Dolování informací z řeči pro obranu a bezpečnost

Honza Černocký
Brno University of Technology,
BUT Speech@FIT group and IT4I center of Excellence

AFCEA - seminář PS07 Inteligence OD DAT KE ZNALOSTEM
Praha, 17.1.2019
Agenda

• Introduction
• Speech recognition
• Speaker recognition
• R&D Challenges
• Some projects
Why and how?

- Speech is the most important modality of human-human communication (~80% of information) ... criminals and terrorists are also communicating by speech.
- Speech is easy to acquire in the scenarios of interest.
- More difficult is to find what we are looking for.
- Typically done by human experts, but always count on:
  - Limited personnel
  - Limited budget
  - Not enough languages spoken
  - Insufficient security clearances

Technologies of speech processing are not almighty but can help to narrow the search space.
Who speaks? John Doe

What gender? Male or Female

What language? English/German/??

What was said? “Hello John!”

Who asked whom? John asked Paul
BUT Speech@FIT 1997-2019
Brno Speech Valley
Working with and for law enforcement

• First cooperation with Czech MoD people back in 2005 (leading to establishing Phonexia)

Projects funded by

• 3 Czech MoI
• 2 DARPA
• 3 IARPA
• EU FP7 ICT-2007.1.4 (Secure, dependable and trusted Infrastructures)

Meeting the LEA guys

• directly

• Regular seminars “Speech technologies” run by MoD for the Czech security & defense community.
Agenda

• Introduction
• **Speech recognition**
• Speaker recognition
• R&D Challenges
• Some projects
Automatic speech recognition (Transcription, speech to text)

Acoustic models
Language model
Pronunciation dictionary

Recognition network

Feature extraction
Evaluation of likelihoods (scores of hypothesis)
“Decoding”

speech → text
Technology behind

• Classification and recognition, nowadays mostly by variants of deep neural networks
Technology behind II

- Decoding from big probabilistic graphs composed of finite state acceptors and transducers => 1-best, sausage (confusion network), lattice
Training

• Acoustic models
  • ... how do speech segments match basic speech units (phonemes)
  • trained on large (>100h) quantities of carefully transcribed speech data

• Language models
  • ... how do the words follow each other
    President George Bush
    President George push
  • Need to be trained on large quantities (Gigabytes) of text from the target domain

• Pronunciation dictionary
  • Translate words into phonemes: dog → d oh g
  • Basis needs to be created by hand, the rest generated using trained grapheme to phoneme (g2p) converter
What to expect?

< 5% word error rate
Spontaneous speech in “reasonable” languages with training data

and then they have one week to retrain their keyword results ... and ... give you might ask why one we there a lot of research or evaluation methods ... the people are trying out what keywords or so it is important to leave a ... sufficient amount of time there as well ...

< 20% WER
Languages we have never heard

uhuh kade sengifowunelwe nguThami
manje ithi angazi e- ekhuluma
nomunye ubhuti wakwamasipala ukuthi
ene usho ukuthi kunabantu ekufanele
baphelelewe ngumsebenzi ngoba
uNomvula emecabanga uzokhokha (())
ngoba yena uzoy ithela uzoyi
uzoyihlulisela ngoba phela kukhona
aba- abaphethe u-Adam angithi

<50% WER (if lucky)
Agenda

• Introduction
• Speech recognition
• Speaker recognition
• R&D Challenges
• Some projects
Speaker recognition

**Verification**
Is this Homer’s voice?

**Identification**
Whose voice is this?

**Diarization** (Segmentation and Clustering)
Where are speaker changes? Which segments are from the same speaker?
Technology

Front end
- Signal processing
- Feature extraction
- Conversion to low-dimensional vector

Low-dim. vectors

Back end
- Vector normalization
- Score computation
- Score normalization

“voice prints”

speech

Score
2-3 years back: i-vectors

- i-Vectors based on Gaussian models.

[Diagram showing the process of i-vector extraction and verification]
Now: x-vectors

- Extraction of vectors by Gaussian models is replaced by a DNN trained to recognize speakers.
Training again

• X-vector extractor: 100s of hours of data with speaker labels.

• Scoring done by PLDA
  • Voice-prints with speaker labels (A, B, C, ...) needed
  • Even 50 speakers help to increase the accuracy by 30%.
What to expect?

