An Overview of Speech Based Interfaces

PART I : Fundamentals of Speech Technologies

From the basics to real-life implementations of speech technologies

Dr. Jean Hennebert, HES-SO
Jean Hennebert CV

- **93**: MSc Engineering, Elec & Telecom, Mons, Belgium
- **98**: PhD, EPF Lausanne
- **96-97**: Visiting researcher, ICSI, Berkeley, USA
- **98-00**: UBS Zurich, System Architect
- **00-04**: UbiCall Communications, Dir. of Technology, USA & Belgium
- **04-now**: Lecturer and Senior Researcher, University of Fribourg
- **07-now**: Prof. Business Information Systems HES-SO, Sierre
Overview

PART I  Fundamentals of Speech Technologies

• Introduction
  – Speech as a source of information
  – Speech production process
  – Speech signal analysis

• Overview of Speech Technologies
  – Audio Coding
  – Speaker Recognition
  – Text to Speech Systems
  – Speech Recognition

PART II  Applications and Business Cases

• Text-to-speech providers and applications
• Telephony Dialog Systems
• Automatic indexation of audio and video documents
• Dictation Systems – Example in the medical domain
• Mobile Services – translation, car systems

• Conclusions - What’s next?
Introduction
- Speech as a source of information
- Speech production process
- Speech signal analysis

Overview of Speech Technologies
- Audio Coding
- Speaker Recognition
- Text to Speech Systems
- Speech Recognition

PART I
SPEECH TECHNOLOGIES
Speech as a source of information

- Speech signal
  - Speech Recognition
    - Textual content
      - “read my mail”
  - Language Recognition
    - Spoken language
      - English
  - Speaker Recognition
    - Speaker ID
      - John Smith
Speech Generation
Speech Generation – Vocal Tract
# International Phonetic Alphabet (1/2)

<table>
<thead>
<tr>
<th>Phonemes (Sons)</th>
<th>Écriture phonétique</th>
<th>Exemples</th>
</tr>
</thead>
<tbody>
<tr>
<td>/a/</td>
<td>barbe</td>
<td>bas</td>
</tr>
<tr>
<td>/e/</td>
<td>école</td>
<td>avec</td>
</tr>
<tr>
<td>/i/</td>
<td>livre</td>
<td>lion</td>
</tr>
<tr>
<td>/j/</td>
<td>il</td>
<td>fille</td>
</tr>
<tr>
<td>/y/</td>
<td>rue</td>
<td>payer</td>
</tr>
<tr>
<td>/u/</td>
<td>u (+ voyelle)</td>
<td>cuisiner</td>
</tr>
<tr>
<td>/wo/</td>
<td>ou</td>
<td>oin</td>
</tr>
<tr>
<td>/o/</td>
<td>ou</td>
<td>au (eau)</td>
</tr>
<tr>
<td>/e/</td>
<td>ou</td>
<td>en</td>
</tr>
<tr>
<td>/æ/</td>
<td>ou</td>
<td>en</td>
</tr>
<tr>
<td>/a:/</td>
<td>ou</td>
<td>en</td>
</tr>
<tr>
<td>/e:/</td>
<td>ou</td>
<td>en</td>
</tr>
<tr>
<td>/æ:/</td>
<td>ou</td>
<td>en</td>
</tr>
<tr>
<td>/ɛ/</td>
<td>ou</td>
<td>en</td>
</tr>
<tr>
<td>/iː/</td>
<td>ou</td>
<td>en</td>
</tr>
<tr>
<td>/uː/</td>
<td>ou</td>
<td>en</td>
</tr>
<tr>
<td>/ʊː/</td>
<td>ou</td>
<td>en</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Graphies fréquentes (Écritures possibles)</th>
<th>Phonèmes (Sons)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a (+ consonnne)</td>
<td>/a/</td>
</tr>
<tr>
<td>e (+ voyelle)</td>
<td>/e/</td>
</tr>
<tr>
<td>i (+ voyelle)</td>
<td>/i/</td>
</tr>
<tr>
<td>y (consonne)</td>
<td>/y/</td>
</tr>
<tr>
<td>u (+ voyelle)</td>
<td>/u/</td>
</tr>
<tr>
<td>ou</td>
<td>/wo/</td>
</tr>
<tr>
<td>oin</td>
<td>/o/</td>
</tr>
<tr>
<td>au (eau)</td>
<td>/æ/</td>
</tr>
<tr>
<td>en</td>
<td>/ɛ/</td>
</tr>
<tr>
<td>en</td>
<td>/iː/</td>
</tr>
<tr>
<td>en</td>
<td>/uː/</td>
</tr>
<tr>
<td>en</td>
<td>/ʊː/</td>
</tr>
</tbody>
</table>

