IMPLICATIONS OF ENERGY DECLINATION FOR SPEECH SYNTHESIS

Jürgen Trouvain, William J. Barry*, Claus Nielsen**, Ove Andersen***

*Institute of Phonetics, University of the Saarland, Germany

**Center for PersonKommunikation, Aalborg University, Denmark

ABSTRACT

This paper examines whether observed phenomena in energy declination can be used to improve the naturalness of synthetic speech. In two production experiments different aspects of intensity fall-off within utterances are analysed including degree of stress, phrase length, phrase boundaries. Energy manipulation was carried out using diphone synthesis as a basis for generating stimuli for perception tests in English and Danish. The results of the listening experiments, in which different versions of a paragraph were ranked for naturalness indicate that amplitude differences can contribute to greater naturalness. However, it is apparent that fine-tuning of amplitude requires good quality synthesis at the more basic prosodic levels.

1. INTRODUCTION

The links in speech production between breath cycle, sub-glottal pressure, acoustic energy, fundamental frequency, and spectral structure [1, 2], and evidence from natural speech indicating their role in the prosodic structuring of phonological phrases [3, 4], suggest that consideration of these parameters should contribute to the naturalness of synthetic speech. The downward trend in both F0 and energy observed within intonational phrases may well be linked to a more general declination tendency which also includes supralaryngeal gestures [5]. Reports of the perceptual significance of F0 and energy declination [6, 7] and of the variation in spectral tilt that accompanies changes in vocal effort [8, 9, 10] further support the assumption that energy differentiation should improve synthetic speech.

2. PRODUCTION EXPERIMENTS

Two production experiments were carried out to ascertain the degree to which energy declination accompanies the declination of F0 during the course of read utterances.

2.1. Production Experiment 1

In a first experiment, two Danish and two English subjects produced two readings each of prosodically comparable, meaningful and nonsense sentences (see Appendix for texts). Each sentence contained five potentially stressed words. Auditory examinations of the recordings showed that these stressable syllables were actually stressed. The vowel in the stressed words was held the same in order to avoid differences due to inherent segment intensity.

Maximum RMS energy was taken from the central five pitch periods to reduce context effects (e.g. from voiceless consonants). Measurements were performed with the standard speech signal analysis software CSL (integration time: 20 ms).

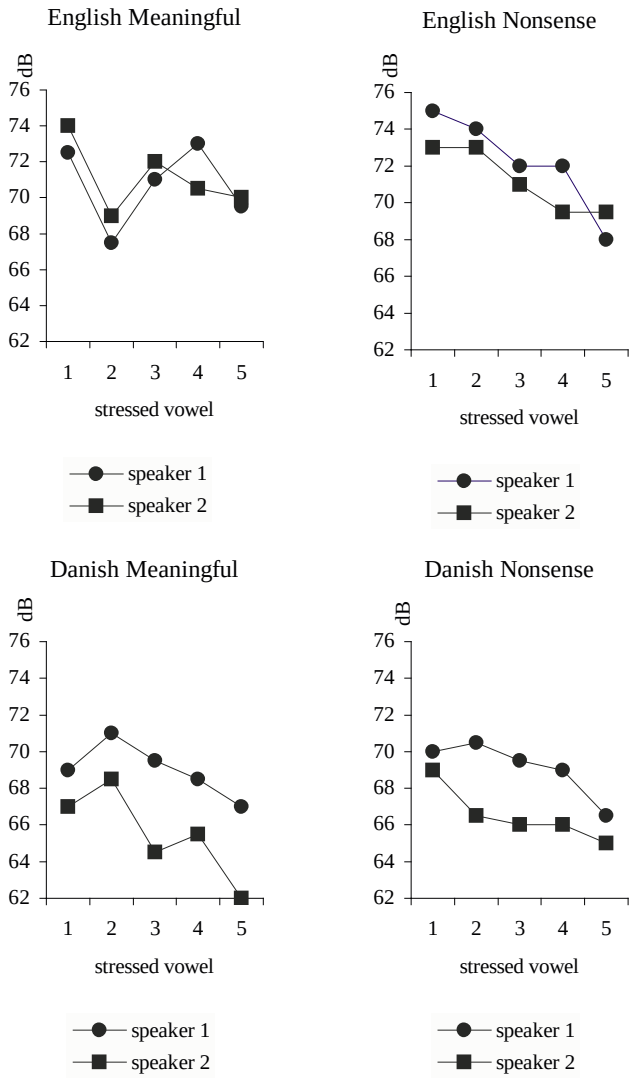


Figure 1: Energy measurements of the stressed syllables in a meaningful and a nonsense sentence condition for the English and Danish speakers (averaged over two readings).

