Mobile Voice Search

Mazin Gilbert, AT&T Labs Research
Alex Acero, Microsoft Research
Acknowledgements

AT&T

• Srinivas Bangalore
• Luciano Barbosa
• Diamantino Caseiro
• Vincent Goffin
• Narendra Gupta
• Junlan Feng
• Michael Johnston
• Andrej Ljolje
• Taniya Mishra
• The Speech Team

Microsoft

• XD Huang
• Ivan Tashev
• Ye-Yi Wang
• Geoff Zweig
• Patrick Nguyen
• Xiao Li
• Xiaolong Li
• Dan Bohus
• YC Ju
• The Tellme Team
Tutorial Outline

- Introduction and Historical prospective
- Advances in voice search technologies
- The mobile revolution
- Mobile voice applications
- Market dynamics and business models
- Technology challenges and the road ahead
Talking Machines: Birth of Spoken Language Technology

Joseph Faber
(1835)

Von Kempelen
(1791)

Homer Dudley
Bell Labs
(1939)
Historical Prospective
Evolution of Speech Services

1992
- Command & control
- Voice dialing
- Directed dialog
- Travel, finance
- DA

2000
- Natural language
- Dialog
- Speech mining
- E-commerce
- Dictation

2008
- Mobile voice search
- Multimodal/LBS services
- Problem solving
- Language translation
- Personal agents

“Traditional Voice Calls”

“Mobility-Game Changer”
411: Directory Assistance (before 2003)

- Operator Assisted DA
411: Directory Assistance

- **Paid 411:**
  - 1984 (AT&T Divesture), Pay per use
  - 2003. Tellme automates a fraction of 411 calls. First large scale deployments of voice search

- **Free 411:**
  - Ad supported
    - (800) FREE-411, 2005
  - Search Engines:
    - (800) BING-411, 2007
    - (800) GOOG-411, 2007
Voice Search in Smartphones

- The arrival of the Smartphones:
  - Better to see the results than hear them
  - Convenient to speak them instead of typing them
  - Send voice through data channel
  - Most people use 411 from their cellphones anyway…
Multimodal voice search (2008)
Tutorial Outline

- Introduction and Historical prospective
- Advances in voice search technologies
- The mobile revolution
- Mobile voice applications
- Market dynamics and business models
- Technology challenges and the road ahead
Mobile Voice Search

Types: “Lawyers open on Saturdays”

Speaks: “nearest pizza restaurant”

Plays “Romano Italian at …”

Natural Language Understanding → Information Retrieval

Speech Recognition

Speech Synthesis

GPS Profiles Business DB
Mobile Voice Search

Types:
“Lawyers open on Saturdays”

Speaks:
“nearest pizza restaurant”

Plays “Romano Italian at …”

Natural Language Understanding

Speech Recognition

Speech Synthesis

Information Retrieval

GPS Profiles Business DB
Google search for "search engines"

Approximately 97,000,000 results (0.13 seconds)

Sponsored link

**Search Engines**

- **Bing**
  Bing is a search engine that finds and organizes the answers you need so you can make faster, more informed decisions.
  - www.bing.com
  - Cached

- **AltaVista**
  AltaVista provides the most comprehensive search experience on the Web...
  - www.altavista.com
  - Cached

**News for search engines**

- Infraction Lands Big First Round For Public Records Search - 9 hours ago
- Privacy is a major issue for people search engines and other sites have been...
- Other search engines oriented around social media sites, such as Spock...
  - Wall Street Journal (blog)

**Web search engine** - Wikipedia, the free encyclopedia

A web search engine is designed to search for information on the World Wide Web. The search results are generally presented in a list of results and are...
  - en.wikipedia.org/wiki/Web_search_engine
Bing

Web

search engines

ALL RESULTS:

1-10 of 127,000,000 results · Advanced

Doggie Web Search
Doggie.com makes searching the Web easy, because it has all the best search engines pieced into one. Go Fetch!
www.doggie.com · Cached page

Web search engine - Wikipedia, the free encyclopedia
Those search engines which do not accept money for their search engine results make money by running search related ads alongside the regular search engine results.
en.wikipedia.org/wiki/Web_search_engine · Cached page

List of search engines - Wikipedia, the free encyclopedia
By content/topic · By information type · By model · Based on
This is a list of Wikipedia articles about search engines, including web search engines, selection-based search engines, metasearch engines, desktop search tools, and web...
en.wikipedia.org/wiki/List_of_search_engines · Cached page

AltaVista
AltaVista provides the most comprehensive search experience on the Web!
www.altavista.com · Cached page

Recommended Search Engines-The Library-University of California ...
Features of recommended search engines in table format...
http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/SearchEngines.html
lib.berkeley.edu/TeachingLib/Guides/Internet/SearchEngines.html · Cached page

WebCrawler Web Search
Offers a single source to search the Web, images, video, news from Google, Yahoo!, Bing, Ask and many more search engines.
www.webcrawler.com · Cached page
How do search engines work?

