Multi-Resolution Spectral Input for CNN-based Speech Recognition

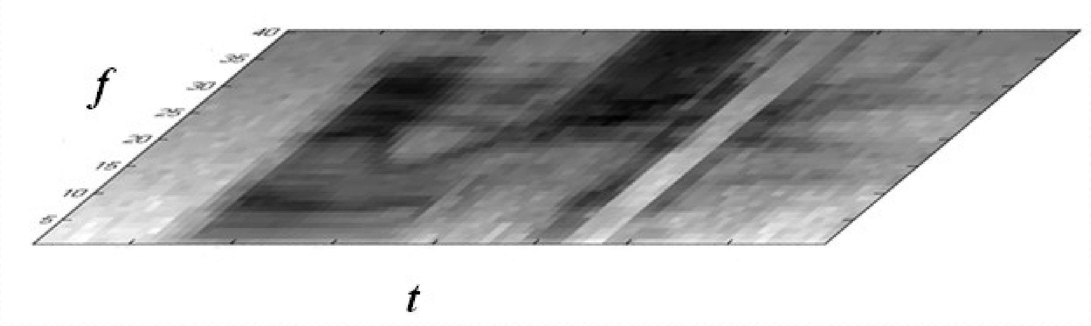
Dr. László Tóth

University of Szeged

Department of Computer Algorithms and Artificial Intelligence

Motivation

- HMMs: the conventional input is the MFCC representation
  - A short-term spectral representation plus a DCT to decorrelate the features
  - The time context is not taken into consideration (only by the “delta” vectors)
- DNNs:
  - DNNs do not require the decorrelation of features (the DCT step)
  - They can efficiently make use of a wider context (9-51 neighboring frames)
- From MFCCs we returned to a spectro-temporal input representation
  - ***f*: 23-40 mel bands**
  - ***t*: 9-51 frames**