In general, the first stressed syllable was stronger than the average of the following ones. The last stressed syllable was weaker than the immediately preceding ones. This can be seen as a weak confirmation of a general energy downdrift [4]. The energy of the unstressed syllables formed a similar but steeper declination pattern on a lower overall level.

Note that in fig.1 the RMS energy of the second "stressed" syllable spoken by the English speakers (the verb "parked") deviates downwards from the line of the other syllables, probably because the verb has a lower degree of stress. Also the penultimate stressed syllable of speaker 1 shows a higher value than expected, presumably due to the nuclear accentedness.

It was also observed that stress distinction in the vowels of the nonsense sentences is accompanied by differences in the formant intensity measured from the spectrum by hand, at least for the English speakers. A steeper slope in the formant amplitudes was observed for unstressed than for stressed vowels. This agrees with a study for German [10] where spectral slope is one of the cues distinguishing word stress.

For all speakers a final energy drop in the last syllable in the nonsense condition was found. Final lowering of *pitch* on the last syllable is valid for three of the four speakers. So, the final energy drop seems a natural concomitant to final pitch lowering.

2.2. Production Experiment 2

In a second experiment, two native speakers of English twice read a dialogue text. The dialogue was used to simulate a more natural speaking situation with more potential phrasing than isolated sentences can deliver. The test sentences (see Appendix) embedded in the dialogues were controlled for:

- Phrase length
- Phrase structure
- Position of phrase boundary

Here again, the words were selected to keep the stressable vowel constant.

Phrase length

How does the declination line change when the length of the declination phrase varies?

For this purpose the test sentences were lengthened step by step by adding a stressed word, from two in the first sentence (S1) to six in the longer one (S4).

Measurement was the same as in the first experiment.

The results (see fig. 2) do not provide a clear answer. For these speakers all phrases have a similar starting point, independent of the length. There is also a tendency, with some exceptions to reach a lower value in a longer phrase.

Interestingly, as in experiment 1 there are again amplitude peaks linked to the nuclear syllables deviating from the declination lines.

In particular sentence realisations of subject 2, a particularly pronounced rhythmic alternation of strong-weak patterns is found. Due to averaging across realisations this pattern is no longer visible in fig. 2.

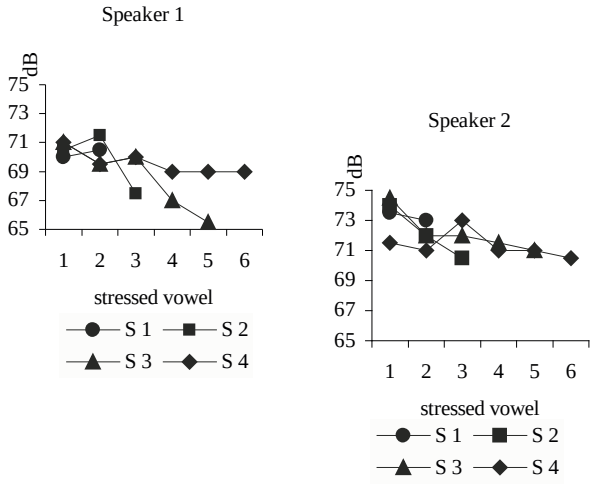


Figure 2: RMS energy of the stressed vowels of declination phrases of different length (S1-4) for both speakers.

Phrase structure

Does syntactic structure affect the form of the declination line?

Two sentences (S4-5) were constructed so that the last two stressed words (out of six) are exchanged. In case of energy declination we would expect that stress no. 5 would have higher values than no. 6, irrespective of the syntax.

