EE691: RnD

Deep Learning for Prominence Detection in Children's Speech

Mithilesh Vaidya (17D070011)
Guide: Prof. Preeti Rao
Done in collaboration with: Kamini Sabu

Goals

- Predict degree of prominence for each word in an utterance based on prosodic features
- Important intermediate step in assessing oral reading ability of children
- Replace handcrafted features used in [1] with features learnt automatically by CNN

Plan to use it as a component for DPP task

Overview

- DPP goal: Assign a single score for entire recording
- Plan to use Prominence detection as a component
- Transfer learnings from this task to the main goal

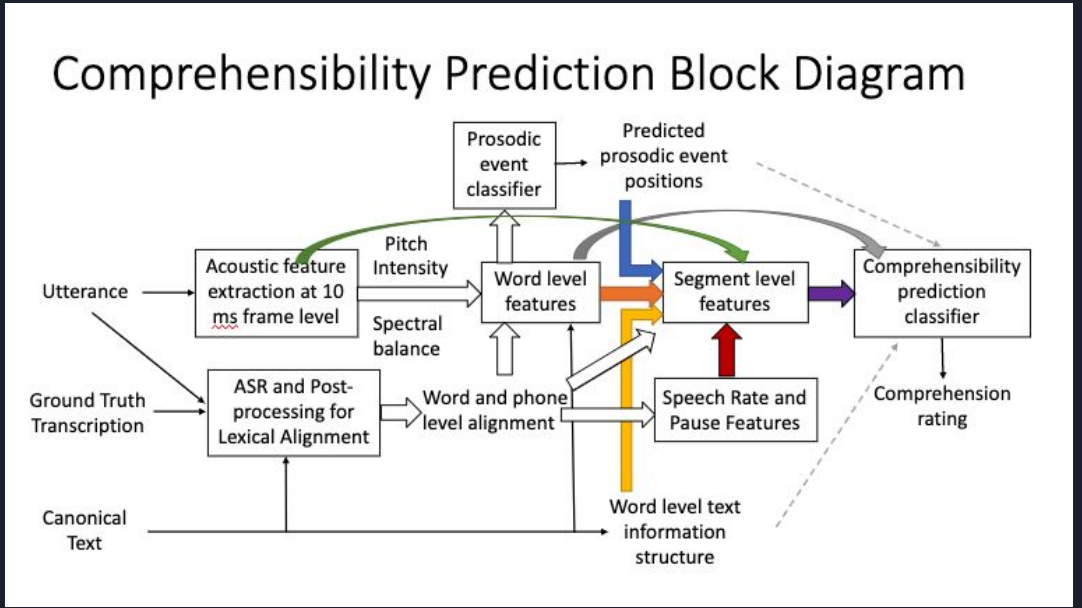


Image Credits: Presentation by Kamini Sabu

Dataset

- 807 recordings or 42,100 words -> 52 words per recording on average
- 35 unique speakers

ASR -> Forced Alignment -> Following features are extracted for each word:

- 159 acoustic features (RFECV analysis) from previous work
- 17 lexical (PoS, phrasing, prominence structure)
- 12 Pause and Duration

For entire recording: 15 acoustic contours per frame

Target: Number of votes (manually rated): between 0 and 7

Evaluation: Pearson correlation between predicted and ground truth score

Approach

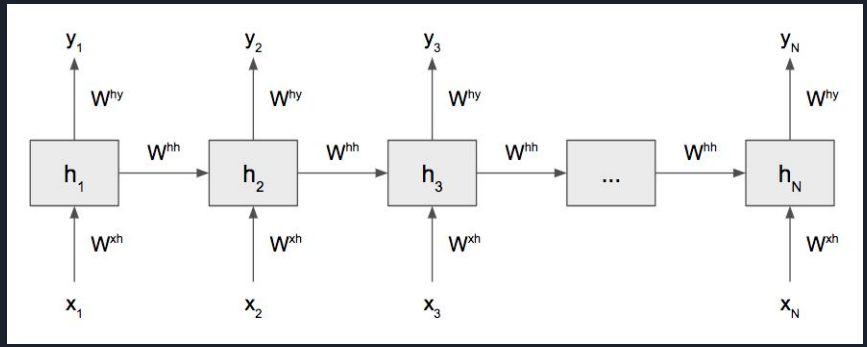
Only RNN:

- Feed one feature vector at each time step and ask RNN to predict regression score
  x_t - feature vector; y_t - scalar normalised score

- Used only word level features
- Tried various combinations of groups
  E.g. RFECV, lexical, RFECV+lexical, etc.

