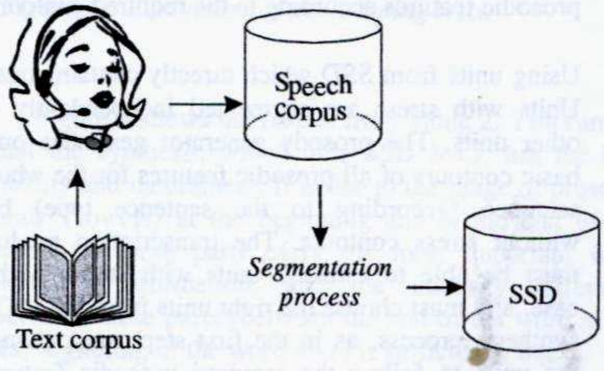

Figure 2: Simply scheme of segmentation process

ARTIC can work with prosodic features. All these are represented as a contour of changes from the base value, and the concatenation method can modify concatenated units according to contour requirements.

There are two basic ways of modelling the stress features in TTS systems:

- Using the contours from prosodic features for stress modelling, especially the contour of intensity and intonation for Czech speech. For a simple scheme of this system see Figure 3. In this case, the prosody generator estimates the basic contours of all prosodic features for the whole sentence (these features depend on the sentence type). The stress features of all stressed phonemes must be added to the basic contours. This means that the generator must be able to estimate phonemes with stress. The transcription module provides transcription to the concatenated units (fenemes in our case) stored in SSD, which are independent of stress. The synthesis module joins units together and changes their prosodic features according to the required contours.

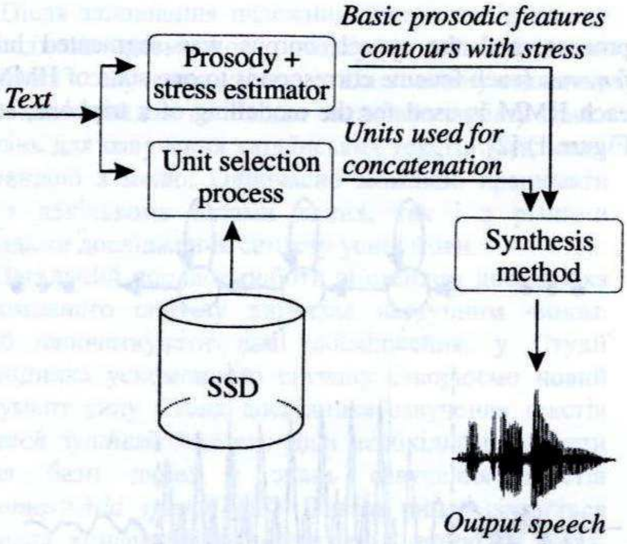
Text
Prosody + stress estimator
Basic prosodic features contours with stress
Unit selection process
Units used for concatenation
Synthesis method
SSD
Output speech

Figure 3: Simple scheme of TTS with stress features generated by a prosody generator.

- Using units from SSD which directly contain stress. Units with stress are segmented independently of other units. The prosody generator generates only basic contours of all prosodic features for the whole sentence (according to the sentence type) but without stress contours. The transcription module must be able to estimate units with stress in this case, and must choose the right units from SSD. The synthesis process, as in the first step, must change the units to follow the required prosodic features contours. A simple system scheme can be seen in Figure 4. In our research we focused on the second approach, as described below.

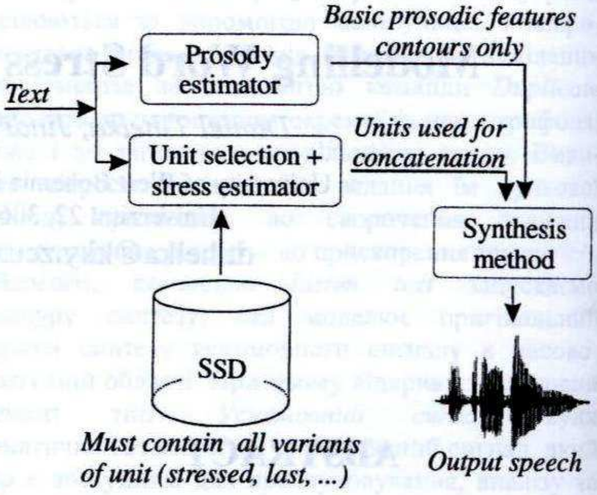
Basic prosodic features contours only
Prosody estimator
Text
Units used for concatenation
Unit selection + stress estimator
Synthesis method
SSD
Must contain all variants of unit (stressed, last, ...)
Output speech

Figure 4: Simple scheme of TTS using units with stress features

3. PHONEMES USED AS PROSODIC UNITS

The statistical approach based on hidden Markov models was used, as mentioned above. The following text describes the positions and types of phonemes chosen for prosody modelling:

1. At first, the vowel in the first syllable of each word with one or more syllables was modelled as *stressed,* which means that this unit contains stress. All other vowels and consonants were modelled as unstressed, because, as we mentioned, stress lies on the first syllable in Czech, and intensity change is usually bigger in a vowel than in a consonant. It can be expected that this simple approach will increase the intelligibility of speech without a large increase in the size of SSD.

