Synthetic regional Danish

Bodil Kyst and Peter Juel Henrichsen
Center for Computational Modelling of Language
Copenhagen Business School
bodilkyst@yahoo.com pjuel@id.cbs.dk

Abstract

The speech technological scene in Denmark is slowly gaining public and commercial attention, and today there is a call for more diverse products than just recognition and synthesis of standard Danish (i.e. the Copenhagen regiolect). In this working paper, we demonstrate how an existing synthesis engine may be tuned to provide a regiolect speaking voice, involving - in its simplest form - prosodic alterations only. We present the first publicly accessible Danish synthetic voice for a non-standard regiolect - in casu Århusian.

1 Introduction

So far synthetic speech in Denmark has been based on the standard variety of Danish spoken in the area around Copenhagen, normally referred to as standard Danish. The purpose of our current project "Synthetic regional Danish" is to develop a Danish artificial voice for the variety of Danish spoken in Århus (the second biggest city).

Many of the local dialects in Denmark have died out. Danish spoken in Århus today (hereafter "Århusian") cannot be described as a genuine dialect, since it hardly differs from Copenhagen standard Danish in terms of vocabulary and grammar, or the sound of the individual phones. Århusian is rather a regional variety Ű a regiolect. A major difference between Århusian regiolect and Copenhagen standard Danish concerns the tone pattern of the stress group. Hence synthetic Århusian must at least diverge from existing applications in the prosodic assignment function (i.e. the function assigning fundamental frequency values to the phonetic segments). In addition, lexical, phonetic, and further prosodic alterations must be considered (e.g. vowel prolongation, schwa-reduction).

Our project has two main objectives: Firstly we want to establish the fact that existing Danish synthesis applications can be modified with relatively little effort to cover Danish regional variants. This fact may be of linguistic as well as commercial interest concerning the localization of speech technology. Secondly, we want to create an experimental environment for testing our theoretical understanding of the distinctive features of the regional variants of spoken Denmark.

In addition to this, there are further perspectives in creating a regiolect-speaking synthetic voice: Firstly it might bring the synthetic voice closer to the user by leading the voice to be associated with local proximity. Secondly it might add to a greater tolerance towards non-standard varieties of Danish.

A version of our Århusian synthetic voice can be tested at: **http://www.id.cbs.dk/~pjuel/TtT_region**

2 Phonetic basics

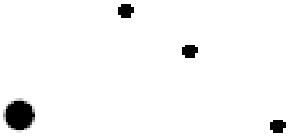
2.1 The Copenhagen stress group tone pattern

The tone pattern of the Copenhagen stress group is described by Nina Grønnum (Grønnum, 1992). The first syllable of the stress group (the tonic syllable) is pronounced in a low tone. The following unstressed syllable (the post-tonic syllable) is pronounced in a high tone and the following unstressed syllables within the stress group are also pronounced in high tones, yet gradually losing altitude according to the overall downdrift of the tones in the stress group. This tone-pattern of the stress group can be described as low-high

S. Werner (ed.), *Proceedings of the 15th NODALIDA conference, Joensuu 2005*, Ling@JoY 1, 2006, pp. 116–123
ISBN 952-458-771-8, ISSN 1796-1114. http://ling.joensuu.fi/lingjoy/

since the tone relation between the first two syllables is low-high. See fig. 1

Fig.1



The big dot is the stressed syllable and the small ones are the unstressed syllables within the stress group.

2.2 The Århusian stress group tone pattern

The tone pattern of the first two syllables of the Århusian stress group is, roughly speaking, the opposite of the Copenhagen tone pattern: In most stress groups the relationship between the tonic syllable and the post-tonic syllable can be described as high-low. Still a considerable part of the stress groups have a low-high tone pattern. As described in Kyst (Kyst, 2004) the tone pattern depends on the segmental composition of the tonic and the post-tonic syllables. Two factors play a determining role: A: Stød and B: The occurrence of non-sonorant elements between the tonic and posttonic vowels:

2.2.1 A. Stød

Stød is a phonation type articulated by a glottal constriction resulting in a sound resembling creaky voice. It is a distinctive feature and its presence/non-presence results in minimal pairs like finner $[f2en!C]$ (noun) (the / is the stød) with stød meaning Finns from Finland vs. finner $[f2enC]$ (noun) without stød meaning finns on a fish (for an introduction to our phone alphabet, see footnote 2).

