TTS: What for?

- Telephone-based applications ($$$)
  - Integrated messaging (fax, email, answering machine)
  - Personal telephone attendant
  - Voice access to databases
    (70% of calls require very little interactivity):
      - Price lists, Cultural events, Weather report, ...

1983: FPMs / TCTS
Long-Term Research in Speech Processing
ASR (HIMARNNET)
TTS (MBROLA)

1997: BABEL Technologies
Speech Products

2000: MULTITEL ASBL
Generic Research

→ 50 people on speech proc.
TTS: What for?

- Multimedia
  - CDRoms
  - Talking books
  - Interactive games

TTS + ASR = DIALOG

- Replace imbricated menus
- Interactive kiosk applications
  (supermarkets, train stations, airports, etc)
- PDAs, Pocket PCs
  (everywhere screen is a problem)
- Monitoring in the industry (stocks, maintenance)

TTS: What for?

- Help to the disabled
  - Speech impairment
  - Sight impairment
    - Artificial voice
    - Automatic reading of electronic documents
    - Automatic reading of paper documents (with OCR)

TTS: What for?

- Man-Machine Communication
  - 20 years ago
  - Now (Lucille’s voice)

TTS: What for?

- Fundamental research
Contents

- TTS: What for?
- Challenges
  - Technologies
    - DSP
    - NLP
  - Currently Available Systems
    - Non-commercial
    - Commercial
- Conclusion: A Bright (Corpus-based) Future?

Challenge 1: Meaning

- Speech is a signal
- Signals carry information (=unpredictable data) from source to receiver
- Complexity of signals = f(complexity of source/receiver), and vice-versa
  - Communication signals
  - Images
  - Biological signals
  - Speech: produced, perceived, and understood by the most complex of all machines

Challenge 2: Naturalness

- We are very sensitive to naturalness

Challenge 3: Coverage

“These speech systems provide excellent examples for the study of complex systems, since they raise fundamental issues in system partitioning, choice of descriptive units, representational techniques, levels of abstraction, formalisms for knowledge representation, the expression of interacting constraints, techniques of modularity and hierarchy, techniques for characterizing the degree of belief in evidence, subjective techniques for the measurement of stimulus quality, naturalness and preference, the automatic determination of equivalence classes, adaptive model parameterization, tradeoffs between declarative and procedural representations, system architectures, and the exploitation of contemporary technology to produce real-time performance with acceptable cost.”

(Allen, 1985)
**Automatic phonetization**

- Dictionary look-up?

To be or not to be, that is the question.

`[210]
\text{t} \ 40
U \ 55 \ 0 \ 173 \ 75 \ 173
b \ 80 \ 10 \ 160
i: \ 198 \ 5 \ 173 \ 75 \ 235

**Automatic phonetization**

- More complex than that!

<table>
<thead>
<tr>
<th>Problem</th>
<th>Example</th>
<th>Level</th>
<th>Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assimilation</td>
<td>nasality or sonority assimilation, vocalic</td>
<td>word/sent</td>
<td>reading style, pronunciation of neighbors</td>
</tr>
<tr>
<td>Assimilation</td>
<td>harmonization</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homophones</td>
<td>the, record, contrast, read, est, couvent,</td>
<td>word</td>
<td>part-of-speech, meaning (rare)</td>
</tr>
<tr>
<td></td>
<td>portions, etc.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schwa deletion</td>
<td>table rouge, je ne te le rediral pas</td>
<td>sentence</td>
<td>syntactic articulation, pronunciation of neighbors, speaking style</td>
</tr>
<tr>
<td>Phonetic</td>
<td>très utile, deux à deux, plat exquis</td>
<td>sentence</td>
<td></td>
</tr>
<tr>
<td>Liaisons</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>New words</td>
<td>proopiomelanortin</td>
<td>word</td>
<td>spelling analogy</td>
</tr>
<tr>
<td>Proper names</td>
<td>your name here ...</td>
<td>word</td>
<td>morphology, analogy</td>
</tr>
</tbody>
</table>

**Intonation**

- Why ups and downs?
  - Stress (word level) ➔ Accent (phrase level) (Phonetization↑)
- Modify slightly ➔ unnatural!
**Intonation**

(a,b) Focus   (c) Finality/continuity
(d) Grouping, using phrase-level accent

**The term prosody refers to certain properties of the speech signal.**

- phonetics
  - **Phoneme Duration**
    - Not constant
    - Not fixed for a given phoneme
    - Linked to intonation
      (longer on accented syllables)

- **Coarticulation!!!**

**Intonation/Duration**

- **Prosody relies on each and every level of linguistic competence of the reader:**
  - **syntax**: Mainly
    - *‘Twas brillig, and the slithy toves Did gyre and gimble in the wabe
      All mims were the borogroves, And the mome raths outgrabe.*
    - Lewis Carroll, Jabberwocky
  - **semantics**:
    - *Time flies like an arrow*
  - **pragmatics**: Personal implication of the reader in his reading
    - *‘John did not come to New York by plane’*
    - Emphasis, contrastive stress, speaking style, ...