“search engines” → Search Engine → Web Index

10 Blue links
URL123
URL112
URL124
...

URL103
URL112
TF: Term Frequency

- **Term frequency**: relative frequency of term \( i \) in document \( j \)
  - \( D1 \): “Information retrieval is powerful”
  - \( D2 \): “Information is good”
  - \( D3 \): “Powerful is good”

<table>
<thead>
<tr>
<th>TERMS</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1=Information</td>
<td>0.25</td>
<td>0.33</td>
<td>0.0</td>
</tr>
<tr>
<td>T2=retrieval</td>
<td>0.25</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>T3=is</td>
<td>0.25</td>
<td>0.33</td>
<td>0.33</td>
</tr>
<tr>
<td>T4=good</td>
<td>0.0</td>
<td>0.33</td>
<td>0.33</td>
</tr>
<tr>
<td>T5=powerful</td>
<td>0.25</td>
<td>0.0</td>
<td>0.33</td>
</tr>
</tbody>
</table>
IDF: Inverse Document Frequency

- IDF is a heuristic that tells how discriminative each term is.
- \( IDF(i) = \log(M|M_i) \)
  - \( M \) is the total number of documents.
  - \( M_i \) is the number of documents that contain term \( i \).

<table>
<thead>
<tr>
<th>TERMS</th>
<th>IDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1=Information</td>
<td>( \log(3/2) = 0.176 )</td>
</tr>
<tr>
<td>T2=retrieval</td>
<td>( \log(3/1) = 0.477 )</td>
</tr>
<tr>
<td>T3=is</td>
<td>( \log(3/3) = 0.0 )</td>
</tr>
<tr>
<td>T4=good</td>
<td>( \log(3/2) = 0.176 )</td>
</tr>
<tr>
<td>T5=powerful</td>
<td>( \log(3/2) = 0.176 )</td>
</tr>
</tbody>
</table>

\( D1: \) "Information retrieval is powerful"

\( D2: \) "Information is good"

\( D3: \) "Powerful is good"
TF-IDF

- In web search, matrix is very sparse:
  - Store as a linked list of sorted TFIDF scores (more compact)
  - Single keyword is a row look up (fast!)
  - Two keywords requires intersecting two rows (still fast!)
  - Index storage requires lots of memory (several machines)

<table>
<thead>
<tr>
<th>TERMS</th>
<th>DOCUMENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D1</td>
</tr>
<tr>
<td>T1=Information</td>
<td>0.044</td>
</tr>
<tr>
<td>T2=retrieval</td>
<td>0.119</td>
</tr>
<tr>
<td>T3=is</td>
<td>0.0</td>
</tr>
<tr>
<td>T4=good</td>
<td>0.0</td>
</tr>
<tr>
<td>T5=powerful</td>
<td>0.044</td>
</tr>
</tbody>
</table>
### The inverted index

<table>
<thead>
<tr>
<th>TERMS</th>
<th>DOCUMENTS</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;information&quot;</td>
<td>&quot;Information retrieval is powerful&quot;</td>
<td>0.044</td>
<td>0.058</td>
<td>0.0</td>
</tr>
<tr>
<td>&quot;retrieval&quot;</td>
<td>&quot;Information is good&quot;</td>
<td>0.119</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>&quot;is&quot;</td>
<td>&quot;Powerful is good&quot;</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>&quot;good&quot;</td>
<td></td>
<td>0.0</td>
<td>0.058</td>
<td>0.058</td>
</tr>
<tr>
<td>&quot;powerful&quot;</td>
<td></td>
<td>0.044</td>
<td>0.0</td>
<td>0.058</td>
</tr>
</tbody>
</table>
• Similarity of a document $d_j$ with a given query $q$ can be measured by the vector cosine measure:

$$\cos(q, d_j) = \frac{\sum w_{ij}w_{iq}}{\sqrt{\sum |w_{ij}|^2} \sqrt{\sum |w_{iq}|^2}}$$

where $w_{ij}$ is the weight of term $i$ in document $j$ and $w_{iq}$ is the weight of term $i$ in query $q$
PageRank (Larry Page, 1998)

\[
PR(p_i; 0) = \frac{1}{N}
\]

\[
PR(p_i; t + 1) = \frac{1 - d}{N} + d \sum \frac{PR(p_i; t)}{L(p_i)}
\]

- Iterate
- In the limit is the solution to Eigenvalue problem
- Random surfer model
- \(d\): damping factor
Document Ranking

- Static Rank measures prior probability for each document
  - Browse Rank (documents browsed by users)
- Ranking uses a machine learned function of static rank, TFIDF, document/query pairs that have been clicked, etc.
- Hundreds of features!!!
From Search to Question/Answering

- Static Questions:
  - What color is the sky?
  - What is swine flu?

- Dynamic Questions:
  - What is the stock price of AT&T?
  - What is playing in the theater near me?

- Distinction not always clear
  - What are the names of the current Supreme Court chief justices?
  - How tall is this year’s Christmas tree at Rockefeller Center?