- Works very nicely for long telephone recordings (EER ~2%) – multiple successes in NIST evaluations.
Bad channels, short speech segments ...
Agenda

• Introduction
• Speech recognition
• Speaker recognition
• R&D Challenges
• Some projects
Multi-linguality

Language(s) with abundant resources

Good ASR

Language with low resources

Reasonable ASR

<table>
<thead>
<tr>
<th>Year 1 (Y1)</th>
<th>Language</th>
<th>ASR Code</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cantonese</td>
<td>IARPA-babel1101-v0.4c</td>
<td>CA</td>
<td></td>
</tr>
<tr>
<td>Pashto</td>
<td>IARPA-babel1104b-v0.4aY</td>
<td>PA</td>
<td></td>
</tr>
<tr>
<td>Turkish</td>
<td>IARPA-babel1105-v0.6</td>
<td>TU</td>
<td></td>
</tr>
<tr>
<td>Tagalog</td>
<td>IARPA-babel1106-v0.2g</td>
<td>TA</td>
<td></td>
</tr>
<tr>
<td>Vietnamese</td>
<td>IARPA-babel1107b-v0.7</td>
<td>VI</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year 2 (Y2)</th>
<th>Language</th>
<th>ASR Code</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assamese</td>
<td>IARPA-babel1102b-v0.5a</td>
<td>AS</td>
<td></td>
</tr>
<tr>
<td>Bengali</td>
<td>IARPA-babel1103b-v0.4b</td>
<td>BE</td>
<td></td>
</tr>
<tr>
<td>Haitian Creole</td>
<td>IARPA-babel1201b-v0.2b</td>
<td>HA</td>
<td></td>
</tr>
<tr>
<td>Lao</td>
<td>IARPA-babel1203b-v3.1a</td>
<td>LA</td>
<td></td>
</tr>
<tr>
<td>Zulu</td>
<td>IARPA-babel1206b-v0.1e</td>
<td>ZU</td>
<td></td>
</tr>
<tr>
<td>Tamil</td>
<td>IARPA-babel1204b-v1.1b</td>
<td>Tam</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year 3 (Y3)</th>
<th>Language</th>
<th>ASR Code</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kurdish</td>
<td>IARPA-babel1205b-v1.0a</td>
<td>KU</td>
<td></td>
</tr>
<tr>
<td>Cebuano</td>
<td>IARPA-babel1301b-v2.0b</td>
<td>CE</td>
<td></td>
</tr>
<tr>
<td>Kazach</td>
<td>IARPA-babel1302b-v1.0a</td>
<td>KA</td>
<td></td>
</tr>
<tr>
<td>Telugu</td>
<td>IARPA-babel1303b-v1.0a</td>
<td>TE</td>
<td></td>
</tr>
<tr>
<td>Lithuanian</td>
<td>IARPA-babel1304b-v1.0b</td>
<td>LI</td>
<td></td>
</tr>
<tr>
<td>TokPisin</td>
<td>IARPA-babel1207b-v1.0c</td>
<td>TP</td>
<td></td>
</tr>
<tr>
<td>Swahili</td>
<td>IARPA-babel1202b-v1.0d</td>
<td>SW</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year 4 (Y4)</th>
<th>Language</th>
<th>ASR Code</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pashto</td>
<td>see Year 1 – progress set</td>
<td>PA2</td>
<td></td>
</tr>
<tr>
<td>Javanese</td>
<td>IARPA-babel1402b-v1.0b</td>
<td>JA</td>
<td></td>
</tr>
<tr>
<td>Igbo</td>
<td>IARPA-babel1306b-v2.0c</td>
<td>IG</td>
<td></td>
</tr>
<tr>
<td>Mongolian</td>
<td>IARPA-babel1401b-v2.0b</td>
<td>MO</td>
<td></td>
</tr>
<tr>
<td>Dhlotu</td>
<td>IARPA-babel1403b-v1.0b</td>
<td>DH</td>
<td></td>
</tr>
<tr>
<td>Guarani</td>
<td>IARPA-babel1305b-v1.0b</td>
<td>GU</td>
<td></td>
</tr>
<tr>
<td>Amharic</td>
<td>IARPA-babel1307b-v1.0b</td>
<td>AM</td>
<td></td>
</tr>
</tbody>
</table>
Encoder-decoder models, attention, adversary networks, embeddings everywhere ...

• Lots of research
  • General machine learning
  • Natural language processing
  • Machine translation
  • Etc etc etc

• Sequence to sequence modeling is ultimately what we need to do!

Vaswani et al. 2017: Attention is All You Need.
Far-field

- microphone arrays
- De-noising, de-reverberation
- re-training of the recognition system on simulated data
Adaptation to user

R&D: Is it working for you?
LEA: No.
R&D: Can you give details on the problem?
LEA: Yes, in one phrase it got almost everything wrong.
R&D: Can we listen to the data?
LEA: No.