In the Århusian stress group stød in the tonic syllable affects the F0 (fundamental frequency = tone height) of the following syllable in a consistent manner.

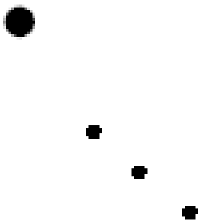
Group 1:
Stress groups with stød (on the tonic syllable) always have a high-low tone pattern:

Tonic syllable: high
Posttonic syllable: low
The following unstressed syllables: low, gradually getting lower

Examples of stress groups of this kind could be:
finner $[f2en!C]$ (noun) meaning Finns from Finland
bruser $[br2u\ :!sC]$ (verb) meaning rushes.

See fig. 2:

Fig.2



2.2.2 B. The occurrence of non-sonorant elements between the tonic and posttonic vowels

Stress groups without stød (on the tonic syllable) sometimes have a high-low tone pattern and sometimes a low-high tone pattern. This depends on the second factor determining the tone pattern of the stress group in Århus, namely the occurrence of non-sonorant elements between the tonic and posttonic vowels:

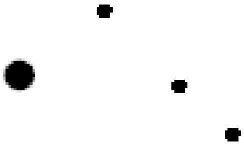
Group 2a:
If only sonorant elements occur between the vowel of the tonic syllable and the vowel of the posttonic syllable, the resulting tone pattern is low-high:

Tonic syllable: low
Posttonic syllable: high
The following unstressed syllables: low, gradually getting lower

An example of this is
finner $[f2enC]$ (noun) meaning finns on a fish

See fig. 3:

Fig.3



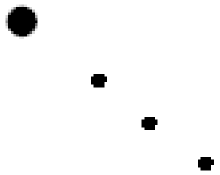
Group 2b:
If at least one non-sonorant element occurs, the resulting tone pattern is high-low:

Tonic syllable: high
Posttonic syllable: low
The following unstressed syllables: low, gradually getting lower (The drop in tone between the tonic and posttonic syllable is not as steep as in the case of syllables with stød (group 1 above). We did incorporate this in our synthetic voice).

An example of this kind of stress group is $[br2u : sC]$ (noun) meaning shower

See fig. 4:

Fig.4



2.3 Differences between natural and synthetic speech

It should be kept in mind that the model of tone patterns outlined above is simplified compared to natural speech in at least two respects:
Firstly in natural speech, the tones of the syllables are not constant in height like steps on a staircase, rather like curves moving up and down. Secondly the low tone in syllables without stød and only sonorant elements between the two vowels (group 2a) is not as low as the low tones of the other types of stress groups. Fig. 5 and 6 below show the tone patterns from recordings of natural speech.[1] These are the patterns that are imitated by the synthetic voice in a simplified way with level tones (fig. 1-4 above).

3 Danish speech synthesis

For reference synthesis application, we selected the so-called TtT Workbench (Tekst-til-Tale, Text-to-Speech), an experimental speech synthesizer for Danish which has been developed at the Center for Computational Modelling (among other places). The TtT system is a largely traditional design based on a pipeline of modules:

3.1 The synthesis application for standard-Danish

1. Text Preparation (expanding abbreviations, converting digits to their written equivalent, etc.) not shown in fig.70

2. Text Analysis (PSG-style grammar rules enriched with prosodic markers for main stress reduction, stød elimination, etc.)

3. Prosody Assignment (phone-based annotation rules: F0, timing, vowel prolongation, schwa reduction, onset of stød depression, etc.)

4. Signal Processing (converting the quantified phone string into sound files using diphone re-synthesis).

Figure 7 below shows the functional parts of the speech engine.
The web-based interface allows the user to control the speech engine using a standard browser (e.g. Netscape, Mozilla, or MS-Explorer). The advanced user can insert and upload a distribution of lexical and grammatical resources, thus defining his own language model.[2] The workbench, however, also allows the user to simply enter ready-made strings of phones (cf. note 1) such as:

User input:	,ensdit2ud,fC,d2z:taleNvisdig, ("Institut for Datalingvistik", Dept. of Computational Linguistics)
User input:	,f2en!Cn0,f2en!C,f2enCn0,mE,f2eNCn0, ("finnerne finder finnerne med fingrene", the Finns find the finns with their fingers)

Words may be comma separated. The commas have no influence on the acoustic rendering,

[1] Recordings from Kyst (Kyst, 2004) Ű stress groups cut out of longer sentences read loud by native speakers.