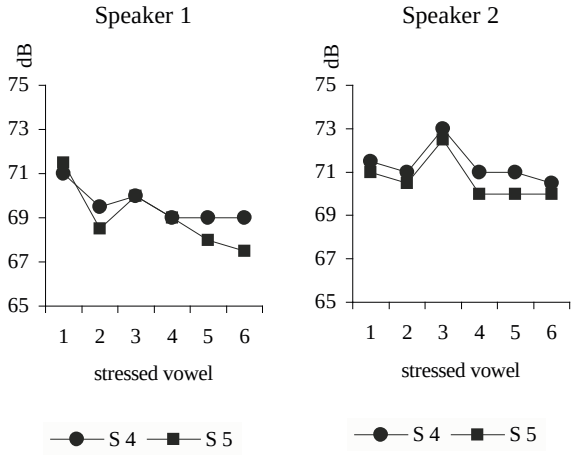


Figure 3: RMS energy of the stressed vowels of declination phrases of same length, but last two stressed vowels are exchanged for both phrases.

Results (presented in fig. 3) show a partial confirmation of our hypothesis. No. 5 is not higher than no. 6 in all cases, but stress no. 6 is never higher than no. 5.

The declination effect for speaker 2 is rather small compared to the first speaker. Again we can observe a nuclear conditioned peak (no. 3, speaker 2), and a de-accented drop (no. 2, speaker 1).

Position of phrase boundary

What is the effect of a phrase break on declination?

At a phrase break, energy declination should result in a rather low energy value for the last stressed syllable in a declination phrase and a rather high value for the first stressed syllable in the next phrase. A phrase boundary would therefore be marked by an energy reset together with features such as a tonal reset, phrase-final lengthening, and pause.

This was examined in S5 and S6, and the results clearly confirm the hypothesis (see fig. 4).

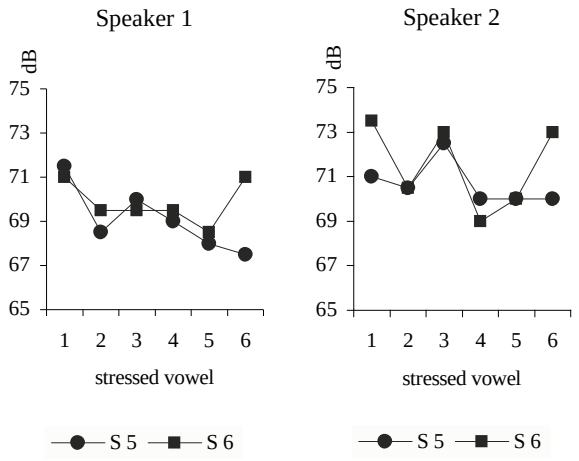


Figure 4: RMS energy of the stressed vowels of declination phrases with different boundary position.

Summarising the second production experiment, the results confirm the previous energy declination findings. They also show a weak correlation between phrase length and energy variation range and a tendency for substructuring including strong-weak alternation of accented syllables. The nuclear accent often deviates upwards from the general declination line of the stronger stress group. Intonation reset is accompanied by a massive energy reset. Finally, there is clear confirmation of the diverging declination line for the unstressed syllables.

3. MODIFICATION OF ENERGY PARAMETERS

A RELP concatenative synthesiser developed in a collaborative project funded by Tele Danmark is used [11]. Originally intended for Danish synthesis, an adaptation to English is also in preparation and was used here. Acoustic parameters to be set for each sound segment are segment duration, F0, and F0 peak alignment. Values for the speech material to be synthesised for the English perception test are based on analyses of recordings with the database speaker (copy synthesis), whereas for the Danish test the rule output of the program was used.

The intensity was manipulated in terms of intensity pre-scaling, uniform energy declination, stress distinctions, and final energy drop, as described below.

3.1. Intensity Pre-scaling

Intensity pre-scaling of the original speech units is performed before run-time. This is done for all sound segments according to their inherent intensity. The need for this step stems from the variation in articulatory effort during the hours of recording the database.

The function of intensity pre-scaling is to obtain a natural intensity relationship between phones on the one hand, and to avoid intensity mismatches at concatenation points between diphones on the other.

Segments to be scaled are divided into four groups:

1. Open and half-open vowels
2. Closed and half-closed vowels
3. Sonorants
4. Obstruents (not scaled)

For details of the prescaling algorithm see [11].

