ON SYNTHESIZING DANISH SHORT VOWELS

O. Andersen[1], N.-J. Dyhr[2], C. Nielsen[1]
(authors in alphabetical order)
[1]*Center for PersonKommunikation, Aalborg University, Denmark*
[2]*Tele Danmark Research and Development, Tele Danmark A/S, Denmark*
E-mail [1]{oa, cn}@cpk.auc.dk, [2]njdyh@tdk.dk

ABSTRACT

Danish has a distinctive vowel length opposition which is realized with little differences in vowel qualities. This paper evaluates the concept of synthesizing short vowels from the corresponding long vowels. Different strategies for synthesizing the short vowels from their long counterpart were tested. The first strategy used the original (homogeneous) time-scaling technique built into the signal generation module, the second relied on a new (steady state) time-scaling algorithm. The two time-scaling strategies for synthesizing short vowels from their long counterparts, were compared to using pre-recorded short vowels in a comprehensive listening test. As a reference for the intelligibility score, a set of identical natural stimuli were included. The results showed a great deviation in intelligibility. The steady state time-scaling algorithm reduced the error rate by 50%, but still introduced a significant reduction in intelligibility compared to pre-recorded short vowels.

1. INTRODUCTION

1.1 Background

This paper presents the ongoing work on a concatenative based text-to-speech (TTS) system for Danish [1], which is being developed in cooperation between 'Tele Danmark Research and Development' and 'Center for PersonKommunikation'. The aim of the work is to develop a flexible, natural sounding TTS system with high intelligibility and with as few concatenative speech units as possible. It is considered an important design criterion that the intelligibility and the naturalness are not achieved at the expense of an excessive number of speech units. However, the reduction of speech units should not reduce the overall quality noticeably. The more speech units the more costly it is to introduce new voices, which may impede the widespread use of TTS systems. It is expected that some large companies using TTS systems will require their own unique voice as a kind of 'trademark'.

1.2 System description

The present TTS system comprises three main modules as shown in figure 1. The text is subjected to a morpho-syntatic analysis, which subsequently is augmented with prosodic information i.e. fundamental frequency and duration of phones. Finally the symbolic information is turned into an acoustic signal by the signal generation module.

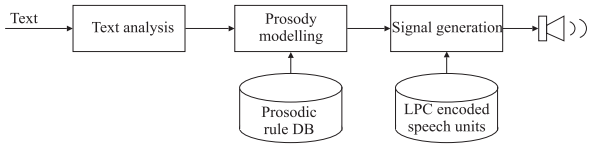

Figure 1. Architecture of the present text-to-speech system

1.3 Speech unit database

The speech unit database employed consisted originally of more than 3800 diphones and triphones extracted from nonsense words spoken by a male speaker. The first step toward reducing the number of speech units focused on the 'stoed', which is a distinctive feature in Danish. It proved possible to implement a signal processing algorithm for simulating the stoed, see [2]. This eliminated the need for pre-recorded 'stoed' units. Consequently, the number of speech units was reduced to 2600. The second step toward minimizing the database concentrated on synthesizing short vowels from their long counterparts. If successful this will enable another significant reduction.

1.4 Long and short vowels

Danish has distinctive vowel length, which means it is possible to make a distinction between different words solely by the duration of the vowel, e.g. *kile* [kˈiːlə] (wedge), *kilde* [kˈilə] (source) and *parken* [pˈɑːgən] (the park), *pakken* [pˈɑgən] (the package). Perceptually most long/short vowel pairs are identical in terms of vowel quality. However, there may be minor acoustical differences especially in fast speech, where the short vowels is slightly centralized compared to their long counterparts [1]. The degree of centralization is to a certain extent dialect and speaker dependent. When the speech unit database was designed in 1994 an informal acoustic analysis was performed on long vs. short vowels in natural words spoken by two speakers. The results at that time indicated that it would not be advisable to exclude the short vowels from the speech unit database. In 1998 substantial testing of both intelligibility and naturalness, see [1], proved it impossible to synthesize short vowels from their long counterparts using the homogeneous time-scaling algorithm, whereas there were no significant differences between pre-recorded short vowels and long vowels manually segmented as short vowels. As the difference between natural long and short vowels are typically as shown in figure 2,

it has been decided to develop a time-scaling algorithm, which reduces duration by deleting pitch periods in the steady state.