This table provides examples of phonetic symbols for various sounds and graphemes in the International Phonetic Alphabet (IPA). Each row details a specific phoneme and its corresponding international phonetic notation, followed by examples illustrating how these sounds are used in words.
### International Phonetic Alphabet (2/2)

<table>
<thead>
<tr>
<th>Consonnes (sons)</th>
<th>écriture phonétique</th>
<th>exemples</th>
</tr>
</thead>
<tbody>
<tr>
<td>[p]</td>
<td>/p/</td>
<td>pain, bebe</td>
</tr>
<tr>
<td>[b]</td>
<td>/b/</td>
<td>terre, dinaniche</td>
</tr>
<tr>
<td>[t]</td>
<td>/t/</td>
<td>carte, quartier</td>
</tr>
<tr>
<td>[k]</td>
<td>/k/</td>
<td>gare, vague</td>
</tr>
<tr>
<td>[g]</td>
<td>/g/</td>
<td>femme, pharmacie</td>
</tr>
<tr>
<td>[f]</td>
<td>/f/</td>
<td>ville, classe</td>
</tr>
<tr>
<td>[v]</td>
<td>/v/</td>
<td>succe, plage</td>
</tr>
<tr>
<td>[s]</td>
<td>/s/</td>
<td>abyme, ton</td>
</tr>
<tr>
<td>[z]</td>
<td>/z/</td>
<td>raison, raison</td>
</tr>
<tr>
<td>[ç]</td>
<td>/ç/</td>
<td>chez, cheval</td>
</tr>
<tr>
<td>[ʒ]</td>
<td>/ʒ/</td>
<td>plage, gare</td>
</tr>
<tr>
<td>[l]</td>
<td>/l/</td>
<td>libre, radio</td>
</tr>
<tr>
<td>[r]</td>
<td>/ʁ/</td>
<td>matin, nageur</td>
</tr>
<tr>
<td>[m]</td>
<td>/m/</td>
<td>m</td>
</tr>
<tr>
<td>[n]</td>
<td>/n/</td>
<td>n</td>
</tr>
</tbody>
</table>
| [ŋ] | /ŋ/ | |}

### Graphies fréquentes (écritures possibles)

| P | t | qu | gu | f | ph | v | s | ç (+) | ch | l | r | m | n | gn |
|---|---|----|----|---|----|---|---|------|----|---|---|---|---|---|---|
| /p/ | /t/ | /k/ | /g/ | /f/ | /ph/ | /v/ | /s/ | /ç/ | /ch/ | /l/ | /r/ | /m/ | /n/ | /ŋ/ |
Figure 1.1: Speech signal of the word *accumulation* (a): waveform, (b) partial waveform, (c) narrowband spectrogram of (a), (d) power spectrum magnitude of (b).
Speech signal is highly variable

• Inter-speaker variability
  – Language / dialect
  – Physiology (for ex gender)
  – Education, social environment

• Intra-speaker variability
  – Age
  – Short-term: tiredness, emotion
  – Medium-term: sickness, drugs

• User cooperation
  – Professional speakers
  – Read speech vs spontaneous speech

• Recording conditions
  – Environmental noise, background conversation
  – Channel variabilities, recording devices
  – Bandwidth
PART I
SPEECH TECHNOLOGIES

Introduction
- Speech as a source of information
- Speech production process
- Speech signal analysis

Overview of Speech Technologies
- Preamble
- Audio Coding
- Speaker Recognition
- Text to Speech Systems
- Speech Recognition
Preamble