Front end
- Signal processing
- Feature extraction
- Conversion to low-dimensional vector

Low-dim. vectors

“voice prints”

Back end
- Vector normalization
- Score computation
- Score normalization

Score

speech

content

No content
Agenda

• Introduction
• Speech recognition
• Speaker recognition
• R&D Challenges
• Some projects
DRAPAK

Information mining in speech acquired by distant microphones

• Czech MoI
• 2015–2020, BUT and Phonexia
• From telephone close-talk speech to far-field microphones, mainly for security and defense applications.
• Consultations with LEA/defense partners, awareness in the community
• Data
• Technology
  • Microphone arrays
  • Data augmentations
  • Denoising, dereverberation
• Transfer to working products (Phonexia)
Real time network, text, and speaker analytics for combating organized crime

- H2020 Security
- Speech, NLP, video and network analysis
- 3 years, 7 MEUR ...
- Thanks PCR for support!

<table>
<thead>
<tr>
<th>#</th>
<th>Organization</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Idiap Research Institute (Coordinator)</td>
<td>CH</td>
</tr>
<tr>
<td>2</td>
<td>Trilateral Research Ltd.</td>
<td>GB</td>
</tr>
<tr>
<td>3</td>
<td>Brno University of Technology</td>
<td>CZ</td>
</tr>
<tr>
<td>4</td>
<td>Phonexia s.r.o.</td>
<td>CZ</td>
</tr>
<tr>
<td>5</td>
<td>SAIL LABS Technology GmbH</td>
<td>AT</td>
</tr>
<tr>
<td>6</td>
<td>Capgemini Technology Services</td>
<td>FR</td>
</tr>
<tr>
<td>7</td>
<td>INTERPOL</td>
<td>FR</td>
</tr>
<tr>
<td>8</td>
<td>Saarland University</td>
<td>DE</td>
</tr>
<tr>
<td>9</td>
<td>KENTRO MELETON ASFALEIAS</td>
<td>GR</td>
</tr>
<tr>
<td>10</td>
<td>Leibniz University of Hannover</td>
<td>GR</td>
</tr>
<tr>
<td>11</td>
<td>Università Cattolica del Sacro Cuore - TRANSCRIME</td>
<td>IT</td>
</tr>
<tr>
<td>12</td>
<td>AEGIS IT RESEARCH UG</td>
<td>DE</td>
</tr>
<tr>
<td>13</td>
<td>Airbus Defence and Space</td>
<td>FR</td>
</tr>
<tr>
<td>14</td>
<td>Police of Czech Republic</td>
<td>CZ</td>
</tr>
<tr>
<td>15</td>
<td>Romanian Minister of Interior</td>
<td>RO</td>
</tr>
<tr>
<td>16</td>
<td>Lithuanian Forensic Science Centre</td>
<td>LT</td>
</tr>
<tr>
<td>17</td>
<td>Police Service of Northern Ireland</td>
<td>UK</td>
</tr>
<tr>
<td>18</td>
<td>ADITESS Advanced Integrated Technology Solutions &amp; Services LTD</td>
<td>CY</td>
</tr>
<tr>
<td>19</td>
<td>Ministry of Interior, Croatia</td>
<td>HR</td>
</tr>
<tr>
<td>20</td>
<td>Netherlands Forensic Institute</td>
<td>NL</td>
</tr>
<tr>
<td>21</td>
<td>Internet of Things applications and Multi-Layer development</td>
<td>GR</td>
</tr>
<tr>
<td>22</td>
<td>Ministry of Public Security – Israel National Police</td>
<td>IL</td>
</tr>
<tr>
<td>23</td>
<td>Hellenic Police</td>
<td>GR</td>
</tr>
</tbody>
</table>
Neural Representations in Multi-modal and Multi-lingual Modeling

• GACR, 2019-2023
• BUT and UFAL MFF Charles University in Prague

Basic research in neural speech processing (SP) and natural language processing (NLP) with accent on multilinguality and multi-modality (speech and text processing with the support of visual information):
  • Systematic study of neural structures for speech and text modeling in multi-modal and multilingual settings.
  • Addressing hierarchy of neural representations, human interpretability, and training under realistic conditions of non-ideal and incoherent data.
Thank you for your attention!

Questions?

cernocky@fit.vutbr.cz
http://speech.fit.vutbr.cz/