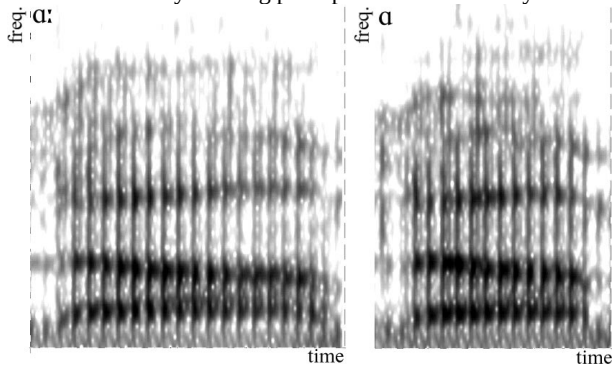

Figure 2. Wideband spectrograms of [ɑː] and [ɑ]. Extracted from the Danish words *larme* [lˈɑːmə] and *lamme* [lˈɑmə] in natural speech.

The purpose of this paper is to investigate whether this steady state time-scaling algorithm, will make it possible to use the long vowels to synthesize the corresponding short vowels.

1.5 Signal generation

The signal generation module in the TTS system is based on pitch-synchronous residual exited linear prediction (LP) analysis. As part of the offline setup of the TTS system an intensity pre-scaling of the speech unit database is performed. The purpose of the pre-scaling is to obtain a natural intensity interrelationships between phones and to avoid intensity mismatches at concatenation points between diphones. The pre-scaling methodology is described in details in [2]. After intensity pre-scaling, each speech unit is LPC-analyzed and stored in an encoded database. At run-time the fundamental frequency manipulations are carried out on the residual signal. The fundamental frequency is modified by reducing or increasing the length of each pitch period. The duration modifications are achieved by repeating or deleting pitch periods. Two types of time-scaling algorithms for reduction of duration have been tested.

A) Homogeneous deletion of pitch periods, (fig. 3A)

B) Deletion of pitch periods in steady state only. (fig. 3B)

Steady states are located by analysis of changes in MEL-frequency cepstral distance measures [3], (fig. 3B). If the desired reduction is larger than the duration of the steady state, the homogeneous time-scaling is applied.

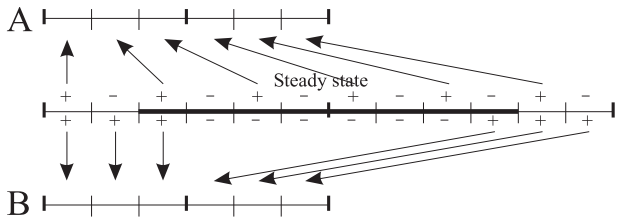

Figure 3. The two methods for deletion of pitch periods at time reduction: A) Homogeneous reduction. B) Steady state reduction. The center line represents the original phone. The thick horizontal line represents steady state.

Finally, the residual signal is filtered with the pitch-synchronously updated LPC-synthesis filters.

2 METHODS

2.1 Test material

The results from an acoustical analysis of long and short vowels in the speech unit database, see [1], revealed that the four most likely candidates for synthesizing short vowels from their long counterparts were [iː eː ɛː ɑː]. 16 minimal pairs containing the following long/short vowel pairs were tested: [iː i, eː e, ɛː ɛ, ɑː ɑ] The vowels were presented in different synthesized versions of genuine Danish words. The vowels were all in stressed position. The test words were phonetically balanced. The test words were imbedded in a carrier sentence: *det er ___ han siger* [de aˀʌ ___ han sˈiːʌ]. As a reference for the intelligibility score an identical material, recorded by the same speaker, who originally recorded the diphone database, was tested as well. The stimuli were generated in the following ways:

I. Natural long vowels.

II. Natural short vowels.

III. Pre-recorded short vowels, homogeneously timed-scaled.

IV. Pre-recorded short vowels, steady state timed-scaled.

V. Long vowels, homogeneously timed-scaled as short vowels.

VI. Long vowels, steady state timed-scaled as short vowels.

VII. Long vowels homogeneously timed-scaled as long vowels.

VIII. Long vowels, steady state timed-scaled as long vowels.

The total material consisted of 128 sentences. In order to avoid any influence from inadequate duration rules, all instances of duration in the synthesized versions were copied from those in the natural material.

2.1 The test

The test took place in acoustically damped environments. The stimuli were synthesized at 16 kHz and made available at 48 kHz on a DAT recorder and presented through headphones. A total of 41 listeners (27 males and 14 females) volunteered for the test. The average age was 38 years. None of the listeners had any known hearing disabilities. Prior to the start of the test the listeners were given a written instruction describing the test procedure.