- **Acceptable neutral prosody?**
Coarticulation !!!

- We don’t even hear the difference! : we think we pronounce twice the same sound.
- We actually have been trained not to hear it, because it is not distinctive in French (cfr tones in Chinese: very hard for a European)
- We mean twice the same phoneme, but pronounce different sounds

This can be generalized: each phoneme corresponds to a continuum of speech sounds, depending on the phonetic context

Challenges : a summary

Intelligible – Natural – Cost effective

- Accurate automatic phonetization (=dictionary look-up)
- Prosody generation (i.e., intonation and phoneme durations) must be "coherent": easy to produce unnatural prosody
- Synthesis of phoneme sequences with corresponding prosody
  - Coarticulation!
  - Segmental quality should be maintained after pitch and duration modification
- Engineering
  - Low design and maintenance cost
  - Low computational and memory cost
  - Easy adaptation to other languages

Coarticulation !!!

- Synthesis : be able to mimic coarticulation! 😊
- (Recognition : be able to overcome it!)

Contents

- Challenges
- Currently Used Technologies
  - NLP
  - DSP
- Currently Available
  - Non-commercial
  - Commercial
- Applications
- Conclusion : A Bright (Corpus-based) Future?
Among the many usable and potentially useful technologies, only some are actually used:

- Genericity (CARTs and n-grams do a lot for NLP)
- Simplicity
  - of development (ex: MBROLA vs PSOLA))
  - of use (ex: n-grams vs formal grammars)
  - coverage (ex: out-of-vocabulary words, nonsyntactic sentences, synthesis of dyphonic voices, etc.)

Let's face it: the best does not always win...

1. Pre-processing

- Text segmentation into broad segmentation units
- From broad to final segmentation units
- Sentence and detection

<table>
<thead>
<tr>
<th>Dr. Jones lives at the corner of Jones Dr. and St. James St.</th>
</tr>
</thead>
</table>

- Dealing with abbreviations
  - German: 'tgl.' = 'taglich', 'tägliche', 'täglichem', 'täglichen', 'täglicher', 'tägliches'
  - Dr. Jones lives at the corner of Jones Dr. and St. James St.

- Recognizing acronyms
  - L.T.S.', 'UFO', FPMs, ...

- Processing numbers
  - '3.14', '2.16 pm', '13:26', '08.11.94', 'the 16th'

Simple Regular grammars do most of the job
(Lex – FSTs)

2. Morphological analysis

- Why?
  - Constrain the size of lexicons
  - Morphological features for syntactic processing
  - Morphologically related pronunciation
    - 'Habermass', 'fem' / 'demonstration' / 'hothouse', ...
  - Word-level stress in free stress languages

- How?
  - morphology = highly language dependent
    - (English verbs: four to eight forms; French verbs: 87 to 41 forms; compound words: hottootsinottoostoonningul)
  - Typically: regular rules, finite state automatons, organized in a language-dependent way

Increasingly: use of graphotactic trained systems
(ex: TNT http://www.coli.uni-sb.de/~thorsten/tnt/)

Or even: brute force: inflected dictionary
3. Contextual analysis

Dogs like to bark

3. Contextual analysis: n-grams

Example: "Dogs like to bark", using bi-grams (n=2)

\[ P(N_{pl}, \text{verb}, \text{Inf} \_ \text{marker}, \text{verb} | \text{dogs}, \text{like}, \text{to}, \text{bark}) \]

\[ = P(\text{dogs} | N_{pl}) \]
\[ P(\text{like} | \text{verb}) \]
\[ P(\text{to} | \text{Inf} \_ \text{marker}) \]
\[ P(\text{bark} | \text{verb}) \]
\[ P(N_{pl} | _) \]
\[ P(\text{verb} | N_{pl}) \]
\[ P(\text{Inf} \_ \text{marker} | \text{verb}) \]
\[ P(\text{verb} | \text{Inf} \_ \text{marker}) \]

\[ \prod_{i=1}^{N} P(w_i | t_i) \]
\[ \prod_{i=1}^{N} P(t_i | t_{i-1}, t_{i-2}, \ldots, t_{i-n}) \]

3. Contextual analysis: n-grams

Sentence \( W=(w_1, w_2, \ldots, w_N) \)

All possible sequences \( T=(t_1, t_2, \ldots, t_N) \)