- Static or dynamic?
  - Depends on the temporal cutoff between static and dynamic
  - Dynamic: answers to which change more frequently than once a year.
Aggregation and Synthesis of Responses

World Wide Web Business Databases User Generated

FAQs Web Crawling/Parsing

Q/A Synthesis

Experts Generated Responses

Q/A Database
Caption/Summarization

**Definitions of caption (n)**

1. **description of illustration**: a short description or title accompanying an illustration in a printed text.
2. **movie or television subtitle**: a printed explanation in a motion picture or on television, especially a translation of dialogue accompanying a scene or an explanation preceding a scene.
3. **heading or subheading**: a heading or subheading in a document or article.

Synonyms: slogan, subtitle, title, description, legend, heading, header, footer.

**Wikipedia Manual of Style (captions)** - Wikipedia, the free encyclopedia

A caption, also known as a cutline, is text that appears below the image. Most captions draw attention to something in the image that is not obvious, such as its relevance to...

**What is caption?** - A Word Definition From The Webopedia Computer Encyclopedia

This page describes the term caption and lists other pages on the Web where you can find additional information.

www.webopedia.com/term/c/caption.html - Cached page - Mark as spam
I had a great dining experience at Michaels. The staff totally walked me through my order and I was so impressed with the presentation of the meal. Just fantastic! I had a blast! I was so impressed with the ambiance. The food and feel of the restaurant were very impressive, especially the atmosphere service and quality of the meal. Seared ahi tuna was one of my favorites. 

Michael’s on main in soquel not only has great cocktails but the food is fantastic! My lady friends and I had a blast last weekend. Thank you for showing us what we have been missing out on. You are just fantastic! Our favorite service is very realistic for the service and quality of the meal. 

The restaurant was empty when our party of 8 was seated and was sparsely populated during our stay. After bungling our cocktails a 20 min wait for the salads then 10 min more we served the entrees then 3 then minutes again for the last 2 meals. The salmon arrived mealy and over cooked as the halibut did. Four of us sub’s sauteed spinach for starch only one received the side. The others received a raw spinach. The salmon was mealy and over cooked as was the halibut. Four of us sub’s sauteed spinach for starch only one received the side. The others received a raw spinach leaves on their plate. When asked the wait person seemed confused. We ordered sides which arrived as we finished our meal. We ordered dessert again the wait person returned to announce they were out of another request. The check arrives and we discover a 20 50 gratuity added for the party of 8. Call this the last ha ha.

My fiancé and I just recently dined here after viewing some possible wedding locations. The staff was awesome the food was phenomenal the ambiance was beautiful and the overall experience lead us to decide to hold our wedding there. I highly recommend this restaurant to anyone staying in or passing through the Santa Cruz area.

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Features</th>
<th>Sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ambiance</td>
<td>decor</td>
<td></td>
</tr>
<tr>
<td>Service</td>
<td>staff</td>
<td></td>
</tr>
<tr>
<td>Food</td>
<td>drinks</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ahi tuna</td>
<td></td>
</tr>
<tr>
<td>Price</td>
<td>fish</td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Template based or NL based text generation
Mobile Voice Search

Types: “Lawyers open on Saturdays”

Speaks: “nearest pizza restaurant”

Plays “Romano Italian at …”

Natural Language Understanding

Information Retrieval

Speech Recognition

Speech Synthesis

GPS Profiles Business DB
ASR - The Big Picture!

Bayes Rule

\[
\hat{W} = \arg \max_w P(W | X) = \arg \max_w \frac{P(W)P(X | W)}{P(X)} = \arg \max_w P(W)P(X | W)
\]
**Goal:** Extract robust features (information) from the speech that are relevant for ASR.

**Method:** Spectral analysis through either a bank-of-filters or through Linear Predictive Coding followed by non-linearity and normalization.

**Result:** Signal compression where for each window of speech samples 30 or so features are extracted (64,000 b/s -> 5,200 b/s).

**Challenges:** Robustness to environment (office, airport, car), devices (speakerphones, cellphones), speakers (accents, dialect, style, speaking defects), packet loss, noise
Acoustic Modeling

**Goal:** Map acoustic features into distinct linguistic units.

**What is a HMM?** Statistical method for characterizing the spectral properties of speech by a parametric random process. A collection of HMMs is associated with a subword unit. HMMs are also assigned for modeling extraneous events.

**Advantages:** Powerful statistical method for dealing with a wide range of data and conditions, and is highly reliable for recognizing speech.
• Model adaptation valuable for personalization and on-line learning

• Maximum a Posteriori (MAP)
  – adds a prior for the parameters $\Phi$
    $$\hat{\Phi} = \arg \max_\Phi [p(\Phi | X)] = \arg \max_\Phi [p(X | \Phi)p(\Phi)]$$

• Maximum Likelihood Linear Regression (MLLR)
  – Transform mean vectors $\tilde{\mu}_{ik} = W_c \mu_{ik}$
  – Can have more than one MLLR transform
  – Speaker Adaptive Training (SAT) applies MLLR to training data as well
Goal:
Map legal phone sequences into words according to phonotactic (or letter-to-sound) rules. For example,

David     /d/ /ey/ /v/ /ih/ /d/

Multiple Pronunciation:
A word may have multiple pronunciations. For example

Data 0.8  /d/ /ae/ /t/ /ax/
Data 0.2  /d/ /ey/ /t/ /ax/

Challenges:
How do you generate a word lexicon automatically?
**Goal:**
Model “acceptable” spoken phrases, constrained by task syntax.