[2] The TtT Workbench phonetic inventory is based on the Danish SAMPA (**www.phon.ucl.ac.uk/home/sampa**), a many-to-one mapping of the IPA (International Phonetic Alphabet) on the Danish sound inventory. Since certain SAMPA symbols are inconvenient for use with regular expressions (as in the TtT server scripts) and for transfer over the Internet (used by the TtT web-interface), we use an alphanumeric SAMPA mapping. Our phone table may be consulted at **www.id.cbs.dk/∼pjuel/TtT**

but they tend to make phonetic strings more readable.

TtT phone table summary (cf. Henrichsen 2001b)

2 is "tryk" (main stress)

! is "stød" (a quick glottal contraction)

: is vowel prolongation

z is the full vowel in e.g. "tabe" (lose)

C is the vowel occurring twice in "kopper" (cups)

0 is schwa as in e.g. "tabe" (lose)

TtT phonetics does not include secondary stress

By pressing the button phon2wav, the client transmits his input to the server, which in turn returns the sound file (in .wav format) produced from the phone string. Most browsers will then allow the user to just click on the link on the answer page in order to listen to the sound file. For pedagogic reasons, the TtT server application adopts a rather conservative style of feedback rejecting, (with a comment), any irregular phone string. Examples of phone strings which are rejected:

- Strings beginning with a semivowel (e.g. R, J, or w) in conflict with the Danish phonotactics.
- Strings with zero instances of symbol 2 (main stress); any utterance must contain a stressed syllable in order to be pronounceable.
- Strings with illegal stød. Only two stød loci are permitted, viz. immediately after a long vowel (as in "ben", [b2e:!n]), and immediately after a short vowel + voiced consonant (as in "bind" [b2en!]).

The TtT workbench is found at **www.id.cbs.dk/∼pjuel/TtT**. Access is free of charge, but advanced users need a password issued by the author. Further details on the access and options of the publicly available Danish synthesis engine are found in Henrichsen (Henrichsen, 2005). (Henrichsen, 2005) contains reprints of a number of early research papers on Danish synthesis.

3.2 The Århusian voice

The current project has involved modifications in especially module 3 and - to a lesser extent - module 2 cf. fig. 7. The single most significant research task has been the quantification of the phonological patterns discussed in sect. II.

For the standard-Danish voice, F0 steps (i.e. fundamental frequency variation for the main vowels in two adjacent syllables) come in two flavors - an "up-step" and a "down-step" - computed as:

$$\begin{aligned}\text{upstep} &= \text{currentF0} + (\text{maxF0} - \text{currentF0})*\text{deltaF0}\\ \text{downstep} &= \text{currentF0} - (\text{currentF0} - \text{minF0})*\text{deltaF0}\end{aligned}$$

where maxF0 (minF0) is the global F0 maximum (minimum) and deltaF0 is the relative step size (set to 0.3 for standard-Danish).

For the Århusian voice, a slightly more elaborate model is engaged, defining "up-step", "small-down-step" and "large_down_step", in accordance with our description of Århusian prosody (see 3.2.b).

$$\begin{aligned}\text{upstep} &= \text{currentF0} + (\text{maxF0} - \text{currentF0})*\text{deltaF0}*0.8\\ \text{smalldownstep} &= \text{currentF0} - (\text{currentF0} - \text{minF0})*\text{deltaF0}*0.5\\ \text{largedownstep} &= \text{currentF0} - (\text{currentF0} - \text{minF0})*\text{deltaF0}*1.0\end{aligned}$$

As seen, large-down-step for Århusian is equal to the general F0-step for standard-Danish, while small-down-step is only half that size, and up-step is somewhere in between, in order to provide a balanced prosodic curve, assuming that small and large downsteps are fairly evenly distributed.

