Multistream recognition of speech

Hynek Hermansky
Center for Language and Speech Processing
The Johns Hopkins University, Baltimore, USA
and
FIT VUT Brno Czech Republic
When decreasing entropy, one should use knowledge.

LOW ENTROPY
The Demon closes door when a slow air molecule comes and lets the fast air molecules to go through.

The Demon must KNOW which molecule is fast and which is slow!

knowledge comes from
- magic
- measurements

When decreasing entropy, one should use knowledge!
machine

message

> 50 kb/s
C = \( W \log_2(S/N+1) \), \( W=5\text{kHz} \), \( S/N+1>10^3 \)
who is speaking, emotions, accent, acoustic environment,....

< 50 b/s
< 3bits/phoneme, < 15 phonemes/s
linguistic message

Information rate (entropy) reduction
• requires knowing what to leave out and how
KNOWLEDGE

- magic
- experts, beliefs, previous experience (hardwired)
- measurements (data)

HARDWIRED

- reusable permanent knowledge
  - no need to re-learn known facts

but

- experts and beliefs can be wrong

DATA

- no knowledge better than wrong knowledge
  - data do not lie

but

- transcribed data are expensive

REUSEABLE AND HARDWAREABLE KNOWLEDGE FROM DATA!
Acoustic Processing in ASR

features (signal processing)
- what we already know (general knowledge)
- alleviate unwanted information
  - wanted information, which is left out is gone forever

classifier (machine learning)
- what we yet do not know (task-specific knowledge)
- typically stochastic (trained on data)
  - unwanted information, which is kept, requires more complex classifiers, trained on more data
Data-driven approaches dominate ASR field

Artificial Neural Networks

- Discriminative nonlinear classifiers introduced to ASR in late eighties of 20\textsuperscript{th} century
- Fewer restrictions on form of input features
- Current hardware advances allow for new revolutionary approaches to ASR

\textit{BIG DATA}

\begin{itemize}
  \item \textbf{deep neural net} \\
  \rightarrow \quad \text{information}
\end{itemize}
Deep Neural Net:
Hierarchical convolutional long-short-memory highway-connected attention-based bi-directional-gated pyramidal temporal-classifying recurrent DNN.

**New DNN structures and their parameters**

New opportunities to verify existing knowledge and to learn new things.

**Data-derived knowledge should be hardwired into future designs!**
Deep Neural Network Based ASR from Raw Speech Signal

Tüske, Golik, Schlüter and Ney 2015

Power spectrum

Convolutions with input speech signal

Remaining fully connected hidden layers of the deep neural networks

Posterior probabilities of generalized tied triphones

Speech
Data-driven two-stage acoustic processing of raw speech signal (spectrum and time-frequency cortical-like filters)

Golik, Tüske, Schlüter and Ney 2015

speech

\[ \text{convolutions with input speech signal} \]

\[ \text{convolutions with time trajectories of power spectra} \]

\[ \text{remaining fully connected hidden layers of the deep neural networks} \]

\[ \text{posterior probabilities of generalized tied triphones} \]
Some examples of mammalian auditory cortical receptive fields

Patil et al 2012
Spectral (simultaneous) masking

**spectral masking:**
detection of signal in one critical band is not influenced by signal in another critical band

Fletcher 1933
any change in the tract shape is reflected at ALL FREQUENCIES of speech spectrum!
Articulatory Bands
French and Steinberg 1949


- 20 frequency bands in speech spectral region
- each band contributes about equally to human speech recognition
- any 10 bands sufficient for 70% correct recognition of nonsense syllables, better than 95% correct recognition of meaningful sentences [Fletcher and Steinberg 1929]
127 different stream combinations in hierarchical MLP structures

evaluate word error for different stream combinations

Hermansky et al 1996
Human Recognition Strategy (and eventually also machines)?

Divide et Impera

- colored noise can be seen as close to white noise in individual bands
- corrupted frequency bands could be left out from further processing
Word error rates of DNN recognizer on Aurora noisy data (relative change in brackets)

<table>
<thead>
<tr>
<th>Auditory Spectral Streams</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.6</td>
</tr>
<tr>
<td>11.0</td>
</tr>
<tr>
<td>(-12.8)</td>
</tr>
</tbody>
</table>

Some of the streams may carry garbage

Train fusing DNN on inputs, which carry no information. During training, randomly set some stream outputs to all-zero.

Similar to feature dropping but here the whole organized sets of features representing streams are being dropped at any given time.
Word error rates of DNN recognizer on Aurora noisy data (relative change in brackets)

- Auditory spectral stream dropping
- 12.6  11.0  9.9
- (-12.8)  (-10.1)

Performance monitoring

Knowing when the result in probability estimation is in error would allow for the selection of the best performing stream combination.

