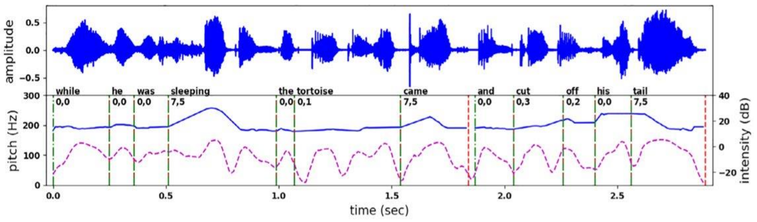
Deep Learning for Prominence Detection in Children's Read Speech

Mithilesh Vaidya, Kamini Sabu, Preeti Rao
Department of Electrical Engineering, Indian Institute of Technology Bombay, India

Motivation

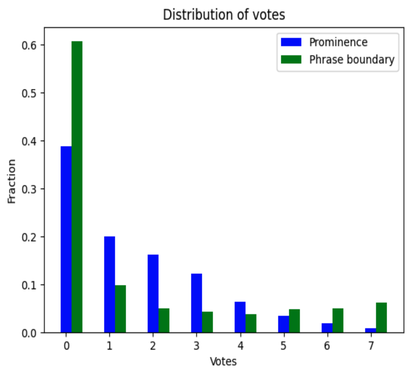
- Humans convey rich non-phonetic information during speech delivery
- Prosody – **suprasegmental** attributes of speech which convey both linguistic and para-linguistic information
- Expressiveness - important prosodic component of spoken language fluency which conveys **linguistic** information
- Proficient readers are good at:
  - Emphasis - Stress on the right words to convey novel info.
  - Phrasing – group words together to convey syntactic info. Phrasing: Spoken language :: Punctuation: Text
- Pitch, intensity, duration – acoustic correlates of prosody
- Typical changes: word lengthening, within-word pitch variation



- Words *sleeping*, *came* and *tail* made prominent by increasing duration (waveform) and varying pitch (blue contour)

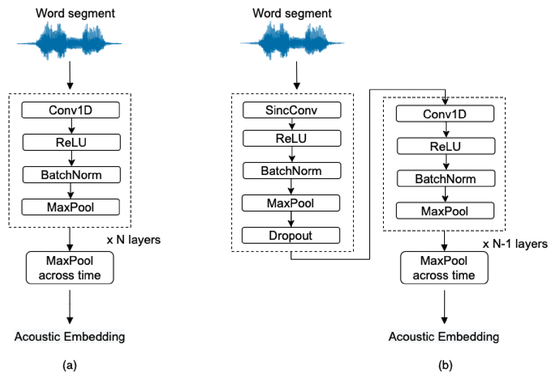
Dataset

- Goal: Predict oral reading fluency of school children
- 41,286 words across 790 utterances of read stories (~52 words/utterance)
- 4 hours 20 minutes of speech at 16 KHz sampling rate
- 35 (middle-school) speakers
- Each word rated for presence/absence of prominence and boundary by 7 naive listeners using RPT methodology (Cole et al., 2017)



- Votes scaled to 0-1; goal is to predict degree of prominence for each word
- Dataset split into 3 speaker non-overlapping folds
- Evaluation: Pearson correlation coefficient

Waveform-based feature extraction



(a) Standard stack of CNN layers:

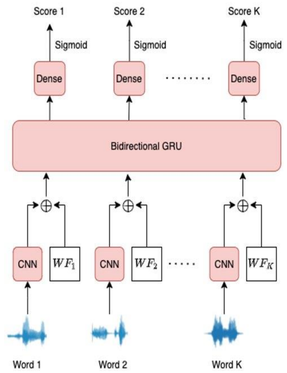
- Extract features directly from word segments
- Optimal conf.: Layers: 4, Kernel width: 51, Pool kernel: 3, Stride: 1, Filters: 32

(b) Additional Sinc layer:

- Sinc convolution (Ravanelli et al., 2018) for constrained band-pass filtering
- Optimal Sinc conf.: Kernel width: 31 , Stride: 2, Pool kernel: 3, Filters: 32

CRNN architecture

- WF: optional word-level features e.g.
  - A34 hand-crafted features (Sabu et. al. 2021)
  - Word embeddings such as *GloVe*
  - Lexical features such as *Part of Speech* tags
  - Information Structure: top-down expectations based on canonical text

