How to build a Speech Synthesis System?

New Media & Language Technologies
Jozef Stefan International Postgraduate School
November 2005

Jerneja Žganec Gros
jerneja@alpineon.com
Speech Synthesis

• Concatenation of prerecorded speech units:
  • small vocabulary, simple syntax
  • limited application domains: naturally sounding output

• Text-to-speech synthesis:
  • automatic conversion of arbitrary text into speech using GTP
  • unrestricted application domain

• Concept-to-speech synthesis:
  • entry: semantic concepts
  • IVR, speech-to-speech translation
Prerecorded speech

database structure

DATE [December 29]
TYPE [maple]
LOC [Javorniki Vrh]
WEIGHT [6.7 kg]
REMARK [po plohi]

template

[Donosi na opazovalnicah DATE TYPE LOC WEIGHT REMARK]
Prerecorded speech

- message construction

- speech segment concatenation
  - continuous transitions
  - sentence intonation
  - nearly natural pronunciation

Donosi na opazovalnicah DEVETINDVAJSETEGA DECEMBRA. JAVORJEVA paša. JAVORNIKI VRH. PLUS ŠEST kilogramov SEDEMDESET dekagramov. PO PLOHI.
TTS approaches

- Modelling the human vocal tract (hvt):
  - mechanical & electrical models of the hvt...
  - formant frequencies: formant TTS...
- Concatenation methods:
  - PSOLA, MBROLA, unit-selection
  - diphones, poliphones...
- HMM-based methods
- this talk: corpus-driven approaches (AlpSynth)
TTS System Architecture

slovene text

rules, pronunciation dict.

intrinsic, extrinsic

tonemic accent, intonation

speech segment database

text preprocessing

grapheme-to-phoneme

duration modelling

F0 modelling

concatenation

slovene speech
Grapheme-to-Phoneme

- ASCII text
- text preprocessing
- pronunciation dictionary
- rules-stress position prediction
- grapheme-to-phoneme
- stress prediction
- GTP transcription
- sequence of punctuation marks and SAMPA transcriptions for words

Special symbols - elimination: \{, (, -, ,, ,$;
- Punctuation marks - usage determination
- Slovene: free stress position
- Rules for automatic stress prediction
- Production rules for automatic grapheme-to-phoneme transcription
Text Normalisation

• **alpha-numerical graphemes**
  - tokenization: merging into words
  - sequences of capital letters:
  
  title / acronym disambiguation
  
  `<AVTOBUSNA POSTAJA>`
  `<ZDA> <NATO>`

• **numerals**
  - cardinal / ordinal (1. torek -> prvi torek)

• **ideograms**
  - $, %, &, (,), +, =, /, <, >, ...
Text Normalisation

- **punctuation marks**
- **grammatical usage (e.g. full stop)**
  - followed by a space AND a capitalized word
  
  <Dopolnil jih je 78. Lepa starost.>
  
  - followed by 2 line feeds (end of paragraph)
  
  - not followed by a numeral or space

- **non-grammatical usage**
  
  - abbreviation stop (as.dr. Simon Dobrišek, dipl.ing.)
  
  - ordinal numeral (Ob 8. uri zvečer.)
  
  - decimal (Cena izdelka je 8.12 SIT.)
TTS System Architecture

slovene text

text preprocessing

rules, pronunciation dict.

grapheme-to-phoneme

intrinsic, extrinsic

duration modelling

tonemic accent, intonation

F0 modelling

speech segment database

concatenation

slovene speech
Graphemes-to-Phonemes

• search in the pronunciation dictionary
• coarticulation corrections
  (word boundaries)
• stress position prediction
  (out-of-dictionary words)
• grapheme-to-phoneme conversion,
  coarticulation corrections
  (out-of-dictionary words)
Pronunciation dictionary

1. Text database

<table>
<thead>
<tr>
<th>Word number</th>
<th>Word number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sveto pismo</td>
<td>152.212</td>
</tr>
<tr>
<td>Mikelj, Veliki Vaz</td>
<td>162.396</td>
</tr>
<tr>
<td>Cankar, Moje ′iv ljenje</td>
<td>26.916</td>
</tr>
<tr>
<td>Slovenec, izbor ′tankov</td>
<td>264.736</td>
</tr>
<tr>
<td>Moj Miklo, izbor ′tankov</td>
<td>150.194</td>
</tr>
<tr>
<td>Jurij, Deseti brat</td>
<td>65.860</td>
</tr>
<tr>
<td>total</td>
<td>822.314</td>
</tr>
</tbody>
</table>

