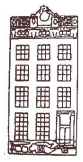
Speech recognition and synthesis



Introduction

Speech technology for "disadvantaged" languages

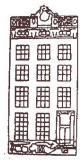
- Language barriers limit access to digital resources
- Speech technology needed for access to services, eg, phone services
- Language often part of national, cultural, and political identity
- Lack of Language and Speech technology will put communities at a disadvantage
- Many speech technology projects for "minority" languages started by single "students" of the language

Introduction

Basic speech technology projects

- Demonstration TTS or ASR systems *can* be build by a single person
- All tools available on the internet for free
- Basic systems for a new language take around 3-6 person months
- Systems and work are modular
- Systems should be constructed iteratively
- Start with an existing system, and change it gradually
- If digital resources are available, *use them!*

See `http://www.fon.hum.uva.nl/IFA-publications/Others/Other_papers.html`



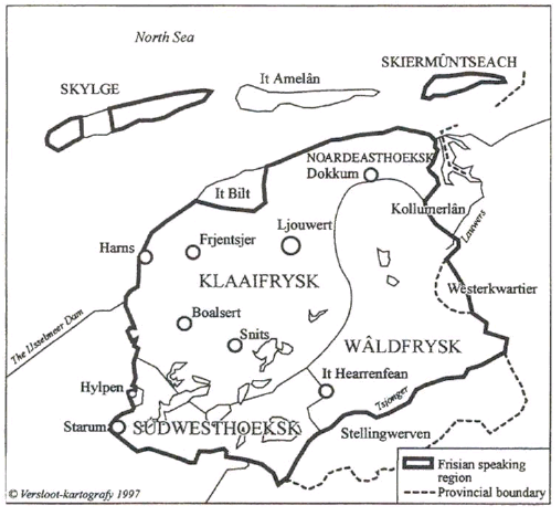
Example 1: A basic Frisian TTS

Master's thesis

- No speech technology available for Frisian
- Language community is organized scientifically
- There is "political" demand for Frisian Language Technology
- Student is a native speaker
- 4 Month thesis project
- Dutch diphones (no time to create Frisian set)
- Aim: "bootstrap" the development of a TTS system

[Dijkstra et al.(2005)Dijkstra, Pols, and van Son]

Example 1: A basic Frisian TTS: West Frisian dialects in the Netherlands



Map 1: Dialect map of Fryslân (Versloot cartography 1997, in: Visser, 1997)

West Germanic language (Indo-European)

- Main dialects: Klaaifrysk, Wàldfrysk, and Sûd-Westhoeksk
- Standard Frisian based on Klaaifrys
- Official status since 1970

[Dijkstra et al.(2005)Dijkstra, Pols, and van Son]

Example 1: A basic Frisian TTS

Total population of *Friesland* $> 634,000$ [Gorter and Jonkman(1995)]

- 55% native speakers (350,000)
- 74% understands Frisian (470,000)
- 65% reads Frisian (410,000)
- 17% writes in Frisian (110,000)

[Dijkstra et al.(2005)Dijkstra, Pols, and van Son]

Example 1: A basic Frisian TTS

Start with an existing system, and change it gradually

- *Frisian* is close to *Dutch* in many respects
- *Nextens* and those that build it were available
- Contacts with the *Fryske Akademy* could supply language help
- A digital pronunciation lexicon could be "borrowed"
- Technical and community support were available

Example 1: A basic Frisian TTS

Language resources and tools

- *Fryske Akademy*
- MBROLA [MBROLA(2005)]
- Nextens [Nextens(2003)]
- Festival [Black and Lenzo(2003a)]
- Pre-publication of "Frysk Hânwurdboek" (Concise dictionary)
- Worldbet [Hieronymus(1994)]
- Enthusiasm from everyone

Example 1: A basic Frisian TTS: Nextens

The architecture of NeXTeNS (Festival):

- **Token Module**: Tokenization
- **POS Module**: Part-Of-Speech tagging
- **Syntactic Module**: Syntax parsing (*disabled*)
- **Phrasing Module**: Phrase break prediction
- **Intonation Module**: Sentence accents
- **Tune Module**: Tune choice needed for ToDI
- **Word Module**: Grapheme-to-phoneme conversion
- **Pauses Module**: Insertion of pause segments
- **Postlexical Module**: Anything left over
- **Duration Module**: Segment and pause-durations
- **Fundamental frequency control**: ToDI $\Rightarrow$ utterance
- **Waveform synthesis**