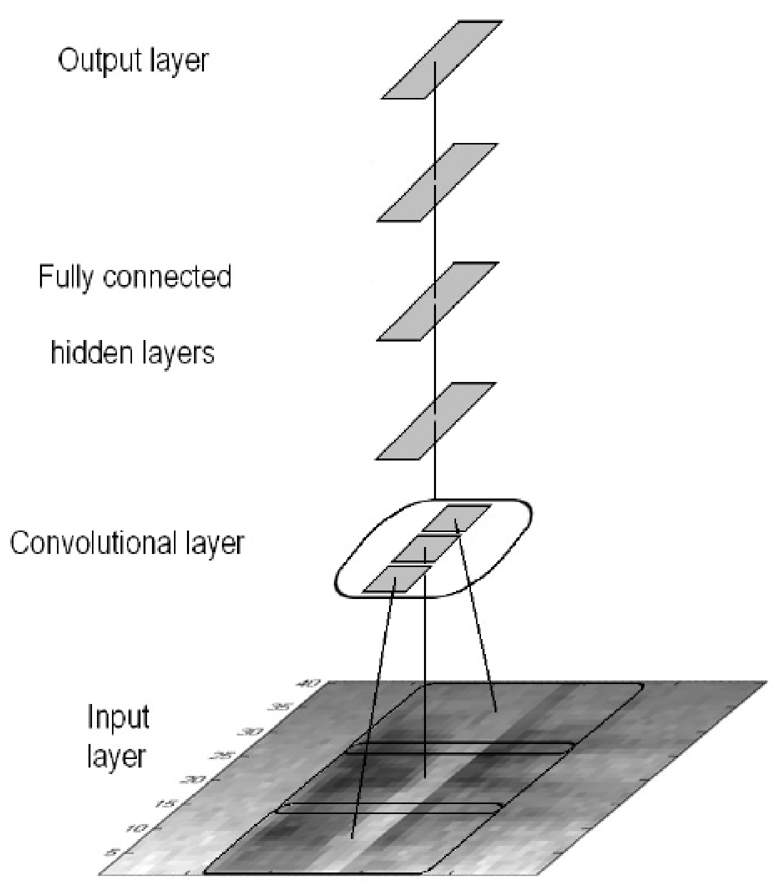


Questions

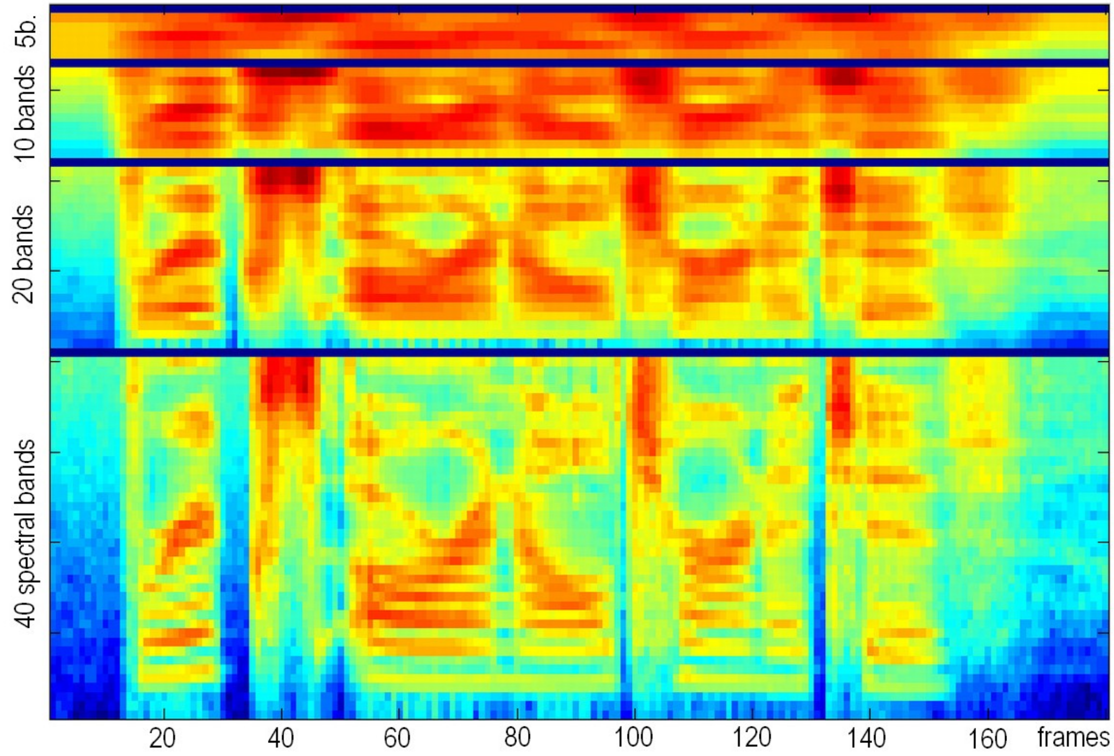
- An early paper used a mel-spectrum input of 40 bands (Mohamed at al, 2012)
  - To be comparable, many following papers used the same input
  - But they gave no explanation why they used 40 spectral bands
  - **QUESTION #1: Is this optimal?**
- Most authors vary the size of the input between 9-51 frames
  - Adding more and more frames introduces less and less extra information
  - However, the number of features increases linearly with the size
  - **QUESTION #2: Would a multi-resolution input help?**
  - Assumptions:
    - It is enough to represent the frames farther away from the center at a lower resolution, as they contain less additional information
    - The neural network can mine the information more efficiently from a smaller set of features (the “curse of dimensionality” problem)

The Convolutional Neural Network

- The structure is the same as that I talked about earlier…
- For simplicity, here we applied the convolution only along frequency
- The baseline system operates with 40 spectral channels
- These are decomposed into 7 convolutional bands
- 1 convolutional layers (with maxout neurons)
- 3 fully connected layers (with maxout neurons)