**Rule-based:**
Hand-crafted grammars. For example, context free grammar (CFG)

\[]($\text{ITINERARY} == \text{flying from } \text{\$CITY to } \text{\$CITY on } \text{\$DATE}$)

**Statistical:**
Compute estimate of word probabilities (N-gram, class-based, CFG).
For example

\[](flying from Newark to Boston tomorrow)
Statistical Ngrams

\[ P(W) = P(w_1, w_2, \ldots, w_n) = P(w_1)P(w_2 | w_1)P(w_3 | w_1, w_2) \cdots P(w_n | w_1, w_2, \ldots, w_{n-1}) = \prod_{i=1}^{n} P(w_i | w_1, w_2, \ldots, w_{i-1}) \]

- Trigrams Estimation

\[ P(w_i | w_{i-2}, w_{i-1}) = \frac{C(w_{i-2}, w_{i-1}, w_i)}{C(w_{i-2}, w_{i-1})} \]

Performance measured in terms of Perplexity

\[ PP(W) = 2^{H(W)} \]
Goal:
Combine information (probabilities) from the acoustic model, language model and word lexicon to find the “optimal” word sequence (highest probability).

Method:
Decoder searches through all possible recognition choices using a Viterbi decoding algorithm.

Challenge:
Efficient search through a large network space is computationally expensive for large vocabulary ASR. For the WSJ 20K vocabulary, this results in a network of $10^{22}$ bytes!
Confidence Scoring

Goal:
Identify possible recognition errors and out-of-vocabulary events to improve the performance of ASR, SLU and DM.

Method:
A confidence score based on a hypothesis likelihood ratio test is associated with each recognized word. For example:

- Label: credit please
- Recognized: credit fees
- Confidence: (0.9) (0.3)

Challenges:
Rejection of extraneous acoustic events (noise, background speech, door slams) without rejection of valid user input speech.
Machines are 5-50 times worse than humans on virtually any recognition task.
Voice Search: business listings
Zweig, Li, Nguyen 2007

Pizza Hut
... Microsoft Corp.

Index

SLM

TF-IDF

ASR

Microsoft

Microsoft Corporation
Bootstrapping Language Modeling for Business Search

- Build an n-gram from Phone book
  - Equally weighted entries
- Specify City/state to narrow down grammar
- Then add weights
  - From 411 operator released numbers or search engine clicks
- Add actual way people speak them:
  - Transcriptions from 411 calls
  - Text queries: requires spell correction!
  - Also help with business listing weights
- Generalize to unseen business listings
- Text normalization (i.e. 7-Eleven)
Database problems when building LM

• We use multiple databases of differing quality
• Need to reach consistency
• Misspellings
  – Restarant, Resterant, Restaraunt, Restruant
  – Italien, Itallian, Italion
  – Too many to enumerate and explicitly fix
• Context-sensitive homophones
  – “Peoples Bancorp” but “People’s United Bank”
  – “Kwick Signs” but “Quick Locks”
Automated Feedback Methods

- **Data addition**
  - What people click on & associated audio
  - Text searches from web

- **Personalization**
  - Per-person / user-profile language models
  - Per-person speaker-adaptive transforms
Click-Driven Automated Feedback

Language Model

Acoustic Model
Mobile Voice Search

Types:
“Lawyers open on Saturdays”

Speaks:
“nearest pizza restaurant”

Plays “Romano Italian at …”

Natural Language Understanding → Information Retrieval

Speech Recognition

Speech Synthesis

GPS Profiles Business DB
Voice over Data Channel

- Microphone
- Amplifier
- Analog/Digital Converter (aka Sampling)
- Codec
- IP Packets
- Data Network
Voice Sampling

• A-to-D
  – discrete samples of the waveform and represent each sample by some number of bits
  – A signal can be reconstructed if it is sampled at a minimum of twice the maximum frequency (Nyquist Theorem)

• Human speech
  – 300-3800 Hz
  – 8000 samples per second

Each sample is encoded into an 8-bit PCM code word (e.g. 01100101) => 8000 x 8 bit/s
The G.711 Standard

• The most commonplace codec
  – Used in circuit-switched telephone network
  – PCM, Pulse-Code Modulation

• If uniform quantization
  – 12 bits * 8 k/sec = 96 kbps

• Non-uniform quantization (logarithmic scaling)
  – 64 kbps rate
  – mu-law
    • North America
  – A-law
    • Other countries, a little friendlier to lower signal levels
  – An MOS of about 4.3
Packetization increases bitrate

<table>
<thead>
<tr>
<th>IP</th>
<th>UDP</th>
<th>RTP</th>
<th>Payload</th>
</tr>
</thead>
<tbody>
<tr>
<td>(20 bytes)</td>
<td>(8 bytes)</td>
<td>(12 bytes)</td>
<td>Variable size depending on codec</td>
</tr>
</tbody>
</table>