- Best sequence of tags \( \hat{T} = \arg \max_T P(T | W) \)

\[ \hat{T} = \arg \max_T \frac{P(W | T) P(T)}{P(W)} \quad \text{(Bayes)} \]

\[ P(W | T) = \prod_{i=1}^{N} P(w_i | t_i) \]

\[ P(T) = \prod_{i=1}^{N} P(t_i | t_{i-1}, t_{i-2}, \ldots, t_{i-n}) \quad \text{(Strong hypothesis 1)} \]

4. Syntactic-Prosodic Phrasing

- Chinks’n chunks

A prosodic phrase = a sequence of chinks (≈ function words) followed by a sequence of chunks (≈ content words)

Example:

I asked them if they were going home to Idaho and they said yes and anticipated one more stop before getting home
4. Syntactic-Prosodic Phrasing

- CART tree

\[ st = \text{time to end of sentence} \]
\[ j3 = \text{tag of word on the right} \]
\[ j2 = \text{tag of word on the left} \]
\[ tr = \text{utterance rate (in words/second)} \]

Classification and Regression trees (CARTs)

<impure» set

Question which splits into best <purified» sets

«pure» sets

Classification and Regression trees (CARTs)

\[ \text{Entropy} = -(1/2 \times 1) - (1/2 \times 1) = 1 \text{ bit} \]

Total entropy after split = \( 0 + 0 = 0 \)
5. Automatic Phonetization

- Rule-based, usually formalized as in generative phonology:
  \[ a \rightarrow [b] / l \rightarrow r \]

- Example: \( s \rightarrow [s] \text{ or } [z] \)
  \[ s \rightarrow [s] / [éanti|hanti] \rightarrow [<V>] \]
  \[ s \rightarrow [s] / [anti|contre|impr.|prime|toume|ultra|psycho|télo] \rightarrow [i] \]
  \[ s \rightarrow [s] / [vrai] \rightarrow [em] \]
  \[ s \rightarrow [s] / [a]para|sinu] \rightarrow [e|o|y] \]
  \[ s \rightarrow [s] / [tran] \rightarrow [a|b|i] \]
  \[ s \rightarrow [s] / [<V>] \rightarrow [<V>] \]

6. Prosody generation: patterns

6. Prosody generation: tones

"..."
6. Prosody generation : tones

- From text to tones

<table>
<thead>
<tr>
<th></th>
<th>ce personnage grossier, te dérange-t-il</th>
</tr>
</thead>
<tbody>
<tr>
<td>WS</td>
<td>. . o . o . o .</td>
</tr>
<tr>
<td>SG</td>
<td>(. . -) (. . -) (. . -)</td>
</tr>
<tr>
<td>IG 1</td>
<td>(. . /LL) (. HH) (. . R/H)</td>
</tr>
<tr>
<td>IG 2</td>
<td>(. . -) (. HH) (. . R/H)</td>
</tr>
</tbody>
</table>

WS = word stress = lexical stress
SG = stress group
IG = intonation group

Phonetization

Synt.-Pros. Phrasing

(only one stressed syllable)

6. Prosody generation : tones

- From tones to F0 (Hz), corpus-based

- Large speech dba, with known intonation groups, F0 and tones
- For each target intonation group :
  - find a list of similar intonation groups (in terms of tones, number of syllables, position in sentence, etc.) in the dba
  - Select the sequence of intonation groups in the dba which :
    - best represents the target groups
    - AND minimizes intonative discontinuities

6. Prosody generation : tones

- From tones to F0 (Hz) by rule :

- If H,L, H+, L- = stressed syllables
- h,l = unstressed syllables)

Bad
Better

Copyright (c)2002 Faculté Polytechnique de Mons - T. Dutoit
## Summarizing NLP for TTS

<table>
<thead>
<tr>
<th></th>
<th>EXPERT-BASED</th>
<th>CORPUS-BASED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preprocessing</td>
<td>Regular Rules</td>
<td></td>
</tr>
<tr>
<td>Morphological analysis</td>
<td>Regular Rules</td>
<td></td>
</tr>
<tr>
<td>Contextual analysis</td>
<td></td>
<td>Bi- and trigrams</td>
</tr>
<tr>
<td>Syntactic-prosodic parsing</td>
<td>Chinks’n chunks</td>
<td>CARTS</td>
</tr>
<tr>
<td>Phonetization</td>
<td>LTS rules</td>
<td>Tries</td>
</tr>
<tr>
<td>Tone assignment</td>
<td>Rules</td>
<td></td>
</tr>
<tr>
<td>Duration generation</td>
<td></td>
<td>CARTs</td>
</tr>
<tr>
<td>Intonation generation</td>
<td>Rule-based tone-to-F0</td>
<td>Corpus-based tone-to-F0</td>
</tr>
</tbody>
</table>