# Evaluating the speech quality of the Norwegian synthetic voice Brage

**Marius Olaussen**
Norwegian Library of Talking Books and Braille
Oslo, Norway
marius.olaussen@nlb.no

## Abstract

This document describes the method, results and conclusions from my master's thesis in Nordic studies. My aim was to assess the speech quality of the Norwegian Filibuster text-to-speech system with the synthetic voice Brage. The assessment was carried out with a survey and an intelligibility test at phoneme, word and sentence level. The evaluation criteria used in the study were intelligibility, naturalness, likeability, acceptance and suitability.

## 1 Introduction

### 1.1 Background

Visually impaired and print disabled students in higher education have a need for adapted literature. In Norway the Norwegian Library of Talking Books and Braille (NLB) is responsible for such adaptation. Over half of the academic literature is produced with TTS. All audio books are produced as DAISY books. To strengthen the services given to students, NLB appropriated a million Norwegian kroner by the Ministry of Education and Research. The library signed collaboration with the Swedish Library of Talking Books and Braille (TPB) to adjust their TTS system Filibuster to Norwegian Bokmål. Bokmål is one of the two written varieties of Norwegian. The second variety is Nynorsk. In late 2009, the Norwegian synthetic voice *Brage* was launched.

### 1.2 The present study

The aim of this study was to evaluate the speech quality of Brage with respect to the suitability to impart academic literature. No similar studies regarding evaluation of synthetic speech quality have previously been carried out in Norway. Four research questions were formulated:

1. How do visually impaired and print disabled students experience the speech quality of Brage assessed by key criteria given in evaluation methodology?
2. How intelligible is Brage compared to other Norwegian synthetic voices?
3. How suitable does Brage seem to be as an imparter of academic literature?
4. How should Filibuster with Brage further develop?

## 2 Brief description of Filibuster TTS

TPB needed a TTS system especially trained for processing textual challenges distinctive of academic texts. In 2007, the Filibuster TTS system was implemented in production with the first Swedish voice *Folke*. The system is based on concatenation with unit selection.

The Norwegian Filibuster system uses a pronunciation dictionary of somewhat 780,000 entries from The Norwegian language resource collection. All entries are transcribed in SAMPA (Wells, 2005). The speech database was recorded at NLB with a manuscript of 15,604 Norwegian utterances created from a text corpus consisting of 10.8 million words from academic literature, newspapers, magazines and official Norwegian reports (Sjölander and Tånnander, 2009). In addition, an English manuscript of approximately 1,150 English utterances from the CMU ARCTIC database was used (Kominek and Black, 2003). 15 xenophones were applied.

## 3 Methodology

### 3.1 A survey

A survey was carried out with a questionnaire designed with the recommendations of the ITU-T evaluation method (Jekosch, 2005). This is a useful method to operationalise key evaluation criteria such as intelligibility, naturalness, likeability, acceptance and suitability.

**Bolette Sandford Pedersen, Gunta Nešpore and Inguna Skadiņa (Eds.)**
**NODALIDA 2011 Conference Proceedings, pp. 336–339**

### 3.2 An intelligibility test

The SUS test (*Semantically Unpredictable Sentences*) was adopted to perform an intelligibility test supplementary to the survey. The SUS test is primarily a sentence level intelligibility test (Benoît, Grice and Hazan, 1995), but has also been applied at word level (Boula de Mareüil et al., 2006). In this study the SUS test was applied at phoneme, word and sentence level to compare Brage to two other Norwegian synthetic voices, which respondents in the survey stated as their favourites: *Kari* (Acapela Group) and *Stine* (Nuance). Like Brage, both of these voices are based on concatenation with unit selection. One of the reasons for choosing the SUS test over other intelligibility tests is the removal of semantic information. Thus the informants cannot use contextual cues to guess the right words.