2. 16,000 most frequent words cover 88.5% input text words

3. SAMPA transcription - manual corrections

<table>
<thead>
<tr>
<th>word number</th>
<th>word number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collocations</td>
<td>17</td>
</tr>
<tr>
<td>Numerals</td>
<td>234</td>
</tr>
<tr>
<td>Words of foreign origin</td>
<td>304</td>
</tr>
<tr>
<td>Acronyms</td>
<td>92</td>
</tr>
<tr>
<td>Proper names</td>
<td>929</td>
</tr>
<tr>
<td>Other frequent words</td>
<td>15,470</td>
</tr>
<tr>
<td>Total</td>
<td>16,215</td>
</tr>
</tbody>
</table>

Cumulative probability [%]

Number of most frequent words and their cumulative probability
Grapheme-to-Phoneme Rules

• **standard words rule set**

• **169 context-sensitive rules**

<table>
<thead>
<tr>
<th>Left context</th>
<th>Grapheme string</th>
<th>Right context</th>
<th>Phonetic transcr.</th>
<th>Example</th>
<th>Rule explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$</td>
<td>er</td>
<td></td>
<td>[@r]</td>
<td>Gaber</td>
<td>@ occurs after each -r not followed by a vowel (Toporisic91, p.49)</td>
</tr>
<tr>
<td>=</td>
<td>m</td>
<td>f</td>
<td>[F]</td>
<td>Simfonija</td>
<td>&lt;m&gt; in front of &lt;f&gt; and &lt;v&gt; is pronounced as a labiodental (Pravopis90, p. 145)</td>
</tr>
</tbody>
</table>

• **names rule set**
TTS System Architecture

- Slovene text
- Rules, pronunciation dict.
- Intrinsic, extrinsic
- Tonemic accent, intonation
- Speech segment database
- Text preprocessing
- Grapheme-to-phoneme
- Duration modelling
- F0 modelling
- Concatenation
- Slovene speech
Duration Modelling

- sequential rule systems (Klatt 73, Van Santen 93)
- neural networks (Campbell 90)
- stochastic modelling (Traber 93), decision trees (Riedi 95), hmms (2000->…)

- two-level approach (Epitropakis 93)
  - **intrinsic** duration modelling
  - **extrinsic** duration modelling
  - adaptation of intrinsic phone duration to extrinsic word duration (Gros 97)
Intrinsic Duration

- phone identity, phone type: C or V
- syllable type: open or closed
- tonic, pretonic, posttonic
- position within the word: initial, medium, final
- phonetic context: CC, VCV

Measurements:

- logatoms in neutral intonation position
Phone Duration

Pair-wise analysis: normal rate - slow rate. Normalised mean duration difference for pairs of phone realisations in the phoneme group context.
Extrinsic Duration

- number of syllables
- word position: phrase initial, medium, final
- requested speaking rate: from slow to normal and fast
- syllable position in a word: initial, medium, final

**Measurements:**
- continuous speech - slow, normal, fast
- duration units!
Syllable Duration

Articulation rate in number of syllables per second is shown for different word positions within a phrase.
Intrinsic to Extrinsic Dur.

- **Curves** $a_i$: linear interpolation between average phone duration measurements at different speaking rates.

- **Curves** $b_i$: horizontal translation of $a_i$ in a way that $b_i$ equals the intrinsic phone duration $t_i$ at normal speaking rate.

- **Equation**: $t_{je} = t_n + \frac{t js - t in}{t n - t p} (t_e - t n)$, $j = 1, 2$.

- **Curve** $c$: sum of $b_i$ over all phones; extrinsic word duration $t_e$ occurs at the speaking rate $x_e$. 
Duration Prediction - Eval.

**Test Base**

<table>
<thead>
<tr>
<th>Speech Rate</th>
<th>No. of Sentences</th>
<th>No. of Words</th>
<th>No. of Phones</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal speech rate</td>
<td>172</td>
<td>1400</td>
<td>5433</td>
</tr>
<tr>
<td>Fast rate</td>
<td>49</td>
<td>607</td>
<td>2351</td>
</tr>
<tr>
<td>Slow rate</td>
<td>60</td>
<td>800</td>
<td>2900</td>
</tr>
</tbody>
</table>

**Statistical Duration Difference Evaluation Between Phone Pairs in Natural and Synthetic Speech**

<table>
<thead>
<tr>
<th></th>
<th>Translation</th>
<th>Proportion</th>
<th>Natural Speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean absolute difference [ms]</td>
<td>10.97</td>
<td>30.20</td>
<td>5.3</td>
</tr>
<tr>
<td>Mean absolute diff. [ms] (stressed vowels)</td>
<td>6.89</td>
<td>33.67</td>
<td></td>
</tr>
<tr>
<td>Standard deviation [ms]</td>
<td>15.24</td>
<td>26.41</td>
<td>8.2</td>
</tr>
<tr>
<td>Standard deviation [ms] (stressed vowels)</td>
<td>13.18</td>
<td>28.07</td>
<td></td>
</tr>
</tbody>
</table>
Duration Prediction - Eval.