Header 40 bytes

<table>
<thead>
<tr>
<th>Window</th>
<th>Header</th>
<th>Payload</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td><strong>G.711 (64 kbps)</strong></td>
</tr>
<tr>
<td>20ms</td>
<td>16 kbps</td>
<td>80 kbps</td>
</tr>
<tr>
<td>50ms</td>
<td>6.4 kbps</td>
<td>70.4 kbps</td>
</tr>
<tr>
<td>100ms</td>
<td>3.2 kbps</td>
<td>67.2 kbps</td>
</tr>
</tbody>
</table>

RTP Header Compression Takes a while to kick in 😞

<table>
<thead>
<tr>
<th>Header</th>
<th>Payload</th>
</tr>
</thead>
<tbody>
<tr>
<td>(2 or 4 bytes)</td>
<td>Variable size depending on codec</td>
</tr>
</tbody>
</table>
Packet Losses add latency

- TCP/IP delay impacted by round-trip time (RTT):
  - RTT of 80ms on wired connection from New York to Seattle
  - RTT of 400ms in 3G brings a 1.4Mbps connection (64kb TCP window) with 2% packet loss down to 220kbps TCP connection throughput 😞

- Weak signal strength often results in packet losses
- Cellular networks give less priority to data channel
- TCP/IP retransmissions imply increased latency
- DSR: Distributed Speech Recognition
- Idea: just send the cepstrum, not the audio
- One window every 10ms => 100 frames/sec
- 13 coefficients/frame
- One float per coefficient => 41.6 kbps
- Scalar quantization:
  - one byte per coefficient => 10.4 kbps
- Vector Quantization => 4.8 kbps
- Properties:
  - Lower bitrate
  - Highest accuracy
  - Robust to packet loss
  - No audio to listen to 😞

<table>
<thead>
<tr>
<th>Fields</th>
<th># bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1,C2</td>
<td>6</td>
</tr>
<tr>
<td>C3,C4</td>
<td>6</td>
</tr>
<tr>
<td>C5,C6</td>
<td>6</td>
</tr>
<tr>
<td>C7,C8</td>
<td>6</td>
</tr>
<tr>
<td>C9,C10</td>
<td>6</td>
</tr>
<tr>
<td>C11,C12</td>
<td>6</td>
</tr>
<tr>
<td>C0</td>
<td>8</td>
</tr>
<tr>
<td>CRC</td>
<td>4</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>48</strong></td>
</tr>
</tbody>
</table>
DSR is robust to Packet Loss

Table 3: Error robustness performance over data channels with transmission errors

<table>
<thead>
<tr>
<th>Channel</th>
<th>Noise 1</th>
<th>Noise 2</th>
<th>Noise 3</th>
<th>Noise 4</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unquantized</td>
<td>97.26</td>
<td>96.19</td>
<td>97.61</td>
<td>98.33</td>
<td>97.35</td>
</tr>
<tr>
<td>error free quantize</td>
<td>97.42</td>
<td>96.28</td>
<td>97.76</td>
<td>98.33</td>
<td>97.45</td>
</tr>
<tr>
<td>TETRA TU50 20dB</td>
<td>97.17</td>
<td>96.01</td>
<td>97.26</td>
<td>98.18</td>
<td>97.16</td>
</tr>
<tr>
<td>GSM EP1</td>
<td>97.39</td>
<td>96.28</td>
<td>97.76</td>
<td>98.33</td>
<td>97.44</td>
</tr>
<tr>
<td>GSM EP2</td>
<td>97.39</td>
<td>96.28</td>
<td>97.64</td>
<td>98.24</td>
<td>97.39</td>
</tr>
<tr>
<td>GSM EP3</td>
<td>91.33</td>
<td>91.41</td>
<td>92.72</td>
<td>93.55</td>
<td>92.25</td>
</tr>
</tbody>
</table>

Figure 4: Performance with channel errors: DSR compared to a mobile speech channel
Evolution of Speech Coder Performance

<table>
<thead>
<tr>
<th>Bit Rate (kb/s)</th>
<th>Speech Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Excellent</td>
</tr>
<tr>
<td>4</td>
<td>Good</td>
</tr>
<tr>
<td>8</td>
<td>Fair</td>
</tr>
<tr>
<td>16</td>
<td>Poor</td>
</tr>
<tr>
<td>32</td>
<td>Bad</td>
</tr>
</tbody>
</table>

ITU Recommendations:
- 1980 Profile
- 1990 Profile
- 2000 Profile

Secure Telephony:
- North American TDMA

Cellular Standards:
Code-Excited Linear Predictive (CELP)

Linear-Prediction-based Analysis-by-Synthesis

- How it works
  - Segment speech into frames (typically 20ms long)
  - Find filter parameter for each frame
    \[ H(z) = \frac{1}{A(z)} \quad A(z) = 1 - \sum_{i=1}^{y} a_i z^{-i} \]
  - Find excitation whose that minimizes prediction error
    - Perceptual weighting
    - More accuracy where speech energy is low
  - Transmit the filter parameter and excitation signal
    - Vector quantization
Cellular Codecs

• **GSM:**
  - **Enhanced Full-Rate (EFR)**
    • 8kHz sampling rate, 12.2 kbps, ACELP
  - **Adaptive Multi-Rate Wideband (AMR-WB, G722.2)**
    • 16kHz sampling rate, 16 kbps