The various F0-steps are assigned using an array of boolean flags, marking each syllable as 'un-voiced', 'with-stød', etc.

4 Project Status

We have concluded a pilot test involving six people with knowledge of Danish dialectology. The subjects were asked to determine the local origin of the voice and whether it was natural or synthetic.[3] Based on these and other results we are currently preparing a test environment

[3] Each test person heard four sound clips, some heard one of each kind and others heard the same clip twice. Each of the four sound clips were played six times in total.

for larger-scale listening tests.

The sentence played to the test persons was:

Den store viser er den der viser minuttallet
$[dEn, sd2o : C, v2i : sC, a, d2En!, dA, v2i : !sC, min2udtal!0D]$
- meaning: the big hand is the one that indicates the minutes

The result of the pilot test can be read from the confusion matrix displayed in fig. 8 below.

All test persons were able to distinguish the natural voices from the synthetic voice, and all natural voices were identified correctly according to their geographic location. The geographic origin of the synthetic Århusian voice was determined correctly in four out of six cases, and the geographic origin of the Copenhagen voice was determined correctly in five out of six cases.

5 Conclusions

Although our experiments are still only in a preliminary stage, we have reasons to believe that synthetic Århusian (or perhaps any Danish regiolect) in within easy reach, given the existing sophisticated Danish applications for speech synthesis proper, and the long record of dialectological research in Denmark. We also note that our experiments - including the small user test presented in section IV - have stirred a considerable public interest, especially among non-Copenhageners. In the next phase of our project we therefore intend to apply for research grants arguing that (i) synthetic regiolects can be developed as minor extensions to existing applications, (ii) west-Danes will certainly give a non-standard Danish synthetic voice a warm welcome.

References

Nina Grønnum. 1992. *The Groundworks of Danish Intonation. An Introduction.*, volume 1. Museum Tusculanum Press, Copenhagen, Denmark.

Peter Juel Henrichsen. 2004. *The Twisted Tongue, Tools for Teaching Danish Pronunciation Using a Synthetic Voice; in Copenhagen Studies in Language 30/2004 (ed. PJH)*, volume 1. Samfundslitteratur, Copenhagen, Denmark.

Peter Juel Henrichsen. 2005. *Tekster til Taleteknologi - med saerlig vaegt paa syntese af dansk talesprog (antology of texts on Danish linguistic research in speech synthesis; several texts are in Danish)*, volume 1. Copenhagen Business School Press, ISBN 87-6340-549-0, Copenhagen, Denmark.

Bodil Kyst. 2004. *Trykgruppens toner i Århusiansk regionalsprog*. Unpublished masterthesis, see www.bodilkyst.dk, Århus University, Denmark.