- phone duration values taken from natural speech
- phone duration values predicted by the 2-level approach

Duration modelling test - results

- preference for synthetic speech with natural dur.
- preference for synthetic speech with modelled dur.
- no difference perceived between the two versions

- 20 test subjects, different professional backgrounds
TTS System Architecture

- slovene text
- text preprocessing
- grapheme-to-phoneme
- duration modelling
- F0 modelling
- concatenation
- slovene speech

- rules, pronunciation dict.
- intrinsic, extrinsic
- tonemic accent, intonation
- speech segment database
F0 Modelling

- initial F0 values
- jump
- jump restrictions
- interpolation
- minor random adjustment

Typical F0 patterns (tonemes):
- barytone acute
- ocsytone acute
- 2-syllabic baritone circumflex
- 3-syllabic baritone circumflex
- ocsytone circumflex

- intrinsic pitch frequency
- syllable position: initial/final/mid
- syllable structure: open, closed
- tonic/pretonic/posttonic syllable

Sentence intonation
TTS System Architecture

- Slovene text
- Text preprocessing
- Grapheme-to-phoneme
- Duration modelling
- F0 modelling
- Concatenation
- Slovene speech

- Rules, pronunciation dict.
- Intrinsic, extrinsic
- Tonemic accent, intonation
- Speech segment database
Speech segment concatenation

4 corpus-driven text-to-speech synthesis

4 speech corpus:

- text selection
  - phonetic transcription of the source text corpus
  - phone frequency analysis
  - algorithm for optimal sentence set selection

- recording

- segmentation and labelling
Corpus-driven TTS

- speech corpus
- optimal speech segment selection (dynamic programming)
- speech segment concatenation and prosodic modifications (TD-PSOLA, MBROLA)
Corpus – elemental units

- allophones
- words
- diphones
- phrases….
- poliphones

longer segments:
- larger corpus
- more natural speech
Speech corpus design

- **text selection**: input reference corpus to resulting text corpus
  - phonetic transcription of the reference text corpus
  - frequency analysis of allophone strings
  - AlpSynth sentence selection method

- **recording**

- **segmentation and labelling**
  - initial automatic segmentation
  - manual fine segmentation
Text corpus: phonetic analysis

- grapheme-to-phoneme transcription of the initial reference text corpus

- frequency analysis of allophone strings:
  - allophones
  - diphones
  - triphones
  - quadphones
Sentence set selection

Allophone frequencies in the reference corpus

Allophone frequencies in the phonetic transcription of the reference text corpus
Triphone string frequencies

number of triphone occurrences in the reference corpus

all triphone occurrences

triphones

triphone frequencies in the phonetic transcription of the reference text corpus
Sentence set selection

- **goal**
  - compact resulting sentence corpus containing all predefined frequent allophone sequences

- **method**
  - cost evaluation for all sentences
  - cost normalization (to sentence length)
  - ranking and selection of evaluated sentences
Sentence set selection

features:

– initial reference text corpus (200,000 sentences)
– resulting compact text corpus (297 sentences)
– rich with different allophone sequences
  • 1,132 different diphones
  • 17,784 different triphones
  • 120,425 different quadphones
  • average sentence length: 34.4 allophones oz. 6 words
Recording

male speaker, laboratory conditions

<table>
<thead>
<tr>
<th>corpus size:</th>
<th>duration</th>
<th>number of words</th>
<th>number of phones</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>all words</td>
<td>different words</td>
</tr>
<tr>
<td>natural speech</td>
<td></td>
<td>all words</td>
<td>different words</td>
</tr>
<tr>
<td>A - recorded natural speech</td>
<td>3622 s</td>
<td>1814</td>
<td>1354</td>
</tr>
<tr>
<td>B - complete logatom corpus</td>
<td>1596 s</td>
<td>2837</td>
<td>2837</td>
</tr>
<tr>
<td>logatom corpus (no diphtongs)</td>
<td>508 s</td>
<td>1169</td>
<td>1169</td>
</tr>
<tr>
<td>logatom corpus (diphtongs only)</td>
<td>1088 s</td>
<td>1668</td>
<td>1668</td>
</tr>
<tr>
<td>C - complete TTS speech corpus (A+B)</td>
<td>5218 s</td>
<td>4651</td>
<td>4191</td>
</tr>
</tbody>
</table>
Segmentation and labeling