2. The vowel in the first syllable (as in the previous case) was modelled as stressed, and the vowel in the last syllable in the words with two and more syllables was modelled as *final*. This means that it lies at the end of the word, and it is possible that the next syllable will be stressed. All the other units were modelled as unstressed. A bigger intensity and intonation contrast between the stressed syllable and the final syllable can be expected, increasing the delimiting feature of stress in Czech. Generated speech could be more intelligible and the size of SSD could still be reasonable.

3. All vowels and consonants in the first syllable were modelled as stressed and all vowels and consonants in the last syllable were modelled as final. All other triphones (i.e. triphones in the middle syllables) were modelled as unstressed. Syllables are modelled

with a greater precision in this approach, as every triphone of the word contains information about its position in the word. We expected that this approach would bring the highest intelligibility and naturalness. Using this approach, the size of SSD will be the largest.

Three-state left-to-right models were used for all described units. These models had the same initialisation of the state mean vectors, covariance matrices and transition matrices before the training process.

3. THE METHOD FOR RESULTS COMPARISON

We set up small listening tests for the comparison of results obtained with different prosodic units. These tests were designed especially for Czech phenomena on words boundaries.

Here are some examples of word pairs which are differentiated from each other by stress positions. The individual words were used in the sentences for the listening tests:

topivo	(fuel)	*to pivo*	(that beer)
tabulka	(a chart)	*ta bulka*	(that roll)
prsten	(a ring)	*prs ten*	(that breast)
jak oběžné	(as circular)	*jako běžné*	(as usual)

Etc.

There were 15 sentences containing one of these words. Each of these 15 sentences was synthesised from:

- SSD without stressed units at first (synthesis output No.1 in the following tables).
- SSD containing units for vowels in the first syllable of the word (synthesis output No.2).
- SSD with units for vowels in the first and in the last syllables of the word (synthesis output No.3).
- SSD with units for all phonemes in the first and in the last syllables of the word (synthesis output No.4).

Naturally, for each synthesis output mentioned, ARTIC had to be able to select the right units for each used SSD.

23 people took part in the listening tests (14 female and 9 male listeners). Every listener heard 10 sentences in all their versions (4 versions per sentence) and had to select the sequence of types of each sentence from the best to the worst. The best type was evaluated by 4 points, the worst by 1 point.

4. RESULTS

Let us compare the sizes of the Speech Segment Database first.

As can bee seen in Table 1, the size of SSD was very similar for syntheses No. 1 and 2, as synthesis No.2 has only 1011 stressed units (triphones, corresponding to vowels at the beginning of the word). The size of SSD for synthesis No.3 is still not very big, as 527 units were added for modelling vowels at the end of the word. On the other hand, the size of SSD used for synthesis No.4 is a little bigger.

SSD used	Number of units in SSD	Size of SSD [MB]
No. 1	9097	17.501
No. 2	10108	20.808
No. 3	10635	31.165
No. 4	11877	47.566

Table.1 SSD sizes for different prosodic units and number of units in the SSD.

Let us now compare the intelligibility and naturalness of speech generated from different SSD. The listening tests (see section 3) were used for comparison, and the results can be seen in the following table.

Place	Synthesis type			
	No. 1	No. 2	No. 3	No. 4
1	30x	40x	160x	0x
2	80x	120x	30x	0x
3	120x	60x	40x	10x
4	0x	10x	0x	220x
No. of points	600	650	810	240

***Table 2:** Total results of the listening test. Each number represents the number of occurrences at a given place.*

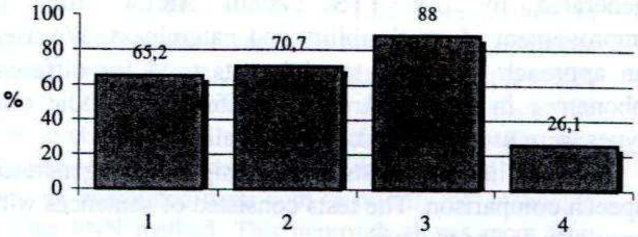

***Figure 5:** Evaluation of the listening tests.*

Figure 5 shows the results from Table 2. You can see that the synthesis, which uses SSD No.3, has the best quality and naturalness. It is due to the usage of prosodic units (vowels) at the beginning and at the end of the word, as these parts carry the most important word prosodic information, and the intensity difference between these parts (between the end of the word *n* and the beginning of the word *n+1*) is sufficiently big.