3.2. Uniform Energy Declination

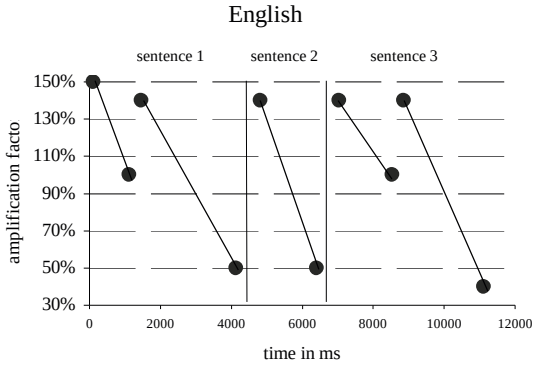
Manipulation for energy declination is performed linearly by interpolating an increase point at the phrase beginning and a decrease point at the phrase end. This results in uniform declination lines where no specific part of a declination phrase is modelled separately. To account for energy differences found in natural speech paragraphs, a simplified model of starting and end points of phrases is proposed. Six levels are assumed. The following rules apply:

1) The first phrase of the first sentence of a paragraph starts with level 1; all other first phrases with level 2;
2) the second phrase of a sentence starts with level 2; the following phrases within a sentence with level 3;
3) the last phrase of the last sentence of a paragraph ends with level 6; all other sentence-final phrases end with level 5;
4) non-sentence-final phrases end with level 4;

The following amplitude factors are applied here for uniform energy declination in our stimuli: level 1 - 150%, level 2 - 140%, level 3 - 120%, level 4 – 100%, level 5 - 50%, level 6 - 40%.

Declination phrase boundaries are set according to the naturally spoken items of the paragraph to be synthesised. Illustrations of stylised declination lines for the three sentence paragraph are shown in fig. 5.

Figure 5: Stylised model of energy declination with six levels.



a) for the English paragraph, b) for the Danish paragraph.

3.3. Stress Distinctions

Relative energy scaling is performed on syllables according to their linguistic importance, not as an overall amplitude modification, but as a change of energy slope. Correlates in analysis are spectral tilt and spectral balance (cf. [10]).

A distinction is only made between stressed and unstressed syllables. Since diphones are recorded in a stressed position only, vowels in unstressed position have to be reduced in energy for the purpose of stress distinction. This process excludes the unstressed portion of vowel-to-vowel sequences as well as Schwa which occurs only in unstressed position and is therefore recorded accordingly.

For de-stressing in terms of spectral re-sloping a filter which lowers intensity by 6 dB from 1 kHz to 8 kHz was applied to the vowels from the baseline signal.

3.4. Final Energy Drop

In addition to the "uniform" declination, a final energy drop is added on the last syllable. According to the observation from natural speech an intensity decrease by 50% is applied to the portion of the last syllable starting at the amplitude maximum up to the last full period. The phrase-final energy dip enhances the effect of reset at the onset of the next phrase.

4. PERCEPTION EXPERIMENTS

Two perception experiments were performed to test the effect of intensity on naturalness. An entire paragraph was selected as speech material since it contains more declination resets than an isolated sentence. Moreover, it represents a more realistic listening condition.

4.1. Stimuli

The procedure for generating the test stimuli is explained with the English example. Different versions of a paragraph were manipulated as follows:

1. natural speech (NSp),
2. NSp with stylised intonation contour,
3. synthetic speech (SSp) without any energy features,
4. SSp with uniform energy declination,
5. SSp with energy declination and accent distinction,
6. SSp with energy declination, accent distinction, and final energy drop.

No. 1 serves as a control stimulus to provide an optimal version against which subjects could judge the less natural versions. This version is recorded with the same voice as for the diphone database used for the synthetic stimuli.

The function of No. 2 is to have a natural counterpart to the synthesised versions sharing the same stylised intonation. F0 resynthesis of natural speech is performed with SFS [12]. No. 2 is thus a copy of No. 1, but with a stylised F0.

No. 3 is the synthetic baseline version. It contains no energy modelling at all and represents what is produced by the current version of the synthesiser. Due to discontinuities at several concatenation points, it was necessary to modify the automatic output to provide an appropriate version for comparison with the natural signals. The analysed segment durations from the natural version (No. 1) and the F0 values used in No. 2 served as input parameters.

No. 4 is a copy of No. 3 with the addition that linear intensity modification across phrases takes place. The six levels specified in 3.2 were applied.

