Danish Text-to-Speech Synthesis Based on stored acoustic segments

Hoequist, Charles
Center for PersonKommunikation, Aalborg University,
Fredrik Bajers Vej 7, DK-9220 Aalborg Øst, Denmark
E-post: ch@cpk.auc.dk

Abstract

As part of a Danish Research Ministry strategy for speech technology R&D, a text-to-speech system has been developed for Danish. In this paper we will look at that part of the initiative which is concerned with synthetic speech. The system used relies on a database of encoded speech segments. Advantages and disadvantages in comparison to signal generation by calculation are discussed.

1. Background

A report [1], prepared under the auspices of the Ministry of Research shows that the synthesis systems used in Denmark today are of such low quality that even understanding the synthesized output requires user training. Obviously, this rules out many promising applications.

For instance, only a limited number of Danes with visual impairments are able to use such equipment for interpreting printed material. People whose sight loss occurs in middle age or later will in all probability not be able to learn to use currently available equipment. Dyslexic, aphasic or illiterate users suffer a similar disadvantage.

The development of high-quality Danish synthetic speech is, however, crucial to the Government's IT policy of making the information society available to all. The Research Ministry therefore entered into a contract with a consortium made up of the Center for PersonKommunikation (CPK), Aalborg University, Aalborg; The Institute for General and Applied Linguistics (IAAS), Copenhagen University, Copenhagen; Tele Danmark A/S (TDK), Taastrup; Tawido ApS (TAW), Aalborg, and Dansk

Taleteknologi A/S (DT), Aalborg. The two university partners - IAAS and CPK - are undertaking research within language and digital speech processing, respectively. It is the CPK work that is covered in this paper.

2. Synthesis System Overview

2.1. System architecture

The architecture of the Text-to-Speech system is as shown in Figure 1, see also [2].

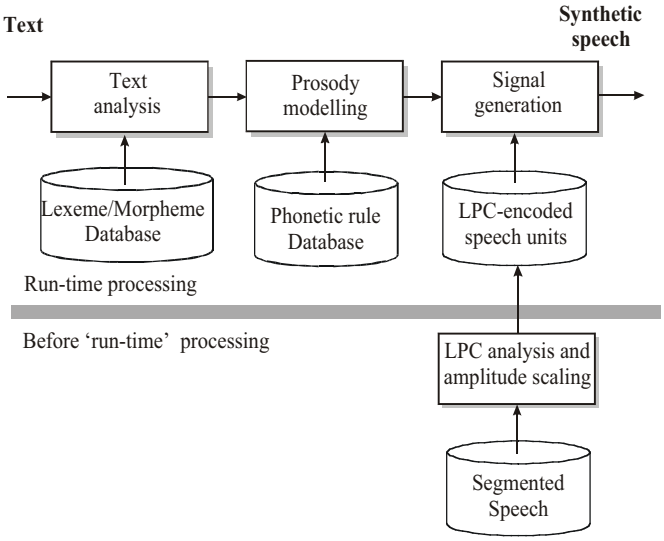

Figure 1. Architecture of the text-to-speech system

Input consists of two parallel streams: text and a subset of MS SAPI tags (see below) which are supported by DST (Dansk Syntetisk Tale).

The process of moving from a standard Danish orthographic input to acoustic output is handled in three stages: text preprocessing and analysis, prosody assignment and sound generation. Note that this is a process description; the corresponding software architecture is covered in section 4.

2.2. Text normalization and analysis

The first step in handling the text is preprocessing to turn all non-lexical forms (digits, abbreviations) into alphabetic representations of the intended spoken string. The output then goes to the text-analysis module, which performs both a morphological and a syntactic breakdown of the input. Words may be accessed as full lexical items (this is always the case for user-added items), broken down into component morphemes, or spelled out. Where the acoustic output is affected by syntactic factors, the output symbol string can include tags, e.g. where a syntactic boundary can trigger a pause.