The test started with 16 dummy sentences to accustom the listeners to the test situation and the synthetic speech. The dummy stimuli were of the same type as the real stimuli. Then the 128 sentences were presented to the listeners. To avoid any positional effect the sentences were presented in reverse order

to 46% of the listeners. The listeners were asked to write the test word. The answer had to be given within 6 seconds. The test lasted approximately 30 minutes and was divided into two sessions with a 2 minutes break between sessions.

3 RESULTS

The error rate among the listeners showed a great deviation, as it varied from 5.8% up to 36.2%. The average error rates for each signal type are shown in table 1.

Type	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**
Error	2.0	2.1	11.0	10.4	51.7	26.0	4.2	3.4

Table 1. Average error rates for each signal type.

To determine whether the differences in error rates in table 1 are statistically significant a Pearson's Chi-square test with Yates' continuity correction [4] was performed. Table 2 shows the Chi-square values. The null hypothesis H_0: TypeX=TypeY can be rejected if the Chi-square value is larger than 6.63 (1% significance level).

Y / X	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**
I		0	42.3	38.4	410	154	5.53	1.88
II	0		40.4	36.5	406	152	4.71	1.39
III	42.3	40.4		0.07	251	47.7	18.9	27.5
IV	38.4	36.5	0.07		260	52.4	16.1	24.2
V	410	406	251	260		90.6	361	382
VI	154	152	47.7	52.4	90.6		116	132
VII	5.53	4.71	18.9	16.1	361	116		0.73
VIII	1.88	1.39	27.5	24.2	382	132	0.73	

Table 2. Results from the statistical analysis. Gray boxes indicate combinations where the H_0 cannot be rejected.

Among all the presented methods for generating synthetic long vowels (VII, figure 4B and VIII, figure 4C) there was no significant difference between the synthetic and the natural reference stimuli (I, figure 4A). It was also found that none of the methods for synthesizing short vowels (III, figure 5B, IV, figure 5C, V, figure 5D and VI, figure 5E) had as few errors as the natural reference stimuli (II, figure 5A). The test proved that the steady state time-scaling algorithm (VI, figure 5E) is significantly better than the homogeneous algorithm (V, figure 5D) when the short vowels are synthesized using long vowels. Neither of the time-scaling algorithms (V, figure 5D and VI, figure 5E) equals the results of short vowels generated from pre-recorded short vowels (III, figure 5B and IV, figure 5C). The test showed no significant difference in intelligibility between the two time-scaling algorithms when the short vowels were generated from pre-recorded short vowels (III, figure 5B and IV, figure 5C).

4 DISCUSSION

The short vowels generated from their long counterparts by the homogeneous time-scaling algorithm (V, figure 5D) had the poorest score in the test, this is in full agreement with the results in [1]. The homogeneous time-scaling does not reduce the steady state as much as desired, this may result in an over-articulated and consequently unnatural sounding short vowel. As the duration in all the synthesized test sentences is copied from the natural material, the result can not be explained by an inappropriate durational relationship between long and short vowels. However, a missing centralization in combination with an over-long steady state for a short vowel, may explain why the synthesized vowels in V are often confused with the corresponding long vowels.

Although the steady state time-scaling algorithm (VI, figure 5E) reduces the error rate by 50% compared to the homogeneous time-scaling (V, figure 5D), it still introduces a significant reduction in intelligibility compared to the pre-recorded short vowels (III, figure 5B and IV, figure 5C). As the duration of steady states of the vowels in VI are almost identical to that of the vowels in IV, a possible explanation may be that the slight centralization normally found in natural short vowels is an important cue to the perception of Danish short vowels.

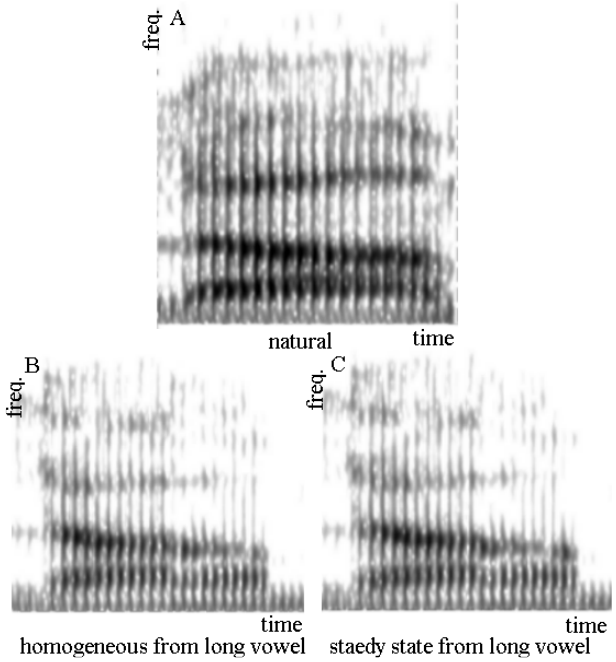

Figure 4. Wideband spectrogram of [ɑː]. Extracted from the Danish word *larme* [lˈɑːmə].