Fig.5 Group 2a vs. group 1, Århus vs Copenhagen

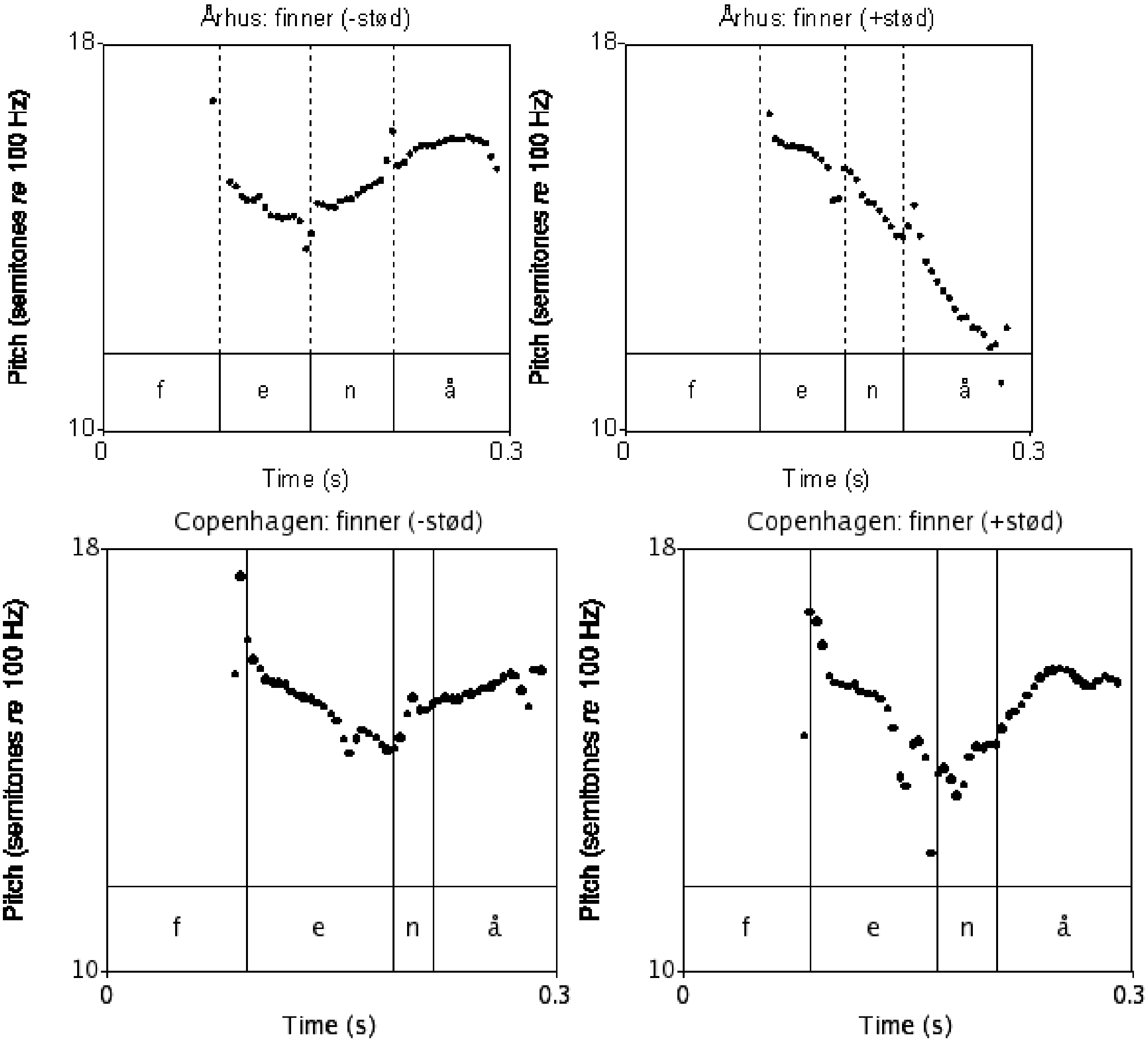
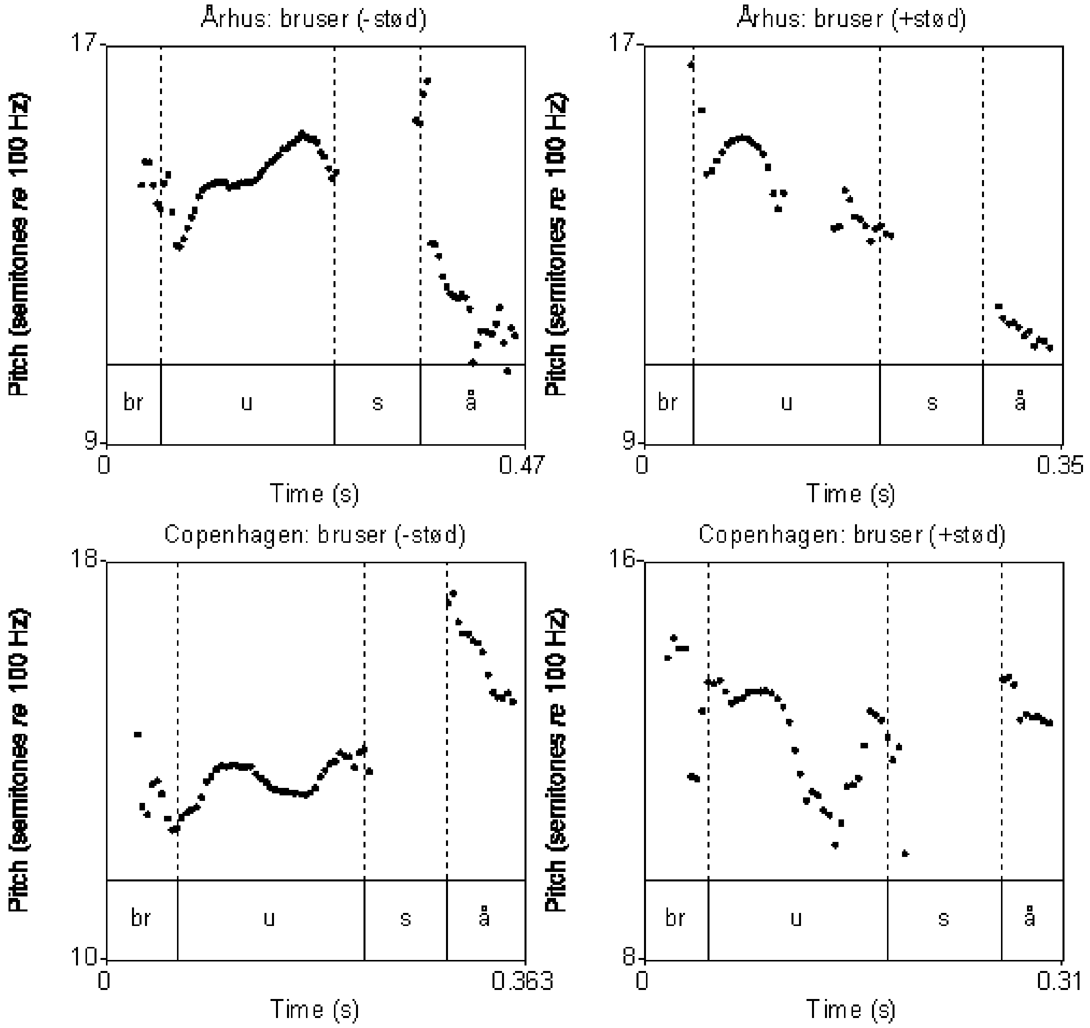