2.3. Prosody assignment

The output string from the text analysis module is passed to the prosody assignment module, which assigns default duration and F0 values to each segment, as well as any pauses or F0 slope changes resulting from text-analysis tags or the presence of stød. Stød in this system is not a prerecorded diphone, but rather a sudden, sharp pitch drop, the acoustic realization of the stød's glottal constriction.

2.4. Sound generation

At this point in many systems, in particular older ones, the segment labels and prosodic instructions would be interpreted in terms of formant parameters, such as target frequencies, amplitudes and slopes, as in the Klatt [3] and Holmes [4] formant synthesizers. The parameters drive the generation of periodic and aperiodic signals to mimic the acoustics of human speech. This synthesis principle is still used in many commercial synthesizers today, not least because of the small memory footprint and the great freedom of control of the acoustic output. However, the quality of the output speech depends heavily on the rule set, which itself takes considerable time and expertise to develop. While copy-synthesized utterances using formant synthesizers show that extremely high quality is in principle possible, rule-based systems available to date are unable to approach this quality.

With the steady drop in cost for computer storage, concatenative synthesis as an alternative to rule-based generation is becoming more popular. Since concatenative synthesis relies on a pre-recorded speech database, the quality (particularly of the voice source) has the potential to be at least as high and generally much higher than rule-based generation. The DST system makes use of such a database.

The output of prosody assignment serves as instructions for selecting which RELP-encoded (Residual Excited Linear Prediction) diphones from the database are to be concatenated. The concatenated sounds are modified in accordance with any F0 or rate modifications passed along in the prosody module output.

In the course of system development, considerable effort has gone into optimizing the database. Some optimizations are purely computational, as in the development of faster search procedures for diphones. Others, which attempt to lower the number of diphones in the database, depend for their acceptance on users' judgment of the quality of the resulting output.

For example, the implementation of stød as a runtime signal adaptation allowed the elimination of some 1200 stød-containing segments from the diphone database, reducing it to its present size of 2600 diphones, with a corresponding reduction in footprint, with little or no loss of intelligibility or quality for listeners [5].

However, an attempt at further storage reduction by creating short vowels from their long counterparts resulted in an overall loss of intelligibility [5], despite the apparent similarity of the short and long vowels' formant structures.

3. Design and Implementation

3.1. Interfacing to applications

Given the desire for a commercially feasible system, the Ministry contract specifies that the synthesis system is to be compatible with Microsoft's Windows 95/98/NT operating systems and that it be usable with a wide range of existing and future applications, e.g. screen readers and internet browsers. It was therefore decided to implement the DST system to support Microsoft's Speech Application Programming Interface (MS SAPI). The system is called from applications as shown in Figure 2. This design makes it possible for existing third-party applications which already use

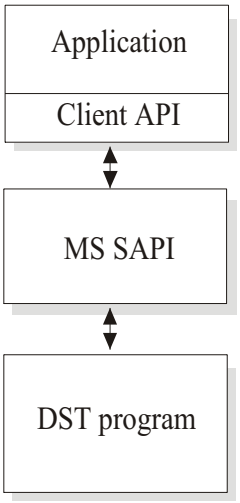

Figure 2. DST interface to MS SAPI compatible applications

MS SAPI for access to and control of a TTS for languages other than Danish to make immediate use of the DST system for Danish TTS. The DST system can also be used from C/C++, Visual Basic or OLE Automation, since MS SAPI complies with OLE COM (Object Linking and Embedding, Component Object Model).

DST system development plans include maintaining compliance with all aspects of future MS SAPI releases

which are relevant to DST functionality. This offers developers of TTS products a stable and publicly-available interface, which in turn increases the likelihood of significant commercial distribution for DST.

In order for the program to be usable with MS SAPI, it is implemented as a DLL (dynamic link library) for running under Windows. In addition, DST relies on an initialization file and databases with language-specific rules and diphones.

At the same time, the DST system has avoided building in Windows dependencies in the code wherever possible. The only platform-dependent module is the interface to MS SAPI (see following section). Use of DST under other operating systems is therefore feasible without too much effort.