The SUS sentences are generated with five syntactic structures (*intransitive, transitive, imperative, interrogative and relative*) limited by a set of syntactic and lexical constraints. The test material mainly consists of high frequent monosyllabic words. The test designers developed SUS generator software, but since the software wasn't supported by later OSs, I decided to develop a new one, in addition to a frequency and part of speech list generator. The monosyllabic frequency lists were based upon all books produced with Brage in the course of one year. All software and both audio and textual test material used in this study are available for download at `www.teksttiltale.no`.

### 3.3 Informants

19 visually impaired and 34 other print disabled Norwegian students in higher education participated in the survey. Most of the students had little experience with TTS (57 % stated they had used TTS for less than a year). In the SUS test, 18 informants participated. To avoid biased results that might not correlate to the actual intelligibility, none of the informants were registered as patrons at NLB nor were print disabled.

## 4 Account of results and discussion

### 4.1 The SUS test

In the SUS test, Brage received on the average higher scores than Kari and Stine. The sentences were distributed in such a way that six informants heard the same sentence with the same TTS. Of a total of 60 sentences all six informants reiterated 26 sentences correctly with Brage, compared to 10 sentences with Kari and 7 sentences with Stine. At word level Stine scored 330 of a total of 408 possible points (8 % less than Kari and 14 % less than Brage). Table 1 shows the distribution of scores at phoneme level.

| Phonemes | Brage | Kari | Stine |
|---|---|---|---|
| **All vowels** | 0.9926 | 0.9558 | 0.8949 |
| - Front | 0.9912 | 0.9561 | 0.8972 |
| - Central | 0.9947 | 0.9520 | 0.9173 |
| - Back | 0.9938 | 0.9604 | 0.8521 |
| **All diphthongs** | 0.9872 | 0.9872 | 0.8077 |
| **All consonants** | 0.9777 | 0.9578 | 0.9225 |
| - Bilabials | 0.9841 | 0.9722 | 0.8948 |
| - Labiodentals | 0.9773 | 0.9621 | 0.9343 |
| - Dentals, alveolars, postalveolars | 0.9761 | 0.9558 | 0.9279 |
| - Retroflexes | 0.9889 | 0.9889 | 0.9333 |
| - Palatals | 0.9849 | 0.9697 | 0.9242 |
| - Velars | 0.9762 | 0.9544 | 0.9028 |
| - Glottals | 1.0000 | 0.9028 | 0.9306 |

Table 1: The SUS test results at phoneme level

A unique feature of the voice quality of Brage, distinguishing this voice from other Norwegian synthetic voices, is the reading speed. To demonstrate this, I carried out a comparison test with eight other Norwegian voices. The test results showed that Brage reads 27 % slower than the average. Kari reads 5 % faster than the average and Stine reads 13 % faster. Since Brage on average scored higher than Kari and Stine in the SUS test, there seems to be a correlation between reading speed and the ability to define word boundaries. Findings indicated that determiners and conjunctions play a role in the intelligibility at sentence level. For instance, there were fewer incorrect reiterations of articles recorded with Brage (3 %), compared to Kari (8 %) and Stine (13 %). Such a possible correlation has also been pointed out in previous studies (Neovius and Raghavendra, 1993).

### 4.2 The survey

#### 4.2.1 Intelligibility

The findings in the SUS test are to be understood as an indication of the overall intelligibility. The user experienced intelligibility seemed, however, to correlate to the findings of the SUS test; 81 % of the respondents in the survey stated Brage generally had either an intelligible or quite intelligible articulation. The user experienced intelligibility appeared to be closely related to how well the respondents thought Brage handled academic terminology within their branch of study. Respondents who studied law, political sciences, economic sciences and business and management found that Brage did not impart terminol-

ogy in their curriculum in any acceptable manner.

When it comes to Norwegian Nynorsk, 75 % of the students stated their curriculum didn't contain elements of Nynorsk. However, despite a low coverage of Nynorsk entries in the pronunciation dictionary (0.2 %), 21 % of the students found that Brage handled Nynorsk either well or quite well. To some extent this also applied to English; about half of the respondents (54 %) considered the English pronunciation to be good.