• Most of the speech technologies are cpu and memory intensive
• The “window” of feasibility of some applications have opened recently – in the past 10 years
• As most of the applications are “statistical” based, the evolution will continue over the next decade
Preamble – Evolution of hardware performances

The number of transistors double every 18 months
In 2004: « IBM will commercialize soon a portable device able to record a full life of speech. »
MScBA
Major in Management of Information Systems

Technologies

- Speech Recognition / Dialogs
- Speaker Recognition
- Speech Synthesis / Prompting
- Speech Coding

Applications

- Over-the-phone applications
- Desktop applications
- Embedded / mobile applications

14 milliard $ market (2001)
Speech coding

• Coding aims at **compressing** the signal
  – For **static** storage on a file
    • **MP3** technology for music
    • **Narrow band** technology for speech (300 Hz-3400 Hz)
  – For **dynamic** transmission
    • Point to point (codecs in your Natel!)
    • Streaming technologies
Speech synthesis and prompting

- “target text” → Full Recording
- “target text” → Concatenation
- “target text” → Text-to-Speech

- Fixed content, ex. games
- Lowly varying content, ex. CFF
- Highly varying content, ex. SMS or email reader

Speech Synthesis / Prompting
Speech Synthesis – TTS

- **Phoneme-based** 🎤
  - smaller footprint – runs in toys or low-end mobile phones – sometimes difficult to understand

- **Non-linear-units based** 🎤
  - larger footprint – needs quite a lot of memory – often server-based - good quality

- **DEMO**: This is a demonstration of a text to speech synthesis. Hey, the weather is pretty good in Helsinki.
Prompting

Pre-recording
- Voice talents are needed
- Concatenation is an option for structured data
- Musical content can be used
  - Corporate earcons
  - Event earcons
- Not dynamic excepted for simple concatenation
Speaker Recognition

Speaker Recognition Tasks

Identification

1:N

"Whose voice is this?"

Verification

1:2

"Is this the voice of Mr. Smith?"

Speaker Tracking

"When spoke Mr. Smith?"

Segmentation

"Who speaks? Who spoke when?"
Speaker Verification as a biometry

Who you are
Fingerprints, voice, iris, ...

What you have
Keys, card, CLI, ...

What you know
Account #, Passwords, ...

Speaker Recognition
Comparison with other biometrics

- Fingerprint, iris
- Speech
- Signature

Physical attributes: Rigid / passive

Performance attributes: Behavioral / dynamic

Speaker Recognition
What identifies a speaker?

- Physical properties
  - Vocal tract shape
  - Vocal cords length

- Behavioral characteristics
  - Speaking rate
  - Prosody

- Higher level information
  - Vocabulary selection
  - Grammatical constructions
  - Hesitation and filler sounds
  - Conversation context
Speech recognition by humans

Diagram showing the process of speech generation and recognition, with details on machine counterparts, message formulation, language code, neural transduction, and acoustic analysis.
Speech recognition by machines

Speech Endpointing

Feature Extraction

Feature Vectors

"I wanna go to dallas"

Speech Recognition

"I wanna go to dallas"

Language Interpretation

Meaning
Destination=Dallas

Dialog Management

TTS: "at what time would you like to fly to Dallas"
Classification of automatic speech recognition systems

- **Application control**
  - Speaker (in)dependent
  - Runs on PDA
  - Web navigation, ...
  - A couple hundred words

- **Dictation System**
  - Speaker dependent
  - Runs on laptop
  - Word processing, ...
  - Up to 50K words

- **Keyword recognition**
  - Word-print based
  - Runs on mobile phones
  - Voice dialing appl., ...
  - A few dozen words

- **Server-side ASR**
  - Speaker independent
  - Runs on big servers
  - Dialog based appl.
  - Up to 10K words
Conclusions

• Talking is a non-intrusive natural gesture
• Strong progress has been done over the past 30 years
• Many areas of applications
  – Dictation systems
  – Indexation - audio mining
  – Voice-activated applications
  – Over-the-phone dialog systems
• The next steps are
  – Multimodality
  – Multilinguality
Further Readings

End of PART I

Questions?