No. 5 introduce a further modification to No. 4. The spectral slope of vowels occurring in a non-stressed position or phrase-final position was modified as described in 3.3.

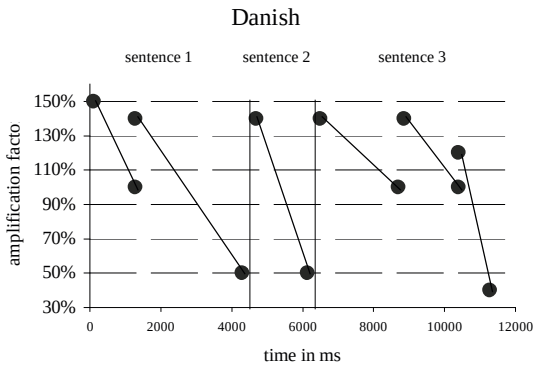
The modification in No. 6 consists of the manipulation of the final energy drop on the last syllable in a phrase (see 3.4), additionally to all other modifications already performed in No. 5.

4.2. The Listening Test

Stimuli were presented as pairs of versions with each pair being presented in both orders (e.g. A-B and B-A). The listeners could listen as often as they wished to the whole utterances and also only to portions of the utterances. The task was to state a preference for one version in each pair.

A consistency examination was performed. If a judgement for a pair didn't match the judgement for that very pair in the reverse order it was noted as inconsistent. An inconsistency rate of 50% (three out of six pairs) or worse resulted in exclusion.

The consistent answers of the remaining judges were recorded, and then ranked for each subject. As a resulting value the average ranking score was calculated.



The hypotheses behind the stimulus generation lead us to expect the following ranking:

1 - 2 - 6 - 5 - 4 – 3, with the synthetic baseline version (no. 3) as worst example and the intensively modified synthetic version (no. 6) as best of the synthetic examples.

4.3. English

From the 18 English native listeners who were tested, five subjects had to be excluded due to inconsistency. The results from the remaining 13 judges can be seen in table 1.

Order	Version	Averaged ranking score
1	no. 1 original	1.23
2	no. 2 stylised	1.77
3	no. 5 spectral slope	4.04
4	no. 6 final drop	4.19
5	no. 4 declin. only	4.54
6	no. 3 baseline	4.92

Table 1: Results of English perception test. Each person's answers lead to a ranking. Rankings of all consistent subjects are averaged.

Our expectations are for the most part confirmed. The natural speech versions are the clear favourites and among the synthetic versions the baseline version without any amplitude manipulation ranks at the end. The fact that the final drop version (No. 6) is not superior to No. 5 can be explained by the selected energy decrease, which may have been too large for the last syllables, introducing a new unnaturalness.

4.4. Danish

Since the English test data showed a massive difference between the natural speech versions and the synthetic ones, putting the synthetic stimuli close together at the "unnatural" end of the scale, the former were removed for the Danish listening test. Thus, only the four synthetic versions for the same paragraph were offered for comparison.

Unlike the English baseline version, which was based on copy synthesis, the rule output of the Danish TTS system served as the Danish baseline version.

Order	Version	Averaged ranking score
1	no. 1 baseline	2.33
	no. 2 declin. only	2.33
3	no. 4 final drop	2.58
4	no. 3 spectral slope	2.75

Table 2: Results of Danish perception test. Each person's answers lead to a ranking. Rankings of all consistent subjects are averaged.

Eleven Danish subjects were tested. Five of the eleven had to be excluded for inconsistency. This rather high figure may be explained by the rather stylised rhythm and melody generated by the prosodic rules which may well have obscured the fine structure tested here.

Among the six consistent subjects, there is no clear picture confirming or rejecting our hypotheses. Three subjects had a clear preference *against* the energy manipulated versions, but two subjects clearly preferred them.

A possible explanation for the different listening preferences might lie in the particular Danish intonation, ending with a high tone on unstressed syllables. These high syllables, and the phrase boundary tones, which are manifested higher in Danish than in English may have a greater impact on intensity than expected.

5. DISCUSSION AND SUMMARY

In two production experiments intensity declination was examined. The described phenomena include level of energy in declination phrases, degree of stress, and a phrase-final energy drop. Manipulation of a synthetic baseline version were carried out according to the observations in natural speech. Cross-linguistic listening experiments for English and Danish indicate that the parameters modified can contribute to an improvement of the naturalness of synthetic speech.