Phone segmentation:
- initial: automatic (HMM)
- fine: manual - SIGMARK©

Pitch marking:
- fine pitch marking: automatic - SIGMARK©
Automatic Labelling

- purpose:
  - basic phonetic research
  - initialisation for the stochastic speech recogniser

- approaches:
  - HMM
  - DTW alignment of natural and synthetic speech

- speech synthesis:
  - diphone inventory

- feature vector:
  - loudness, 11 mel-cepstrum coefficients
Automatic Labelling

Segmented and labelled synthesised speech signal
Automatic Labelling

Average frame match between manual and automatic segmentation

<table>
<thead>
<tr>
<th>group</th>
<th>frames</th>
<th>hmm</th>
<th>synt</th>
<th>diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>01F</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>vowels</td>
<td>25237</td>
<td>87.2%</td>
<td>84.1%</td>
<td>-3.1%</td>
</tr>
<tr>
<td>sonorants</td>
<td>10452</td>
<td>68.8%</td>
<td>73.4%</td>
<td>+4.6%</td>
</tr>
<tr>
<td>nonsonorant</td>
<td>all</td>
<td>16538</td>
<td>88.1%</td>
<td>93.1%</td>
</tr>
<tr>
<td></td>
<td>fricatives</td>
<td>5677</td>
<td>88.8%</td>
<td>93.9%</td>
</tr>
<tr>
<td></td>
<td>plosives &amp; affricates</td>
<td>10861</td>
<td>87.7%</td>
<td>92.6%</td>
</tr>
<tr>
<td></td>
<td>all</td>
<td>52227</td>
<td>83.8%</td>
<td>84.9%</td>
</tr>
<tr>
<td>01M</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>vowels</td>
<td>21971</td>
<td>83.9%</td>
<td>81.6%</td>
<td>-2.3%</td>
</tr>
<tr>
<td>sonorants</td>
<td>10317</td>
<td>65.9%</td>
<td>75.8%</td>
<td>+9.9%</td>
</tr>
<tr>
<td>nonsonorant</td>
<td>all</td>
<td>13623</td>
<td>85.3%</td>
<td>92.7%</td>
</tr>
<tr>
<td></td>
<td>fricatives</td>
<td>4659</td>
<td>84.0%</td>
<td>92.2%</td>
</tr>
<tr>
<td></td>
<td>plosives &amp; affricates</td>
<td>8964</td>
<td>86.0%</td>
<td>93.0%</td>
</tr>
<tr>
<td></td>
<td>all</td>
<td>45911</td>
<td>80.2%</td>
<td>83.6%</td>
</tr>
</tbody>
</table>
Plans for further work

- Reduction of spectral discontinuities
- Optimization of the speech segment selection procedure
- Selection of optimal intra-segment concatenation locations
- Further upgrades of the speech corpus
Evaluating TTS Systems

Jekosch93, Pols94, JEIDA95, Klaus03, ITU-T Recs

First experiment
  – intelligibility
  – naturalness

Second experiment
  – ITU-T Rec. P.81
  – ITU-T Rec. P.85
Text Selection

Text types:

- newspaper text (daily newspaper, 264,763 words)
- The Bible (152,212 words)
- SUS (semantically unpredictable sentences)
  - basic pattern structures: Subject - Verb – Adverbial,
    Subject – Transitive Verb - Object, etc.
  - *Hrast gleda morje*
  - word lists from the MULTEXT-EAST lexicon
    (morpho-syntactic descriptions)

Text selection methods:

- 4 text selection methods as proposed by LDC and COCOSDA
Text Selection Methods

1. Random selection

2. Minimum word frequency
   - determine number of occurrences (frequency) of each word in the text corpus
   - for each sentence, determine the frequency of the least frequent word
   - sort sentences in descending order by least frequent word frequency
   - randomly select from the top 1, 5, or 10% of this sorted list
Text Selection Methods

## Overall word frequency
- determine number of occurrences (frequency) of each word in the corpus
- for each sentence, add the log frequencies of all its words
- sort sentences in descending order by log frequency sum
- randomly select from the top 1, 5, or 10% of this sorted list

## Overall trigram frequency based selection
Design of the experiments

- Laboratory conditions
- 2 sessions, preliminary training session
- Various evaluators
- Questionnaire