Fig.6 Group 2b vs. group 1, Århus[4] vs. Copenhagen



[4] Stress groups with short tonic vowels (and at least one non-sonorant element between the two vowels) also have this high-low tone pattern Ű see (Kyst, 2004)

Fig.7

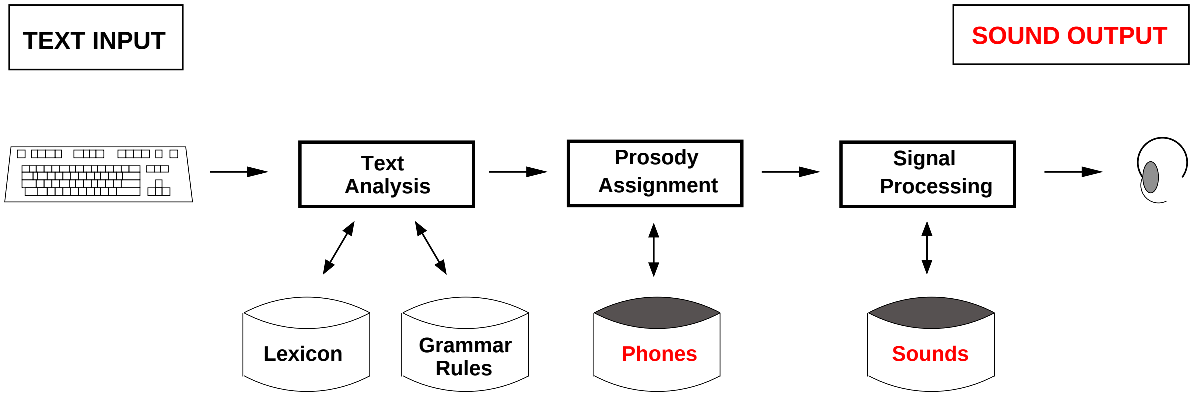


Fig.8

Played / Heard	Natural Århus	Natural Copenhagen	Synthetic Århus	Synthetic Copenhagen
Natural Århus	6/6			
Natural Copenhagen		6/6		
Synthetic Århus			4/6	1/6
Synthetic Copenhagen			2/6	5/6