Furthermore, a correlation between experience and speech perception was observed. Among the respondents who considered the overall articulation of Brage to be either unintelligible or quite unintelligible (20 %), 82 % stated that they had been using a Norwegian TTS for less than three years or not at all. Similar observations have also been made in previous studies (Francis, Nusbaum and Fenn, 2007).

Addressing particular textual challenges, the students stated that Brage did not process digits and numeral phrases, homographs and foreign proper names in a satisfactory manner.

#### 4.2.2 Naturalness

57 % of the respondents liked the voice of Brage either well or quite well. Respondents who liked it less or not at all also report they had less experience with speech synthesis. Although many liked the voice itself, 45 % of the respondents thought Brage was unnatural or quite unnatural. None of these students, however, had used speech synthesis for more than six years. In comparison, none of the respondents who had used a Norwegian speech synthesis for seven years or more believed Brage to be unnatural.

86 % of the students stated it was important or quite important that synthetic speech resembles human speech to a technologically possible extent. Nonetheless, 6 % reported that it didn't matter at all. Interestingly, 8 % preferred synthetic speech to human, particularly justifying this with the shorter production time, the potential of larger quantum of academic literature and the direct access to the book content electronically.

Furthermore, 22 % of the respondents found the prosody of Brage to be good, while 26 % had no remarks at all. Prosodic weaknesses specially pointed out concerned stress (15 %), rhythm and intonation (24 %) and incorrect reproduction of syllables (4 %). 60 % of the respondents who thought Brage sounded either unnatural or quite unnatural, had remarks concerning prosodic characteristics. This indicates the importance of prosodic characteristics for the user experienced naturalness of a synthetic voice.

#### 4.2.3 Likeability

34 % of the students thought it was either pleasant or quite pleasant listing to Brage over time, while 30 % stated it was ok, and 36 % found it either unpleasant or quite unpleasant. This may be due to a number of things. Firstly, the individual preferences seemed to vary, particularly regarding reading speed. Similar observations were done in other evaluation studies (Furui, 2007). But still there doesn't seem to have been carried out any studies addressing the cause of such variation.

The practise of speech synthesis seemed to play a key role in the assessment of user experienced likeability; all the students who did not prefer synthetic adaptation (71 %) had been using Norwegian TTS for less than a year. This corroborates the importance of encouraging the use of TTS to a larger extent, in order to get positive user experience (Francis, Nusbaum and Fenn, 2007).

Regarding concentration problems, 83 % stated they would strive more to retain focus when reading texts adapted with synthetic speech, compared to texts adapted with human speech. More experienced students seemed to exert less than those with less experience.

#### 4.2.4 Acceptance

About half of the students (45 %) preferred human to synthetic speech. Students who reported they had more experience, however, showed a greater acceptance for such adaptation.

45 % stated they had good confidence in Brage as imparter of academic literature, while 28 % had some confidence and 26 % little or no confidence whatsoever. The lack of confidence was justified in particular by user experienced intelligibility together with naturalness and likeability. It is therefore crucial to improve the intelligibility, for instance by finding an effective way to ensure that frequent terms in academic literature of various branches of study is pronounced correctly.

## 5 Conclusions

### 5.1 Summary

Assessment of the suitability to impart academic literature should be carried out as a sum of the other key evaluation criteria (Jekosch, 2005; King, 2007). In this regard, Brage seemed to im-

part academic literature in an overall acceptable manner. 15 respondents preferred Brage over other synthetic voices. This is five times as many compared to the other voices. It is interesting that 9 of the 15 respondents preferring Brage, also reported that they had only made use of a Norwegian TTS for less than a year. However, only 4 of these 9 students stated they had knowledge of other Norwegian synthesizers. Thus, this shouldn't necessarily be understood as an acceptance for Brage, but rather for TTS in general.

When it comes to characteristics, a unique feature with Brage is the slow reading pace, which would result in distinct word boundaries, but also frustration among users preferring to read faster.