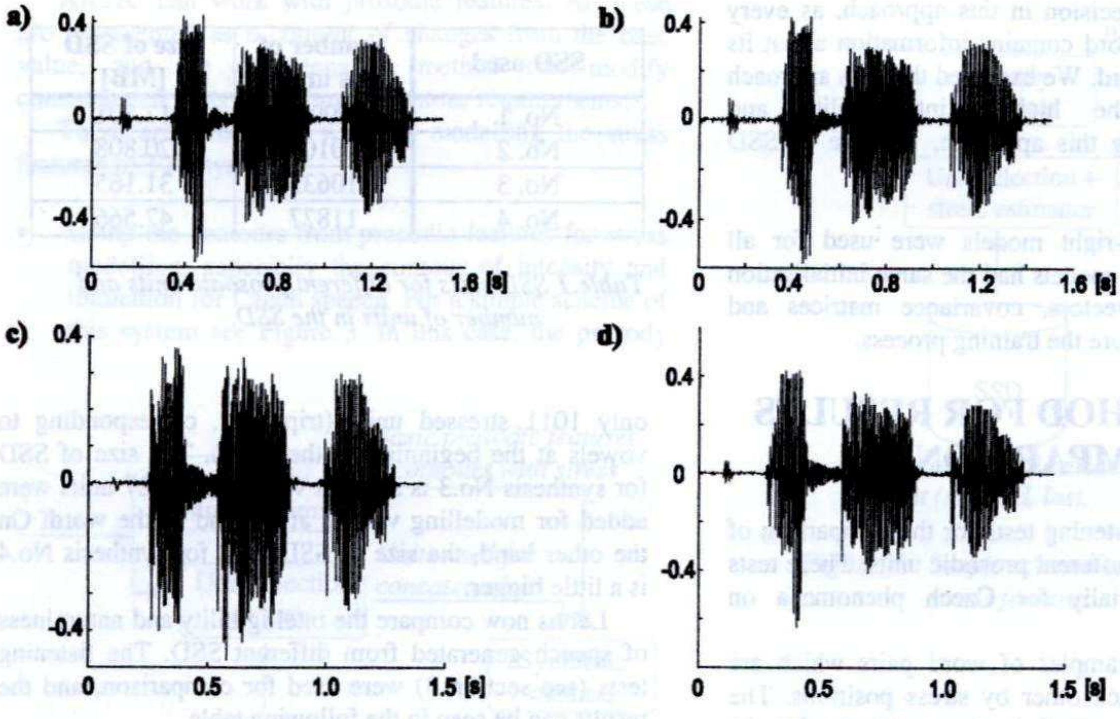

Figure 6: ***Speech signals for the words "tabulka" (a, c) and "ta bulka" (b, d) from SSD without prosodic units (a, b) and with stressed and final vowels as prosodic units (c, d).***

Using SSD No.4 for speech generation, the best results were expected. However, this speech turned out to be the worst. It is due to the small number of units used for training the HMMs, as every unit from SSD No.1 had 3 variants in the SSD No.4.

7. CONCLUSION

We tried to add the stress of the words to the speech generated by our TTS system ARTIC for the improvement of intelligibility and naturalness. We used an approach based on special units used for different phonemes in the word. Three different prosodic unit types were used: stressed, final and all others.

Special listening tests were designed for generated speech comparison. The tests consisted of sentences with specially designed words.

The listening tests showed that this approach brought higher intelligibility and naturalness, especially when vowels in stressed and final syllables of the word were modelled independently of other phonemes.

Stress modelling directly by a prosody generator is planned as our future work.

AKNOWLEDGEMENT

This research has been supported by the Grant Agency of the Czech Republic No. 102/02/P134 and by the Ministry of Education of the Czech Republic No. MSM 235200004.

REFERENCES

1. Matoušek J., Psutka J. and Krůta J.: "*Design of Speech Corpus for Text-to-Speech Synthesis.*" -In: Proceedings of the EUROSPEECH 2001, vol. 3. Aalborg, Denmark, 2001, pp. 2047-2050.

2. Matoušek J. and Psutka J.: "*ARTIC: A New Czech Text-to-Speech System Using Statistical Approach to Speech Segment Database Construction.*" -In: The Proceedings of ICSLP2000, vol. IV. Beijing, China, 2000, pp. 612-615.

3. Palková Z.: "*Fonetika a fonologie češtiny (Phonetics and phonology of Czech).*", Karolinum, Prague 1994.

4. Donovan R.E., Woodland P.C.: "*A Hidden Markov-Model-Based Trainable Speech Synthesiser.*", Computer Speech and Language, 1999.