3.2. Identification of modules

The synthesizer can be broken down to a number of individual modules as shown in Figure 3. This division makes it possible for parallel development of individual modules and their separate testing before integration into the product.

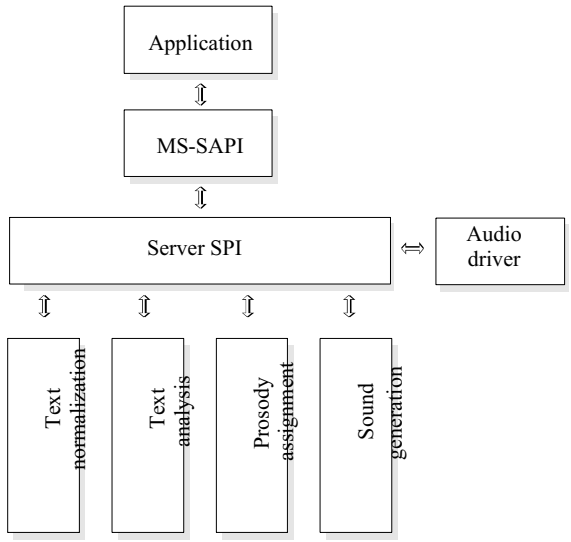

Figure 3. DST module architecture

Briefly, the modules are:

- **SPI** (Service Provider Interface), which serves as a connection between MS SAPI and the other modules in the system. The SPI interprets queries from applications (*Clients*), creates sentences out of a section of blocks of text and calls the underlying modules. MS SAPI tags are sent as a parallel stream via the processing interface, where every module in the processing path can inspect them for possible activity required in the particular module.
- **Text normalization,** which converts recognized abbreviations, dates, telephone numbers, etc. to their orthographic full forms.
- **Text analysis,** which maps orthographic strings to entries in the lexicon wherever possible.
- **Prosody assignment**, which calculates and annotates duration and pitch (F0) changes for the individual segments.
- **Sound generation,** which concatenates stored diphones to create an output audio stream.
- **Audio driver,** installed as a part of the Windows OS, is used to play out the synthesized speech signal.

4. Quality Measures

The foremost reason for the Research Ministry to initiate the project was to support the development of a high-quality Danish text-to-speech product. Quality is in this context to be interpreted as intelligibility and naturalness. The first commercial version of the synthesizer will at least be on a par with a demo version, which has been assessed as described below [6].

4.1. Intelligibility

The intelligibility test included synthetic speech from the DST system and as a reference natural speech.

The 32 test subjects listened to a total of 1600 words from the two categories. Each word was embedded in a carrier sentence: "*Der er [test word] de siger" ("It is [test*

word] they are saying") The percentages of words misheard are illustrated in Figure 4.

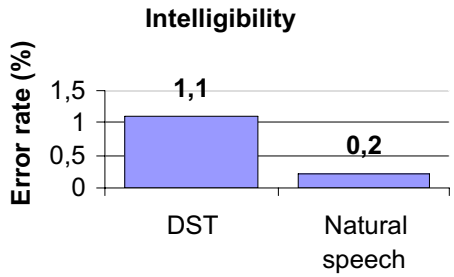

Figure 4 Word error rates for DST and natural speech

As expected, natural speech comes out with the highest intelligibility with an error rate of 0.2%, However, the DST system is also demonstrating high performance with only 1.1% errors.

4.2. Naturalness

The naturalness of the system was assessed by asking the same 32 subjects used in the above test to evaluate the naturalness of an utterance on a MOS (Mean Opinion Score) scale with values from 1 to 5. The higher the MOS is the more natural the utterance is. The test subjects were given speech from three different categories: natural speech, synthetic speech produced by the Infovox system 230, synthetic speech produced by the DST system. Figure 5 summarizes the results of the naturalness test. The naturalness of the DST system comes out with a score of 2.29, roughly midway between the Infovox score of 1.1 and a real speech score of 4.63.