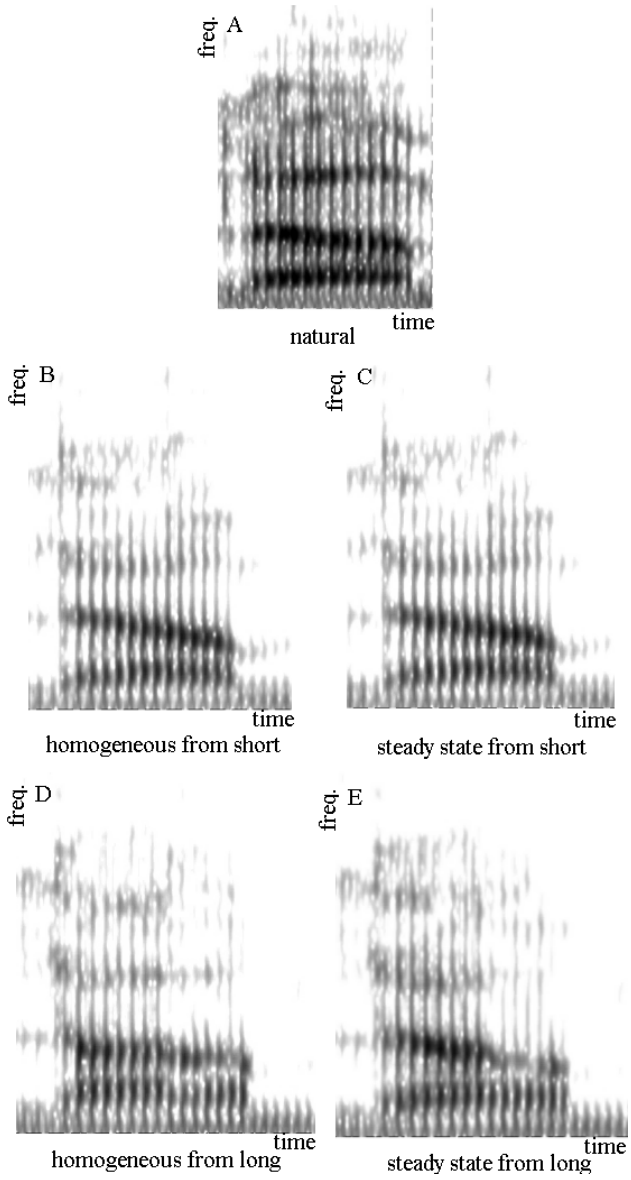

Figure 5 Wideband spectrogram of [ɑ]. Extracted from the Danish word *lamme* [lˈɑmə].

5 CONCLUSION

The results of the present investigation show that it will not be possible to synthesize the short Danish vowels from their long counterparts, without introducing a significant reduction in the intelligibility of the short vowels. It is concluded, that the loss of intelligibility is too costly, and it is therefore not considered advisable to exclude pre-recorded short vowels in the present and future Danish speech unit databases.

6 SAMPLE FILES

The following files are included on the CD-ROM.

File name	Description
0257_01.wav	Type I sample.
0257_02.wav	Type II sample.
0257_03.wav	Type III sample.
0257_04.wav	Type IV sample.
0257_05.wav	Type V sample.
0257_06.wav	Type VI sample.
0257_07.wav	Type VII sample.
0257_08.wav	Type VIII sample.

Table 3. The included sound files.

An online demonstration of the synthesizer is available at: http://www.cpk.auc.dk/syntese

7 REFERENCES

[1] O. Andersen, N.-J. Dyhr, I. S. Engberg, C. Nielsen, 'Synthesizing Short Vowels from their Long Counterparts in a Concatenative Based Text-to-Speech System' In proceedings of The Third ESCA/COCOSDA Workshop on Speech Synthesis. 1998.

[2] Jesper Jensen, Claus Nielsen, Ove Andersen, Egon Hansen, Niels-Jørn Dyhr, 'A Speech Synthesizer with Modelling of the Danish Stoed', In proceedings of 3'rd IEEE Nordic Signal Processing Symposium, June 8-11, 1998.

[3] Lawrence Rabiner, Biing-Hwang Juang, 'Fundamentals Of Speech Recognition', Prentice Hall Signal Prossing Series, 1993.

[4] Martin Bland, 'An introduction to Medical Statistics', Oxford University Press, 1996.