Example 1: A basic Frisian TTS

Building a Frisian TTS

- Construct Frisian Worldbet phonetic alphabet [Hieronymus(1994)]
- Convert pronunciation lexicon to Worldbet
- Phrasing, Tune, Pause: Use Dutch (small adaptations)
- Tokenization: Enter Frisian numbers and abbreviations
- POS: Translated Dutch *Function* wordlist
- POS: Use only *Content/Function* word difference
- Intonation: Accent every other *Content* word

Example 1: A basic Frisian TTS

Building a Frisian TTS: Word module

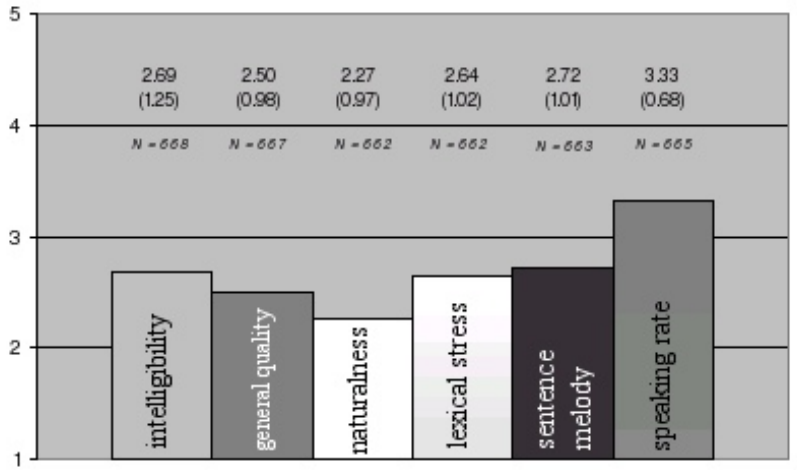
- Pronunciation lexicon
- Letter-to-Sound rules, eg,
  `(VOWEL [- g] VOICEDC = - G)`
- Syllable stress rules, i.e. strong/weak syllables
- Map complex sounds, eg, nasalized vowels and triphthongs

Example 1: A basic Frisian TTS

Building a Frisian TTS: Other modules

- Postlexical: Adapted Dutch rules
- Postlexical: Map Frisian worldbet to Dutch SAMPA symbols
- Duration: Shorten schwa, change duration long vowels
- Fundamental frequency: Adapt Dutch ToDI module
- Waveform synthesis: Map each "Frisian" phone to the "nearest" Dutch MBROLA phone

Example 1: A basic Frisian TTS: Evaluation



Mean judgments for 20 test sentences

- End evaluation over WWW with 32 native subjects
- 10 short ($<$ 13 words) 10 long ($\geq$ 13 words)
- Example of short and long sentence
- 6 qualities on a 5 point scale (higher is better/more rapid)

Example 1: A basic Frisian TTS

	short $N \approx 331$	long $N \approx 331$	total $N \approx 662$
intelligibility	2.57 (1.25)	2.80 (1.24)	2.69 (1.25)
quality	2.51 (0.99)	2.50 (0.97)	2.50 (0.98)
naturalness	2.31 (0.97)	2.22 (0.97)	2.27 (0.97)
lexical stress	2.67 (1.05)	2.58 (0.99)	2.64 (1.02)
sentence melody	2.79 (0.99)	2.64 (1.02)	2.72 (1.01)
speaking rate	3.30 (0.65)	3.35 (0.71)	3.33 (0.68)

Mean judgments (standard deviation)

- Mean ratings *below 3* (mid-point)
- *Naturalness rated* lowest
- Sentence length did not change ratings
- Ratings were above *1!*
- Note: This was done using a Dutch diphone set

Judgments on a 5 point scale, higher is better. For speaking rate higher is more rapid

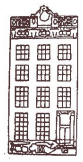
Example 2: Digit recognition in two languages

Kinyarwanda: Official language of Rwanda

- Niger-Congo Language
  `http://www.nvtc.gov/lotw/months/september/niger.html`
- 7-8 million native speakers
- Many Rwandese are monolingual
- Recognizer build by *Muhirwe Jackson* for his Master of Science thesis [Jackson(2005)]
- Computer Science of Makerere University, Kampala, Uganda
- Implements the tutorial digit recognizer from HTK

[Jackson(2005)]
[Young et al.(2004)Young, Evermann, Hain, Kershaw, Moore, Odell, Ollason, Povey, Valtchev, and Woodland]