Findings:

- RFECV performed slightly better than Random Forest in [1]
- Adding lexical features gives a huge jump in performance

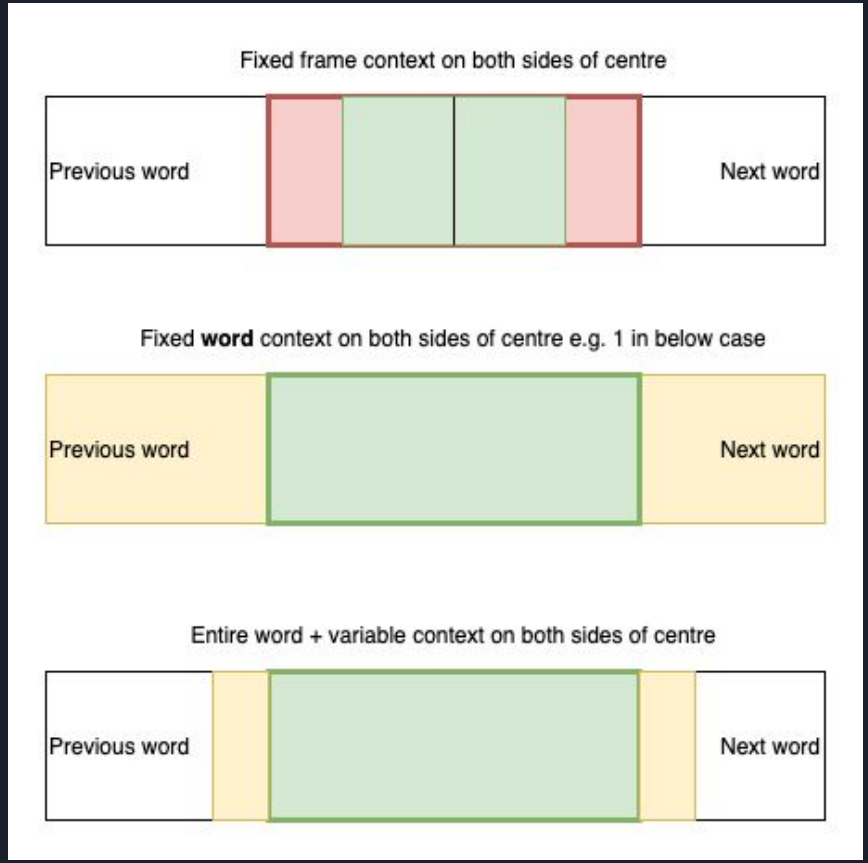


CNN

- Have acoustic contours for entire recording
- How do we extract word-level relevant frames as input to CNN?
- Tried 3 approaches

Typical values:

1. 30, 40
2. 1, 2
3. 15, 20

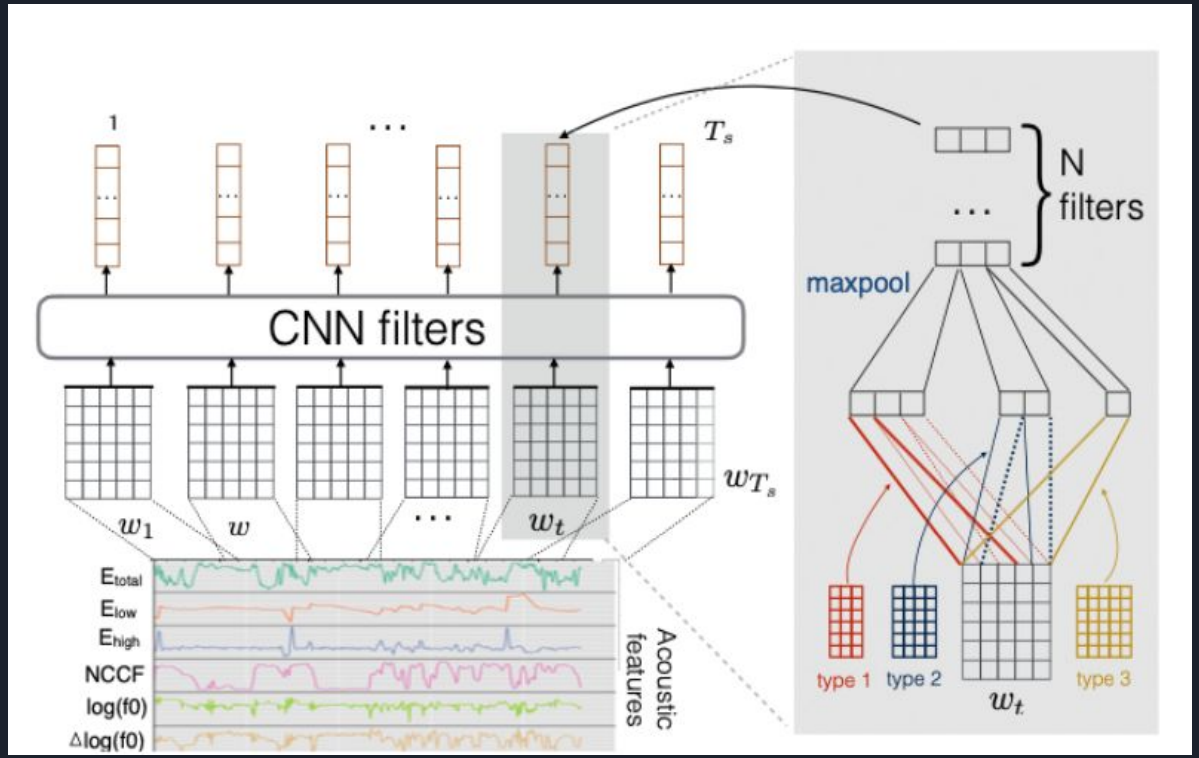


CNN

Features augmented with:

- Positional encoding:
  - 1-bit: boolean for current word [2]
  - 3-bit: 100 for previous, 010 for current, 001 for next
  - 5-bit: pause between words is also considered
- Syllable numbering: Use syllable boundaries and supply 7-d one-hot encoded vector
- Transformer-style positional encoding

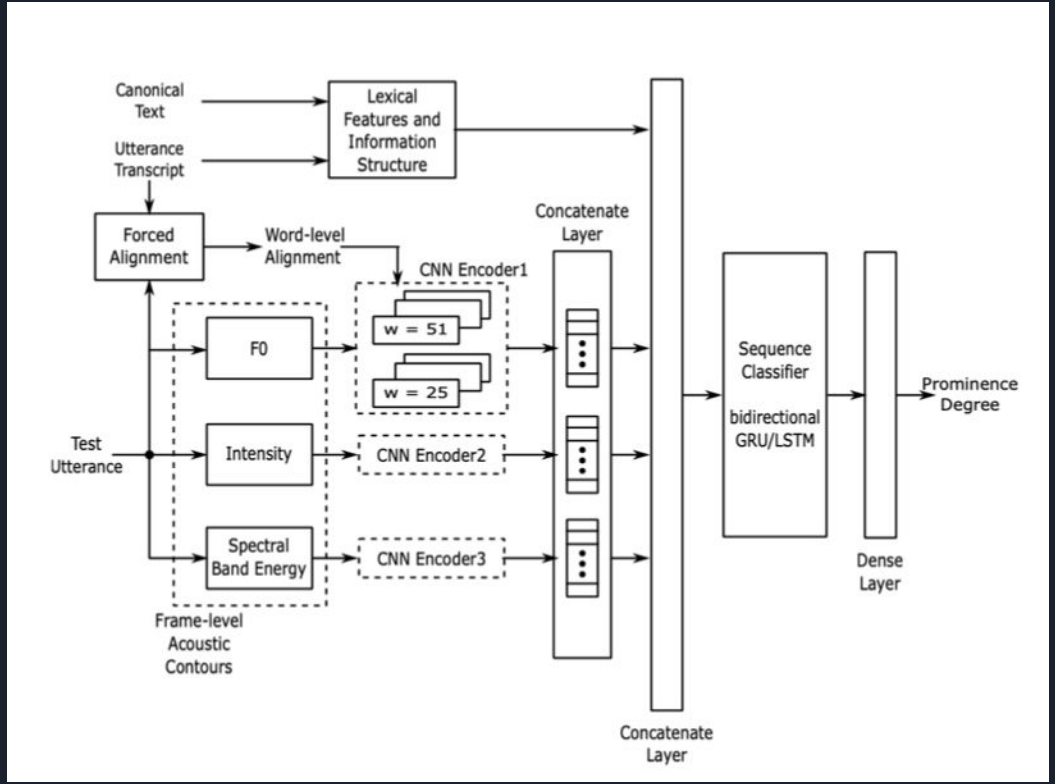
Architecture



From Trang et. al. [4]