• **CDMA**
  - **Enhanced Variable Rate Codec B (EVRC-B)**
    • 8kHz sampling rate, (8.55 kbps, 4.0 kbps, 0.8 kbps), RELP
Microphone Arrays

Microphone arrays useful when driving or gaming
Audio Pipeline in Kinect

- Loudspeakers
- Microphone Array
- MAEC
- MicArray
- AES/NS
- Xbox console
- Audio Playback
- Codec
- Voice Chat
- Speech reco
- Voice Interface
- Other Audio
- Source separation
- Speaker ID
Audio Pipeline in Kinect

Loudspeakers → Microphone Array → MAEC → MicArray → AES/NS → Audio Playback

Game sounds → Codec → Voice chat

Speech reco → Voice Interface

Other Audio → Source separation Speaker ID

Kinect

Xbox console
Directional Microphones

- Acoustical design
  - Using the enclosure shape to increase the microphones directivity

- Optimized microphone array geometry
  - Non-equal spacing, covers the entire bandwidth
Mono Acoustic Echo Cancellation

- Acoustic echo cancellation
  - Mono AEC – part of each speakerphone

\[ y(t) = x(t) \ast h(t) + s(t) + n(t) \]

\[ s(t) + n(t) \]

\[ \text{output} \]

Loudspeaker

Microphone
Multichannel Acoustic Echo Cancellation

- Acoustic echo cancellation
  - “Stereo AEC has a non-uniqueness problem that presents a fundamental limitation” (Sondhi et al. Bell Labs, 1995)
Multichannel Acoustic Echo Cancellation
Ivan Tashev 2008

- Acoustic echo cancellation
  - “Stereo AEC has a non-uniqueness problem that presents a fundamental limitation” (Sondhi et al. Bell Labs, 1995)

- Multichannel AEC
  - Use calibration pulses, lock mixing filters, use one adaptive filter
  - Reduces 15-20 dB echo
  - Entire audio pipeline: ~35 dB
• Adaptive beamformer
  – Acts as a steerable directional microphone
  – Can suppress interferers as well
  – Reduces 3-6 dB noise

• Spatial filtering
  – Sound source localization per frequency bin
  – Suppresses sounds outside desired direction range
  – Suppresses 6-12 dB noise
Mobile Voice Search

Types:
“Lawyers open on Saturdays”

Speaks:
“nearest pizza restaurant”

Plays “Romano Italian at …”

Natural Language Understanding -> Information Retrieval

GPS Profiles Business DB

Speech Recognition

Speech Synthesis
Concatenative TTS

Message Text → Text Analysis, Letter-to-Sound, Prosody → Assemble Units that Match Input Targets → Store of Sound Units (Inventory) of Recorded Speech → Speech Waveform Modification and Synthesis

Dictionary and Rules

Speech

Alphabetic Characters

Phonetic Symbols, Prosody Targets

Text Analysis

Phonetic Analysis

Prosodic Analysis

Speech Synthesis
Mobile Voice Search

Types:
“Lawyers open on Saturdays”

Speaks:
“nearest pizza restaurant”

Plays “Romano Italian at …”

Natural Language Understanding

Information Retrieval

Speech Recognition

Speech Synthesis

GPS Profiles Business DB
Natural Language Understanding

Lexical → Semantic → Syntactic → Pragmatic → Acoustic/Phonetic

Voice Search Applications
- Question/Answering
- Web Search
- Video Search
- Speech Translation
- Business Search

$C^* = \arg \max_C P(C | W^*)P(W^* | X)$
Semantics

When does the flight leaving Boston at 7:30 arrive in Dallas?

REQUEST:
ARR|DEPT TIME

DESTINATION:
Dallas

DEPARTURE TIME:
730 | 1930

ORIGIN:
Boston

REQUEST:
ARRIVAL TIME

DEPARTURE TIME:
7:30

AMBIGUITY

Previous dialog: I want to leave in the morning.
SLU - The Big Picture!

From ASR/DM (text, lattices, n-best, history)

- Text Normalization
  - Morphology, Synonyms

- Parsing/Decoding
  - Extracting named entities, semantic concepts, syntactic tree

- Interpretation
  - Slot filling, reasoning, task knowledge representation

To DM (concepts, entities, parse tree)

Database Access
Text Normalization

Reduce language variation (lexical analysis)

• Morphology
  - Decomposing words into their “minimal unit of meaning (or grammatical analysis)

  ![Diagram of morphemes, prefix, suffix, affixes, stem, un_happy_ness]

• Synonyms
  - Finding words that mean the same (hello, hi, how do you do)

• Equalization
  - disfluencies, non-alphanumeric, capitals, non-white space characters, etc.
Mapping textual representation into semantic concepts using knowledge rules and/or data.

• **Entity Extraction (Matcher)**
  - Finite state machines (FSM)
  - Context Free Grammar (CFG)
  - Classifiers

• **Semantic Parsing**
  - Tree structured meaning representation allowing arbitrary nesting of concepts
  - Segment an utterance into phrases each representing a concept (e.g. using HMMs)

• **Classification**
  - Categorizing an utterance into one or more semantic concepts.
Entity Extraction through Finite-state Composition

- check
- my
- January
- second
- bill

ε:<month> January ε:<month> ε:<day> first ε:<day>

December

thirtieth

check my <month> January <month> <day> second <day> bill
Interpret user’s utterance in the context of the dialog. Interpretation is encoded through a set of rules.