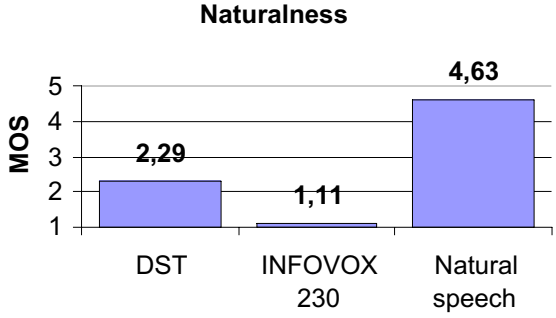

Figure 5: Naturalness (MOS) of DST, natural speech and Infovox

5. Future Plans

The project is moving ahead with plans for further quality improvements. These fall into two categories: first, the removal of artifacts based on the synthesis method, and second, a better and more robust modeling of natural speech as derived from text.

The primary artifact in concatenated synthesis is of course the potential for a discontinuity in the signal at every concatenation point. This is currently addressed by the preprocessing stage, where concatenation points are placed in low-amplitude sections of the signal. Work is underway to investigate the value of various types of signal smoothing at run time as well.

Modeling of natural speech occurs at various stages of the system. A later release will improve both text rules, to handle misspelled input, and the current prosody rules, to better model the pitch contour and pause structure of utterances. The possibility of expanding the diphone inventory to handle voicing variation in realizations of the Danish /r/ is being investigated, as well as some re-recording of the base speech for the diphone encoding. The re-recording is intended to address gaps in the original recordings, which did not adequately account for presence or absence of stress on a target diphone.

An additional area of research is the construction of the segment databases themselves. Creating a segment database requires less specialized knowledge and experience than the development of a rule system for formant synthesis. The disadvantage is that the database is dependent on the original recordings, and any segments not present there, or present but with low quality, usually require a new recording session and rebuilding of the database. This places a premium on having as much recorded and tagged material as possible to choose from. Unfortunately, the available speech databases are not geared toward this need. Most are designed to supply material for benchmarking speech recognizers, and there is little tagging of the kind common in text databases, where the database contexts are tagged. Speech databases would become very useful for concatenative synthesis if tagged with transcriptions and even analysis parameters.

6. Conclusions

The Danish Research Ministry has concluded that Danish synthetic speech is lagging behind comparable countries in quality and degree of market penetration.

To remedy this situation the Ministry has given a consortium the task to develop a new generation of Danish text-to-speech engines. Specific quality measures in terms of intelligibility and naturalness are intended to ensure that the system will surpass what is now available for Danish. The synthesizer is compliant with the MS SAPI interface. Hence, many applications as for instance screen readers, talking e-mails etc. can immediately be used together with the synthesizer. The above factors together with an attractive pricing will ensure another Ministry objective: widespread deployment of high-quality Danish text-to-speech.

7. References

[1] Dansk Syntetisk Tale 1996 (februar). Udarbejdet for Forskningsministeriet af Hjælpemiddelinstituttet.

[2] Jensen, J., Nielsen, C., Andersen, O., Hansen, E., and Dyhr, N.J. 1998.. A Speech Synthesizer with Modelling of the Danish "Stoed". In IEEE Nordic Signal Processing Symposium (NORSIG '98). Vigsø, Denmark.

[3] Klatt, D. H. 1980. Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, 67:971-995.

[4] Holmes, J.N. 1985. A parallel-formant synthesizer for machine voice output. In F. Fallside and W.A. Woods (eds.). *Computer Speech Processing*. London:Prentice-Hall: 163-187.

[5] Andersen, O., Dyhr, N.-J., Nielsen, C. 1999. On Synthesizing Danish Short Vowels. In *Proceedings of the XIVth International Congress of Phonetic Sciences* (ICPhS '99). San Francisco:2291-2294.

[6] Bagger-Sørensen B. (1997). Testrapport (ter) for FRITSYN (Lyttetest) version 1.0. Tele Danmark, Udviklingsområdet.