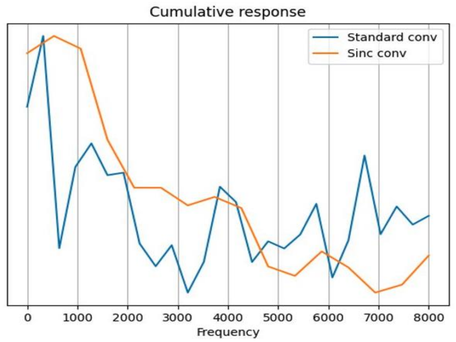


- 3-layer, 256-dim. BGRU for modelling utterance-level dependencies
- Fully-connected layers (512 -> 128 -> 1) + Sigmoid for final prediction
- Loss: MSE
- LR: 0.001, Batch Size: 64, Optimizer: Adam

Results

No.	Input	Acoustic Model	Layer 1 (type, width, stride)	Pearson correl.
1.	A34	BGRU	-	**0.726**
2.	Wav	CRNN	Standard, 51, 1	0.692
3.	Wav	CRNN	Sinc, 51, 1	0.712
4.	Wav	CRNN	Sinc, 31, 2	**0.721**
5.	A34 + Wav	CRNN	Sinc, 31, 2	0.735

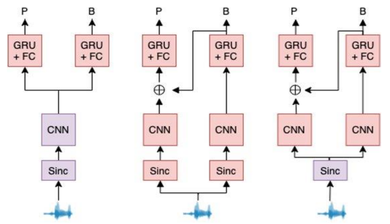
- 1: Handcrafted features baseline performance
- 2->3: Benefit of Sinc layer at input
- 3->4: Tuning of Sinc width and stride
- 4->5: Complementary information in A34 and CNN-based feature extraction



- Both capture peaks around: 200 Hz (Pitch) and 1100 Hz (First formant)
- Standard conv response is noisy as compared to Sinc -> sign of overfitting
- Sinc does a better job at capturing overall spectral envelope

Multi-task Learning

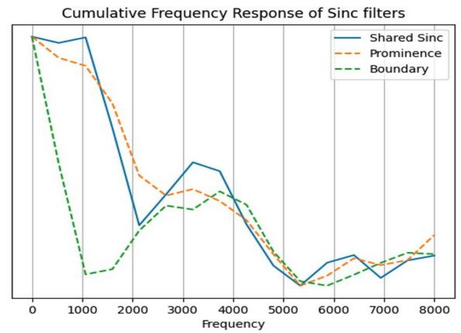
Exploit dependencies between prominence and phrase boundary



- Knowledge of phrasing is a reliable cue for prominence (applicable for English)
- Sharing of low-level feature extractors reduces overfitting
- Loss = Convex combination of prominence MSE and phrasing MSE loss

No	MTL variant	Pearson correl.
1.	Tuned Sinc (Without MTL)	**0.721**
2.	Shared Sinc	0.727
3.	Conditioned	0.727
4.	Shared Sinc + Conditioned	**0.740**
5.	Row 4 + A34 + A27	0.757
6.	Row 5 + GloVe	0.813

- 1 -> 2/ 1-> 3: Marginal improvement
- 1 -> 4: Noticeable improvement due to combination of sharing Sinc and conditioning
- 4 -> 5: Complementary info. in A34 and Wav
- 5 -> 6: Complementary info. in lexical and acoustic



- Shared Sinc closely follows Sinc for prominence
- Sinc for boundary seems to be only capturing a peak near 3500 Hz
- By sharing Sinc, boundary predictions improve since pitch and intensity are crucial for phrase boundary

Conclusion

- Constrained (Sinc) filters better than unconstrained kernels which overfit on our task and dataset
- Conditioning on phrase boundary in the presence of a shared Sinc layer boosts performance
- Significant complementary information in lexical features such as word embeddings and POS tags

References:

1. Sabu, Kamini, and Preeti Rao. "Prosodic event detection in children's read speech." Computer Speech & Language 68 (2021): 101200.
2. Cole, Jennifer & Mahrt, Tim & Roy, Joseph. (2017). Crowd-sourcing prosodic annotation. The Journal of the Acoustical Society of America. 141. 3910-3910. 10.1121/1.4988814.
3. Ravanelli, Mirco, and Yoshua Bengio. "Speaker recognition from raw waveform with sincnet." *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018.