- **Ellipses and anaphora**
  
  "What about for New York?"

- **History Mechanism**
  - Communicated messages are differential.
  - Removing ambiguities

- **Database Look-up**
  - Retrieving or checking entities
Building Accurate Voice UI is hard

• Traditional Context Free Grammar (CFG):
  
  
  <one-of>
  
  <item> business search </item>
  <item> search </item>
  <item> biz search </item>
  <item> driving directions </item>
  <item> directions </item>
  <item> traffic </item>
  <item> tell me my choices </item>
  <item> What are my options </item>
  
  ...
  
  </one-of>

• Easy to write but fragile
Data driven speech understanding

- ASR engine
- n-gram
- IR
- Inverted index

Runtime

Training

Build Index

Directives: <directions>
Directions please <directions>
Business <business>
Business search <business>
How bad is the traffic <traffic>
Traffic <traffic>

Count n-grams

Uh uh directions

<directions>
Example-based SLM

- Interpolation of
  - Large general domain bigram model
  - Small domain specific bigram model through backoff state
- Robust SLM with little in-domain data
Mobile Voice Search

Types:
“Lawyers open on Saturdays”

Speaks:
“nearest pizza restaurant”

Plays “Romano Italian at …”
User-initiative dialogs

• Pros:
  – Can result in a shorter call
  – Can feel more natural
  – Useful when too many choices

• Cons:
  – Requires expensive expertise
  – Could lead to user frustration: system appears human but caller can’t use full natural language
Who manages the Dialog?

Directed Dialog
- “Who would you like to contact?”
- Finite State Machine
- Simple CFG
- MSConnect

Initiative

User Initiative Dialog
- “What can I do for you?”
- Ngrams
- Windows Airlines

Reservations → Flight Status → Baggage Claim → Special Announcements
airline traveler journey: a trip

At each stage:

What are the callers immediate needs?
Which set of tasks do they want to perform?
How can we use what we already know to shorten the process?
Design for the user
Tellme circa 2000

United

TellMe

NWA

American Airlines

0:42 min

1:40 min

[Stop]
Mobile Voice Search

Types:
- "Lawyers open on Saturdays"
- "nearest pizza restaurant"
- "Romano Italian at ..."

Natural Language Understanding

Information Retrieval

GPS Profiles Business DB

Multimodal Inputs

Speech Recognition

Speech Synthesis

Multimedia Output
- Going beyond voice: Multimodal interfaces
  - Combine graphical user interfaces with spoken dialog
- Sample application: Interactive city guide
  - Airport arrival lounge, American Express office, Tourist center …
  - Restaurant and subway information for Washington, DC. and NYC
  - Delivering DC version to AT&T LGA for visitors to AT&T Innovation Center
- Flexibility: All commands can be speech, pen, or multimodal
  - “How do I get to here?” + <circles location>
- Mobility: Runs on both tablet pc and public kiosk
1. Multimodal Integration and Understanding Using Finite-state Language Processing Techniques
   - Multimodal grammar captures speech understanding, gesture understanding, and multimodal integration in a single model
   - Compiles to an efficient finite-state device
   - Interprets speech, pen, and multimodal inputs

2. Multimodal Testbed Platform

3. MATCH: Multimodal Access To City Help
   - Provides an interactive guide and navigation for information-rich urban environments (NYC, DC)
   - **Understands**: speech, pen, and multimodal input
   - **Generates**: dynamic multimodal presentations combining TTS with graphics and gestures
Speech-centric Multimodal
Kinect: Gesture Recognition with 3D camera
Kinect: Speech Recognition

• Speech recognition
  – Complementary to gesture
  – Want to talk to your animal
  – Voice control without on-screen buttons
  – Access long lists

• From headsets to hands free
  – Needs relatively good quality audio!
  – Loud gaming sounds from Xbox
  – Noise and reverberation in the room
Kinect: Voice Control
Types: “Lawyers open on Saturdays”

Natural Language Understanding

Information Retrieval

GPS Profiles Business DB

Spech Recognition

Multimodal Inputs

plays "Romano Italian at ..."

Speech Synthesis

Multimedia Output

speaks: “nearest pizza restaurant”
Web Standards

CLIENT: HTML+Speech (W3C XG)
SALT
X+V

SRGS
SISR

EMMA: Extensible Multimodal Annotation

MMI Architecture SCXML

Speech Reco

Gesture Reco

NLU

Interaction Manager

InkML
Tutorial Outline

- Introduction and Historical prospective
- Advances in voice search technologies
- The mobile revolution
- Mobile voice applications
- Market dynamics and business models
- Technology challenges and the road ahead
Explosion of Mobile Devices
Explosion of Multimedia Data
Information On Demand
Anywhere & Anytime

Record Armageddon on CBS tonight

Market Revolution: Speech and multimodal interfaces are natural communication modalities for mobility across multiple screens.
Tutorial Outline

- Introduction and Historical prospective
- Advances in voice search technologies
- The mobile revolution
- **Mobile voice applications**
- Market dynamics and business models
- Technology challenges and the road ahead
• **Differentiated User Experience:** Speech is faster, more accurate, easier, can multitask (e.g., speak and drive), fun to use, and meet regulatory mandates on hands-free, eyes-free.