Academic literature spans a wide range of disciplines. Findings in the study indicate that Brage seemed to be less suited to impart texts within certain disciplines. Prior to a documented acceptable degree of coverage of academic terminology, it is recommended to adapt less suited academic literature with human speech rather than synthetic, until the system has been improved. In the wake of this recommendation, tools for mapping out the coverage of the most frequent academic terminology within different branches of study have recently been developed, providing statistical overview of new and missing entries. A coverage test was carried out with a test material of 70 academic books, divided between seven different branches of study. Results from this test showed that the coverage of terminology within law (42 %), economic sciences, business and management (38 %), and political sciences, was somewhat higher compared to the coverage in academic literature within the branches of study which the respondents in the user survey believed Brage imparted well or in an acceptable manner. This finding seems to indicate that the terminological coverage has less impact on how suitable students find Brage to be imparting their syllabus than previously presumed. These tools are presently being expanded with automatic phonetic transcription and morphological annotation suggestions.

### 5.2 Further work

Any given academic text will almost invariably contain words not yet listed in the pronunciation dictionary. An intermediate solution to this challenge could be to develop a spelling feature in the DAISY player ensuring the reader access to the text. Furthermore, since 45 % of the informants stated Brage was either unnatural or quite unnatural, initiatives to increase the prosody should be prioritised. One possible solution is phrase splicing (Donovan et al., 1999).

## References


Benoît, C., Grice, M. and Hazan, V. 1995. *The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences.* Speech Communication, Vol. 18, 1996:381-392.

Boula de Mareüil, P., d'Alessandro, C., Raake, A., Bailly, G., Garcia, M-N. and Morel, M. 2006. *A joint intelligibility evaluation of French text-to-speech synthesis systems: the EvaSy SUS/ACR campaign.* Proc. of LREC2006, Genoa, Italy.

Donovan, R. E., Franz, M., Sorensen, J. S. and Roukos, S. 1999. *Phrase Splicing and Variable Substitution using the IBM Trainable Speech Synthesis System.* Proc. of ICASSP'99, Phoenix, AZ.

Francis, A. L., Nusbaum, H. C. and Fenn, K. 2007. *Effects of Training on the Acoustic-Phonetic Representation of Synthetic Speech.* Journal of Speech, Language and Hearing Research, Vol. 50, Nr. 6:1445-1465.

Furui, S. 2007. *Speech and Speaker Recognition Evaluation.* In Dybkjær, L., Hemsen, H. and Minker, W. 2008. Evaluation of Text and Speech Systems. Text, Speech and Language Technology, Vo. 37. Springer, New York.

King, M. 2007. *General Principles of User-Oriented Evaluation.* In Dybkjær, L., Hemsen, H. and Minker, W. (ed.). Evaluation of Text and Speech Systems. Text, Speech and Language Technology, Vol. 37:125-161. Springer, New York.

Kominek, J. and Black, A. 2003. *CMU ARCTIC databases for speech synthesis.* Report CMU-LTI-03-177. Language Technologies Institute. Carnegie Mellon University, Pittsburgh. http://festvox.org/cmu_arctic/cmu_arctic_report.pdf.

Jekosch, U. 2005. *Voice and Speech Quality Perception. Assessment and Evaluation.* Springer. Heidelberg, Germany.

Neovius, L. And Raghavendra, P. 1993, *Evaluation of comprehension of KTH text-to-speech with 'listening speed' paradigm*, STL-QPSR, Vol. 34:21-30.

Sjölander, K. and Tånnander, C. 2009. *Adapting the Filibuster text-to-speech system for Norwegian Bokmål.* Proc. of FONETIK 2009. Available at http://www.ling.su.se/fon/fonetik_2009/036%20sjolander_tannander_fonetik2009.pdf.

Wells, J. 2005. *SAMPA computer readable phonetic alphabet.* Department of Speech, Hearing and Phonetic Sciences, University College London. http://www.phon.ucl.ac.uk/home/sampa/.



ISSN 1736-6305 Vol. 11
http://hdl.handle.net/10062/16955