However, improving naturalness clearly demands more than copying the amplitude patterns found in natural speech. Prosody is clearly a highly complex structure and some aspects (rhythm and intonational tune) dominate in the overall impression.

Although the amplitude and spectral properties investigated here are reliably present in natural speech, they cannot improve acceptability independent of the dominant prosodic properties.

We thank Niels-Jørn Dyhr from Tele Danmark for helping us generate the Danish baseline stimulus.

6. REFERENCES

1. Lieberman, P. (1977): *Speech Physiology and Acoustic Phonetics*. New York: Macmillan.

2. Ladefoged, P. (1967): Stress and respiratory activity. In: *Three Areas of Experimental Phonetics*. Oxford: OUP.

3. Fant, G. (1997): The voice source in connected speech. *Speech Comm.* 22, 125-139.

4. Strik, H. & Boves, L. (1995): Downtrend in F0 and Psb. *J. Phonetics* 23, 203-220.

5. Vayra, M. & Fowler, C.A. (1992): Declination of supralaryngeal gestures in spoken Italian. *Phonetica* 49, 48-60.

6. Pierrehumbert, J. (1979): The perception of fundamental frequency declination. *J. Acoust. Soc. Amer*. 66, 363-369.

7. Terken, J. (1991): Fundamental frequency and perceived prominence of accented syllables. *J. Acoust. Soc. Amer*. 89, 1768-1776.

8. Campbell, W. N. (1995): Loudness, spectral tilt, and perceived prominence in dialogues. *Proc. ICPhS 95*, Stockholm, Vol. 3, 676-679.

9. Campbell, N. & Beckman, M. (1997): Stress, prominence and spectral tilt. *Proc. ESCA Workshop on Intonation: Theory, Models and Applications*, Athens. 67-70.

10. Claßen, K., Dogil, G., Jessen, M., Marasek, K. & Wokurek, W. (1998): Stimmqulität und Wortbetonung im Deutschen. *Linguistische Berichte* 174, 202-245.

11. Jensen, J., Nielsen, C., Andersen, O., Hansen E. & Dyhr, N.-J. (1998): A speech synthesizer with modelling of the Danish "stød". *Proc. IEEE Nordic Signal Processing Symposium* (Norsig '98), 121-124.

12. Huckvale, M. (1988): *Speech Filing System*. Phonetics & Linguistics, University College London.

7. APPENDIX

7.1. Production Experiment 1

Sentences recorded in isolation for 1st production experiment. Syllables which might contain any stress are underlined:

Mar(1)tin has parked(2) the car(3) in the garage(4) for his fa(5)ther. (English Meaningful)

Baa(1)baa eats daa(2)daa with laa(3)laa in gaa(4)gaa on a naa(5)naa. (English Nonsense)

Mar(1)tin har ta(2)get sin fars(3) bil til mar(4)kedet med Carl(5). (Danish Meaningful)

Baa(1)baa får daa(2)daa med laa(3)laa i gaa(4)gaa på en naa(5)naa. (Danish Nonsense)

7.2. Production Experiment 2

Sentences from the dialogue text for 2nd production experiment:

S1 So Mar(1)tin parked(2) it.

S2 So Mar(1)tin parked(2) the La(3)da.

S3 So Mar(1)tin parked(2) his fa(3) ther's La(4)da.

S4 So Mar(1)tin Lar(2)kin parked(3) the car(4) in his fa(5)ther's gar(6)den. That was fine.

S5 So Mar(1)tin Lar(2)kin parked(3) the car(4) in the gar(5)den for his fa(6)ther. That was just as good.

S6 So Mar(1)tin Lar(2)kin parked(3) the car(4) in the gar(5)den. For his fa(6)ther that was fine.

7.3. Perception Experiment

Paragraph for perception experiment:

English:

Since last Monday the weather has been unusually bad for the time of year. It's been raining continously. The forecast tells us though we can expect sunshine for the next few days.

Danish:

Siden sidste mandag har vejret været usædvanligt dårligt for denne årstid. Det har regnet konstant. Vejrudsigten fortæeller os imidlertid, at vi kan forvente solskin i de nærmeste dage.