Performance monitoring:
requires estimation of performance of a classifier without knowing what the correct result is.
“good” posteriogram – derived from speech data similar to its training

“bad” posteriogram – derived from corrupted speech data
How “clean” is a posteriogram?

$$M(\Delta i) = \frac{\sum_{i=0}^{N} D(p_i, p_{i+\Delta \tau})}{N}$$

$\Delta i$ – time delay
$D(.)$ – symmetric KL divergence

clean data
noisy data

0 250 ms $\Delta i$
Quality of speech signal from microphone array

from Bernd T. Meyer

performance monitoring module

speaker

noise source

M-Measure

Azimuth

Quality of speech signal from microphone array from Bernd T. Meyer

performance monitoring module

speaker

noise source

M-Measure

Azimuth
How “similar” is the estimator performance on its training data and in the test?

Mesgarani et al 2011

DNN auto-encoder, trained on output of the estimator when applied to its training data

- **Training of probability estimator**
  - Trains on output of the estimator when applied to its training data.
  - Targets from labels.

- **Training of performance monitor**
  - Trains to minimize $(p' - p'')^2$.

- **Performance monitor in use**
  - Evaluates $(p'_{test} - p''_{test})^2$.

![Diagram showing training processes and performance evaluation](image)
Word error rates of DNN recognizer on Aurora noisy data (relative change in brackets)

<table>
<thead>
<tr>
<th>auditory spectrum</th>
<th>spectral streams</th>
<th>stream dropping</th>
<th>performance monitoring</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.6</td>
<td>11.0</td>
<td>9.9</td>
<td>9.6</td>
</tr>
<tr>
<td>(-12.8)</td>
<td>(-10.1)</td>
<td>(-2.8)</td>
<td></td>
</tr>
</tbody>
</table>

Word error rates of DNN recognizer on Aurora noisy data (relative change in brackets)

<table>
<thead>
<tr>
<th>Band</th>
<th>DNN1</th>
<th>DNN2</th>
<th>DNN3</th>
<th>DNN4</th>
<th>DNN5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-3 Bark</td>
<td>12.6</td>
<td>11.0</td>
<td>9.9</td>
<td>9.6</td>
<td>7.9</td>
</tr>
<tr>
<td>4-6 Bark</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7-11 Bark</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12-15 Bark</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16-19 Bark</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Multiple parallel noise-specific streams

- speech
  - clean
  - car
  - crowd
  - ship1
  - ship2

pick the best stream

performance monitor

phoneme error rates noisy TIMIT

<table>
<thead>
<tr>
<th>train / test</th>
<th>clean</th>
<th>car</th>
<th>crowd</th>
<th>ship1</th>
<th>ship2</th>
</tr>
</thead>
<tbody>
<tr>
<td>multi-style</td>
<td>23.0</td>
<td>24.9</td>
<td>39.4</td>
<td>42.0</td>
<td>43.0</td>
</tr>
<tr>
<td>matched</td>
<td>20.7</td>
<td>22.8</td>
<td>37.0</td>
<td>38.1</td>
<td>37.6</td>
</tr>
<tr>
<td>oracle (cheating)</td>
<td>18.4</td>
<td>20.5</td>
<td>34.7</td>
<td>34.5</td>
<td>31.8</td>
</tr>
<tr>
<td>multi-stream with</td>
<td><strong>20.9</strong></td>
<td><strong>22.9</strong></td>
<td><strong>36.8</strong></td>
<td><strong>36.6</strong></td>
<td><strong>36.8</strong></td>
</tr>
</tbody>
</table>

Mallidi et al ASRU 2015
Many ways of seeing the signal

Number of neurons

- **Apex**: 100M
- **Base**: 100K

Speed of firing

- **Apex**: 10Hz
- **Base**: 1kHz
Concept of multi-stream recognition

- stream forming
- performance monitoring
- stream selection
- fusion

EXTRACTED INFORMATION

- different streams
  - modalities,
  - frequency bands,
  - spectral and temporal resolutions,
  - levels of prior knowledge

SIGNAL
THANKS

Sri Harish Mallidi  Nima Mesgarani  Tetsuji Ogawa  Samuel Thomas  Feipeng Li

Ehsan Variani  Phani Nidadavolu  Vijay Peddinti  Bernd T Meyer
Regarding the database:
The training set consists of 14 hours of multi-condition data, sampled at 16 kHz. Total 7137 utterance from 83 speakers. Half of the utterances were recorded by the primary Sennheiser microphone and the other half were recorded using one of a number of different secondary microphones. Both halves include a combination of clean speech and speech corrupted by one of six different noises (street traffic, train station, car, babble, restaurant, airport) at 10-20 dB signal-to-noise ratio.

The test set consist of 14 conditions, with 330 utterances for each condition. The conditions include clean set recorder with primary Sennheiser microphone, clean set with secondary microphone, 6 additive noise conditions which include airport, babble, car, restaurant, street and train noise at 5-15 dB signal-to-noise ratio (SNR) and 6 conditions with the combination of additive and channel noise.

Regarding the features:
From signal extract 63 Mel filterbank energies
At a given frame, take 11 frame context (-5, +5)
In each subband project the 11 frame context onto 6 dct basis