[HTK(2002)]



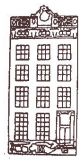
Example 2: Digit recognition in two languages

Dutch: Official language of the Netherlands

- West Germanic language, 21 million native speakers
- Masters of Science course for AI students
- University of Amsterdam
- Speech Technology project
- 1 month, 6 students [Adriaans et al.(2004)Adriaans, Heukelom, Koolen, Lentz, de Rooij, and Vreeswijk]
- Implements the tutorial telephone application from HTK

[Adriaans et al.(2004)Adriaans, Heukelom, Koolen, Lentz, de Rooij, and Vreeswijk]
[Young et al.(2004)Young, Evermann, Hain, Kershaw, Moore, Odell, Ollason, Povey, Valtchev, and Woodland]

[HTK(2002)]



Building a basic ASR system

Tasks

1. Building the task grammar
2. Constructing a dictionary for the models
3. Recording the data.
4. Creating transcription files for training data
5. Encoding the data (feature processing)
6. (Re-) training the acoustic models
7. Evaluating the recognizers against the test data
8. Reporting recognition results

Building a basic ASR system

```
# Kinyarwanda
$digit=RIMWE | KABIRI | GATATU | KANE | GATANU | GATANDATU | KARINDWI | UMUNANI
| ICYENDA | ZERO;
(SENT-START [ $digit ] SENT-END)

# Dutch
$digit = EEN | TWEE | DRIE | VIER | VIJF | ZES | ZEVEN | ACHT | NEGEN | NUL;
$name = [ ROB ] (VAN SON) | [ FRANS ] ADRIAANS | [ TOM ] LENTZ | [ MARIJN |
MARINUS ] KOOLEN | [ ORK ] (DE ROOIJ) | [ MARKUS ] HEUKELOM | [ DAAN ]
VREESWIJK;
( SENT-START ( DRAAI <$digit> | BEL $name) SENT-END )
```

Task Grammars

- Define digits and names
- Define grammar on vocabular
- Square brackets enclose optional items

Building a basic ASR system

Construct pronunciation dictionary

- Make a word list of all words in the training corpus or a suitable text corpus
- Transcribe the words by hand or use a TTS system (eg, *Nextens*)
- Feed the lexicon to HTK [HTK(2002)]

Building a basic ASR system

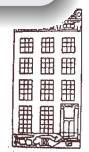
Generate prompts and record utterances

- Use task grammar to generate random prompts
- Record as many users as possible reading the prompts
- Better, subjects repeat synthesized (TTS) prompts
- Transcribe all prompts and all sentences in the corpus

Building a basic ASR system

Training

- Transcribe and (feature) encode utterances
- Feed as much speech as possible to the HTK training
- Kinyarwanda uses 3 male and 3 female speakers, 150 sentences
- Words were hand-labeled
- Dutch uses 1000 labeled sentences from the IFAcorpus (4 male, 4 female speakers)
- Dutch recorded 150 task sentences from 4 male speakers (total 600)
- Recorded utterances were transcribed automatically
- Put all files in correct format and fire up HTK training [HTK(2002)]



ASR evaluation: Kinyarwanda

Subject	Words correct	Substitution errors	Percentage
Subject 1	9	1	90%
Subject 2	8	2	80%
Subject 3	8	2	80%
Subject 4	8	2	80%

Live data recognition results

- 4 New subjects
- Read out all 10 numbers
- HTK self-test results (*not* live):
- Sentence Recognition Rate: 92.00% (N=50)
- Word Recognition Rate: 94.87% (N=156)

ASR evaluation: Dutch

TRAINED ON	TESTED ON
IFA + Domain	Domain, training speakers
IFA + Domain	Domain, 'unknown' speaker
IFA + Domain	New sentences, training speakers
IFA + Domain	New sentences, new speaker

Testing procedures

- Two corpora: IFA corpus and Domain corpus
- Testing using randomly selected sentences
- Test set not used during training

ASR evaluation: Dutch

Left Out %	WORD RECOGNITION (%)	SENTENCE RECOGNITION (%)
10	99.71	91.38
20	99.46	92.31
50	99.67	89.93
80	99.66	89.18

Testing on random sentences

- Leave out random sentences and train
- Test randomly selected sentences
- Smaller training set affects Sentence Recognition most