Overall architecture

- Separate CNN bank for each group of acoustic contours
- Number of filters per kernel: 8
- Kernel widths: 25 and 51



Training details

- All features are normalised
- Optimiser: Adam/AdamW with LR = 0.003
- NVIDIA GeForce RTX2080 GPU with 12 GB of graphics memory
- Batch size: 500 (based on capacity of GPU)
- MSE between scaled ground truth (0-7 -> 0-1) and sigmoid output from model is minimised
- Pearson correlation used as the early stopping metric
  Test performance every 8 epochs
  If difference below 0.005, test every epoch and early stop on validation
- Mean and standard deviation is reported

Results

Table 1: *Performance of various models with set of 34 acoustic features. (* indicates sd < 0.01)*

Model	# layers	# units	Correlation	F-score
RFC	-	-	0.69*	0.63*
GRU	2	96	0.68	0.63
LSTM	2	256	0.69	0.63
BGRU	2	96	0.70	0.64
BLSTM	2	256	0.71*	0.64*

Table 6: *Performance of CNN encoding concatenated with different word-level features as RNN input. (* indicates sd < 0.01)*

Features	Correlation	F-score
A34	0.70	0.64
CNN	0.69	0.63
CNN + D-P12 + A10	0.71	0.64
CNN + D-P12 + A10 + L + I	0.77*	0.68
A34 + L + I	0.79*	0.69*

Table 2: *Performance with addition of lexical and information structure features. (* indicates sd < 0.01)*

Features	Correlation	F-score
A34	0.70	0.64
A34 + L	0.75*	0.67*
A34 + L + I	0.79*	0.69*

1. RNN slightly better than RF
2. Adding L and I gives big boost
3. A34 ~ CNN

   Last 2 rows -> still some scope for better CNN

Other experiments

Unlabelled data:

- Added high-confidence RF results on unlabelled recordings to the dataset
  Performance dropped
- Speaker embedding:
  - Train a separate model to predict speaker
  - Use both labelled and unlabelled data since we only need speaker identity
  - 167-class classification problem
  - Use bottleneck layer representation as additional feature vector
  - Performance dropped
  - Plausible reason: not learning any new features; same as CNN filters

Future scope

- Different weight initialisation schemes, random seeds and other DL hacks
- Better stopping criterion and training methodologies
- Jointly train speaker and main network (Multi-task learning)

- Multi-modal attention to weigh CNN embedding, word-level features, lexical, etc.
- Semi-supervised and transfer learning for training feature extractors
- Deeper CNNs (Conv + ReLU + BatchNorm) with skip connections
- Replace RNNs with SOTA models like transformers
- Alternate positional encodings

References

[1] Kamini Sabu and Preeti Rao. “Prosodic event detection in children’s read speech”. In: Comput. Speech Lang. 68 (2021), p. 101200. doi: 10.1016/j.csl.2021.101200. url: https://doi.org/10.1016/j.csl.2021.101200.

[2] Sabrina Stehwien and Ngoc Thang Vu. “Prosodic Event Recognition Using Convolutional Neural Networks with Context Information”. In: Proc. Interspeech 2017. 2017, pp. 2326–2330. doi: 10.21437/Interspeech.2017-1159. url: http://dx.doi.org/10.21437/Interspeech.2017- 1159.

[3] “Posistional encoding in Transformers”. In: kazemnejad.com (). url: https://kazemnejad.com/blog/transformer_architecture_positional_encoding/.

[4] T. Tran, S. Toshniwal, M. Bansal, K. Gimpel, K. Livescu, and M. Ostendorf, “Parsing speech: A neural approach to integrat- ing lexical and acoustic-prosodic information,” in Proceedings of NAACL-HLT, New Orleans, Louisiana, 2018.