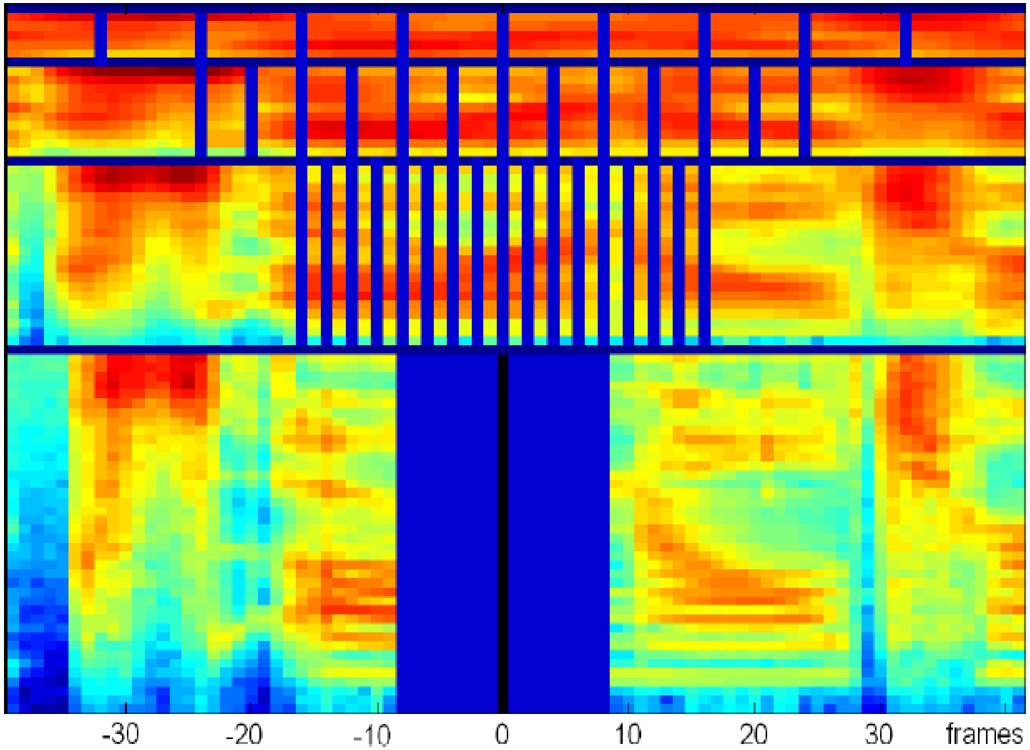


The Multi-Resolution Input



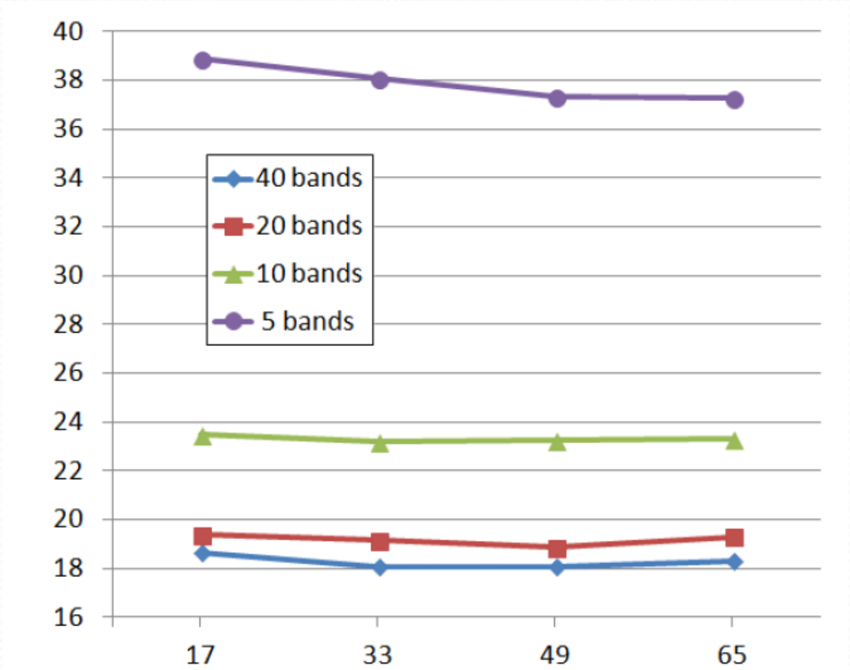
- **4 different spectrograms with decreasing resolution**
  - **40-20-10-5 spectral bands**
  - **Window size in time: 25-50-100-200 ms**
    - **The frames will be downsampled by the CNN**

Illustration of Downsampling



- **Input1: 17 frames of context, 40*17=680 features**
- **Input2: 33 frames of contect, 20*33=660 features**
- **Input3: 49 frames of contect, 10*49=490 features**
- **Input4: 65 frames of contect, 5*65=325 features**