<table>
<thead>
<tr>
<th>Koda poslušalca</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>IME IN PRIIMEK</th>
<th>ženski</th>
<th>moški</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPOL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>STAROST</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NARODNOST</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MATERIN JEZIK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IZOBRAZBA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MOREBITNE SLUŠNE MOTNJE</td>
<td>srednja</td>
<td>višja</td>
</tr>
<tr>
<td>STE ŽE KDAJ PREJ SLIŠALI TA SINTETIZATOR</td>
<td>da</td>
<td>ne</td>
</tr>
<tr>
<td></td>
<td>da</td>
<td>ne</td>
</tr>
</tbody>
</table>
Experiment

- TTS system
  - ITU-T Recommendations
  - 21 evaluators

- acceptability of the synthetic speech for the application
- naturalness of pronunciation
- subjective impressions of the synthetic speech
Acceptability

ITU-T Recommendation P.85
(a method for subjective performance assessment of speech voice output devices)

Application domain - automatic information retrieval
(for comparison with the test of the S5 TTS system – Gros97)

Message templates

CARRIER, flight number FLIGHT_NO, arriving from DEP_LOC, is about to land at ARR_LOC at ARR_TIME.

Adria Airways, flight number JP743, arriving from Frankfurt, is about to land in Ljubljana at 13:30.
Do you think this TTS system could be used in an automatic information dialog system for airline timetable retrieval?

- YES
- NO

Comments:
Naturalness

ITU-T Recommendation P.81
(Telephone quality subjective transmission tests - Modulated noise reference unit)

Voice sources
- corrupted natural speech (SNR 5dB, 10dB, 15dB, 30dB)
- speech synthesiser

MOS opinion scales
- overall impression
- listening effort
- comprehension problems
- articulation
- voice pleasantness

![Pie chart showing MOS opinion scales]

- better than natural speech corrupted with noise (10dB) 79%
- worse than natural speech corrupted with noise (5dB) 14%
- better than natural speech corrupted with noise (5dB) and (10dB) 7%
## Subjective impressions

### ITU-T Recommendations P.80 and P.85

"Methods for subjective determination of transmission quality"

“A method for subjective performance assessment of the quality of speech voice output devices”

<table>
<thead>
<tr>
<th>MOS scale</th>
<th>Overall impression</th>
<th>Comprehension problems</th>
<th>Articulation</th>
<th>Speech rate</th>
<th>Voice pleasantness</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>excellent</td>
<td>never</td>
<td>yes, very clear</td>
<td>much faster than preferred</td>
<td>very pleasant</td>
</tr>
<tr>
<td>4</td>
<td>good</td>
<td>rarely</td>
<td>yes, clear enough</td>
<td>faster than preferred</td>
<td>pleasant</td>
</tr>
<tr>
<td>3</td>
<td>fair</td>
<td>occasionally</td>
<td>fairly clear</td>
<td>preferred</td>
<td>fair</td>
</tr>
<tr>
<td>2</td>
<td>poor</td>
<td>often</td>
<td>no, not very clear</td>
<td>slower than preferred</td>
<td>unpleasant</td>
</tr>
<tr>
<td>1</td>
<td>bad</td>
<td>all the time</td>
<td>no, not at all</td>
<td>much slower than preferred</td>
<td>very unpleasant</td>
</tr>
</tbody>
</table>
Subjective impressions

- MOS rating scales:
  - overall impression, listening effort, comprehension problems, articulation, pronunciation, speech rate and voice pleasantness
- overall quality of the synthetic speech
- evaluation of individual components of the TTS system:
  - grapheme-to-phoneme: pronunciation dictionary
  - prosody modeling:
    - tonemic accent patterns
    - segment duration prediction methods
Subjective impressions

Segment duration prediction evaluation:
- segment duration of the synthetic speech
  - taken from natural speech
  - automatically predicted by the two-level approach (Gros et al, 1997)

![Bar chart showing preference for natural and synthetic duration of speech at different speaking rates (normal, slow, fast). The chart indicates a preference for natural duration in the 'normal' and 'slow' speaking rate conditions and for synthetic duration in the 'fast' speaking rate condition. There is no difference perceived between the two versions.]

The chart illustrates that the percentage of evaluators preferring natural duration increases as the speaking rate decreases, with the highest percentage in the 'normal' speaking rate condition. Conversely, the percentage preferring synthetic duration is highest in the 'fast' speaking rate condition. There is no preference difference perceived between the two versions in the middle (slow) speaking rate.
Conclusion

- Slovenian TTS system performance evaluation
- pleasant, quite natural speech,
  sufficiently rapid, not overarticulated
- further work: prosody, concatenation,
  lexical stress assignment
- Slovenian TTS: demo applications