ASR evaluation: Dutch

LEFT OUT SPEAKER	WORD RECOGNITION (%)	SENTENCE RECOGNITION (%)
Tom	99.57	85.71
Markus	99.78	72.60
Ork	99.43	89.13
Frans	99.78	81.63

LEFT OUT PERCENTAGE	WORD RECOGNITION (%)	SENTENCE RECOGNITION (%)
12	99.41	92.86
25	99.80	90.57
50	99.84	89.35

Top: Testing on a new speaker, Bottom: Testing on new sentences

New speakers are worse than new sentences

- More speakers needed for independence
- Sentence recognition drops sharply
- New speaker *and* new sentences
  Recognition: Word - 99.57%, Sent - 84.35%

Conclusion

Simple TTS and ASR can be done in a few months

- Free tools are available
- People like it when their language is used
- Recording speech is the most laborous step
- More speech is better, as is more text
- Pronunciation dictionaries are crucial

Further Reading I

Frans Adriaans, Markus Heukelom, Marijn Koolen, Tom Lentz, Ork de Rooij, and Daan Vreeswijk.
Speech Technology Project 2004 Building an HMM Speech Recogniser for Dutch.
Technical report, Masters of AI, Faculty of Science, University of Amsterdam, 9 July 2004.
URL http://www.fon.hum.uva.nl/IFA-publications/Others/Other_papers.html.

Guillermo Aradilla, Jithendra Vepa, and Hervé Bourlard.
Improving speech recognition using a data-driven approach.
IDIAP-RR 66, IDIAP, Martigny, Switzerland, 2005.

Alan W. Black and Kevin A. Lenzo.
Festvox.
Web, 2003.
URL http://festvox.org/.
Festival speech synthesis.

P. Boersma.
Praat, a system for doing phonetics by computer.
Glot International, 5:341–345, 2001.
URL http://www.Praat.org/.

P. Boersma and D. Weenink.
Praat 4.2: doing phonetics by computer.
Computer program: http://www.Praat.org/, 2004.
URL http://www.Praat.org/.

Further Reading II

M. De Wachter, M. Matton, K. Demuynck, P. Wambacq, R. Cools, and D. Van Compernolle.

Template-Based Continuous Speech Recognition.
IEEE TRANSACTIONS ON AUDIO, SPEECH & LANGUAGE PROCESSING, 15(4):1377–1390, 2007.
ISSN 1558-7916.

Jelske Dijkstra, Louis C.W. Pols, and R.J.J.H. van Son.

Frisian TTS, an example of bootstrapping tts for minority languages.
In *Proceedings of the 5th ISCA Speech Synthesis Workshop - Pittsburgh*, pages 97–102, 2005.
URL `http://www.ssw5.org/`.

FSF.

GNU General Public License.
Web, June 1991.
URL `http://www.gnu.org/licenses/gpl.html`.

D. Gorter and R.J. Jonkman.

Taal yn Fryslân op 'e nij besjoen.
Fryske Akademy, Ljouwert, 1995.

James L. Hieronymus.

Ascii phonetic symbols for the world's languages: Worldbet.
Web, 1994.
URL `http://www.ling.ohio-state.edu/~edwards/worldbet.pdf`.

Further Reading III

HTK.

Hidden Markov Model Toolkit.
Web, December 2002.
URL http://htk.eng.cam.ac.uk.
Toolkit distribution.

Muhirwe Jackson.

Automatic speech recognition: Human computer interface for Kinyarwanda language.
Master's thesis, Faculty of computing and Information Technology, Makerere University, August 2005.
URL http://www.fon.hum.uva.nl/IFA-publications/Others/Other_papers.html.

MBROLA.

The MBROLA Project.
Web, 2005.
URL http://tcts.fpms.ac.be/synthesis/.
Synthesis.

Nextens.

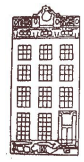
NeXTeNS: Open Source Text-to-Speech for Dutch.
Web, 2003.
URL http://nextens.uvt.nl/index.html.

Keiichi Tokuda, Heiga Zen, and Alan W. Black.

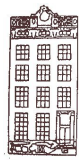
An HMM-based speech synthesis system applied to English, 2002.
URL http://www.cs.cmu.edu/~awb/papers/IEEE2002/hmmenglish.pdf.

Further Reading IV

S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland.
The HTK Book.
Cambridge University Engineering Department, December 2004.
URL `http://htk.eng.cam.ac.uk`.
Part of the HTK distribution.



Appendix A



Copyright License