Evaluation - Separately



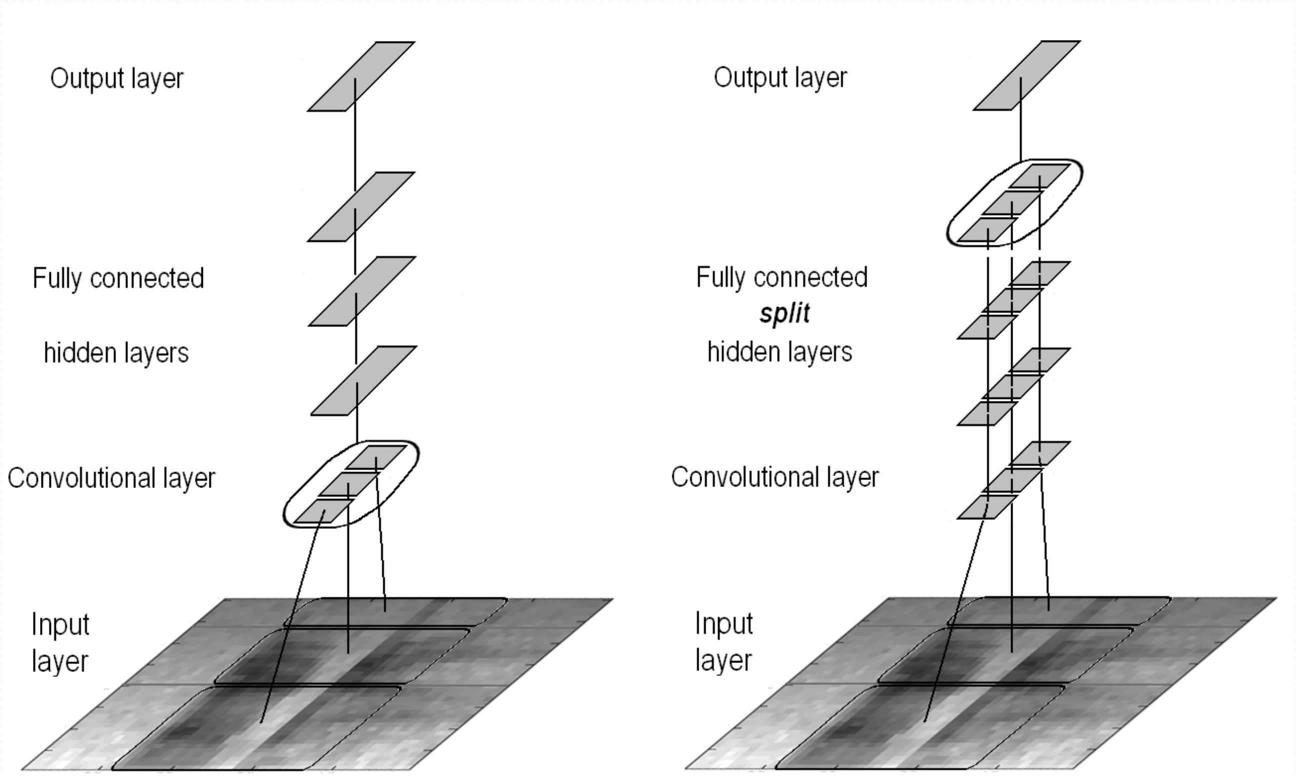
- 40 bands: gets worse beyond 33 frames
- 20 bands: gets worse beyond 49 frames
- 10 bands: stays stable up to 65 frames
- 5 bands: keeps improving with adding more frames

Evaluation – Multi-Resolution Input

Input spectrogram(s)	Context size (frames)	PhER	
		Dev.	Test
40-band spectrum	17	16.2%	18.6%
40+20 bands	17+33	16.0%	17.9%
40+20+10 bands	17+33+49	15.6%	17.5%
40+20+10+5 bands	17+33+49+65	15.6%	17.7%
40+20+10 bands	33+49+49	15.9%	17.3%
40+20+10+5 bands	33+49+49+65	15.8%	17.6%
40+20+10 (17+33+49) plus dropout		15.1%	17.0%

- 40 → 40+20 → 40+20+10 bands: keeps improving
  - But adding the 5-band representation does not help
- Different frame counts (33+49+49+65):
  - These are the optimal sizes from each separate system
  - The observations are similar, no significant improvement
- The best model was trained again with dropout → further improvement

Varying the place of combination



- Left: the model used so far
  - Dedicated convolutional filters for the 4 input types, joint hidden layers
- Right: a model with **split** hidden layers
  - The fusion of information is delayed until the output layer

Varying the place of combination

Place of combination	PhER	
	Dev.	Test
1st fully conn. layer	15.6%	17.5%
2nd fully conn. layer	15.6%	17.6%
3rd fully conn. layer	15.7%	17.5%
Output layer	15.8%	17.5%

- The optimal place for information fusion is not obvious
- Delaying the fusion brought some improvement in the "Split Temporal Context" framework earlier (Tóth, ICASSP 2015)
- However, in this case there was no performance difference between the various models

Summary

- We varied the resolution of the input spectrogram for CNNs
  - This involved both the time and the frequency resolution
- We experimented with combining the various types of input, which resulted in a multi-resolution input
- The best 40-band system (with 33 frames) gave 18.1%, while the best multi-resolution system gave 17.5% accuracy
  - This is a relative improvement of 3.3%
  - Applying dropout, the relative improvement was 4% (17.7%→17.0%)
- We also experimented with splitting the hidden layers, but with no positive outcome

Thank you for your attention!