• **Market transition:** Speech is a transitional technology for mobility. IC’s faster, memory larger, phones smaller. But fingers are not getting smaller. Need screen, need to eliminate typing.
Examples of Mobile Applications
Searching a manual by voice

Ju 2008

How much does Xbox cost?

Page 12
SMS in Cars (Ford SYNC, Hyundai/Kia UVO)

- SMS are commonly used
- But sending SMS while driving is dangerous
  - and illegal in many countries
- Ford SYNC reads SMS using TTS
- Most SMS only require short replies
<table>
<thead>
<tr>
<th>FORD SYNC Canned SMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>I’LL BE LATE</td>
</tr>
<tr>
<td>MEETING CANCELLED</td>
</tr>
<tr>
<td>CAN'T TALK RIGHT NOW</td>
</tr>
<tr>
<td>CALL ME</td>
</tr>
<tr>
<td>WHERE R YOU?</td>
</tr>
<tr>
<td>I NEED MORE DIRECTIONS</td>
</tr>
<tr>
<td>THANKS</td>
</tr>
<tr>
<td>I AGREE</td>
</tr>
<tr>
<td>I DISAGREE</td>
</tr>
<tr>
<td>I’M STUCK IN TRAFFIC</td>
</tr>
<tr>
<td>C U IN 5(10,15,20) MINUTES</td>
</tr>
</tbody>
</table>
SMS Dictation using voice search
Ju, 2009

SMS Database
I’m stuck in traffic
Call U later
... See U soon

Lunch?

SLM

ASR
No I can get lunch
today out of next week

Index

TF-IDF

Not today, next week
Press the button and then use speech to reply the message

Try Another SMS
AT&T/ChaCha Mobile Service
Powered by AT&T WATSON

- Human assisted question/answering (~4%)
- Average 20 transactions per user per month

500+ million Q&A
Vlingo Voice Search
Powered by AT&T WATSON

- Top-Level Phone UI using advanced speech technology
- User presses a button, speaks what they way, application handles
- Unconstrained Speech Technology to remove application constraints
- MultiModal UI to handle errors and disambiguation
Your Personal Agent

- **E-Reading**
  - Natural sounding text to speech
  - Language understanding and voice search

- **Interactive Dialog**
  - “What is the review for this book?”
  - “Find my presentation for GoTV from last week”
  - “Read Alice in the Wonderland”
  - “Italian restaurants along this street”
Making video archives searchable and browsable
Spoken language is a natural interface to access and search television content, and to order products and services.
Multimodal Voice Remote for IPTV
Tutorial Outline

- Introduction and Historical prospective
- Advances in voice search technologies
- The mobile revolution
- Mobile voice applications
- Market dynamics and business models
- Technology challenges and the road ahead
Market Landscape

• Information access
• Legal
• Customer care
• Medical and healthcare
• Messaging
• Language translation
• Unified communication
• Social media
• Gaming
• Ordering
Market Drivers for Mobile Services

- **Improved User Experience**
  - Smart phones have limited keyboard and screen size

- **Technology Advances**
  - Mobility, bandwidth, cloud computing, statistical modeling from massive data

- **Application Ecosystem**
  - Enabling developers to build applications

- **Revenue & Efficiency**
  - Subscription or ad-based business model.

- **Regulatory Requirements**
  - Text messaging banned while driving in some states

---

Mobile Voice Services
Technology Advances

- Increased Bandwidth 3G/4G LTE
- Algorithmic advances for large scale data and model creation
- Pervasive cloud computing promising lower cost deployments
- New complementary enablers such as location based information
- Penetration of smart phones with large screens
- Personalized user interfaces
- Standardized interfaces
Business Models

- One time fee per download
- Subscription-based
- Ad-sponsored
- Fulfillment
- Freemium
- Per usage charge
- Free (part of another service)
Enabling Innovation Ecosystem

Lowering the barrier to entry for developers

Consumer & Enterprise Customers

Developer Community

API

Developer Center

Speech As A Service

SMS  TTS  AGPS  Storage
ASR  Email  Presence
Search  Translation  NLU
Tutorial Outline

- Introduction and Historical prospective
- Advances in voice search technologies
- The mobile revolution
- Mobile voice applications
- Market dynamics and business models
- Technology challenges and the road ahead
Big Research Challenges

- Open and robust speech recognition
- Multimodal understanding (speech and gesture)
- Multimedia presentation (speech and visual)
- Understanding and translating open speech
- Interactive and personalized dialog
- Real-time analytics and computational advertizing
Vision of the Future
Free Downloads

- ChaCha: Question Answering (chacha.com)
- Speak4it: Business Search (speak4it.com)
- YPMobile: Local Guide (yp.com)
- Buzz.com: Personal Social Network (buzz.com)
- Vlingo: Web search and SMS (vlingo.com)