Sinc Net for speaker Recognition

Jagabandhu Mishra

IIT Dharwad

outline :-

1. speaker recognition ✓
2. Earlier approaches ✓
3. End-to-End approaches
4. Proposed approach
5. Experimental setup
6. Result & discussion

1. Speaker Recognition

- Speaker Recognition
  - Speaker Identification (SI)
  - Speaker Verification (SV)

This approach

- SI :- Identify from N speakers
- SV :- Verify if genuine speaker or not (true/false)

2. Earlier approaches

- speech signal → Feature extraction (of E) → Modeling → decision
- FE → MFCC largely used

⇒ Why it is good for Speaker Recognition ??

2. Feature approaches (continue)

- modeling :- Statistical | Neural network

GMM

GMM-UBM - I-vector

(LSTM, GRU) → vector

(X-vector)

→ ANN
→ RBM
→ FFNN
→ CNN
→ RNN
→ Attention TDNN

How to design an end-to-end network to perform SR ??

2. End-to-End approaches

- filterbank → CNN - (H-U)
- spectrogram → CNN - (don't know what learned)
- Raw speech → CNN

- will all this captures appropriate info ?? (for SR). (Ans:- we don't know)

- Intuition:- Can we Put some constraint on the CNN filter. ??

4. Proposed approach

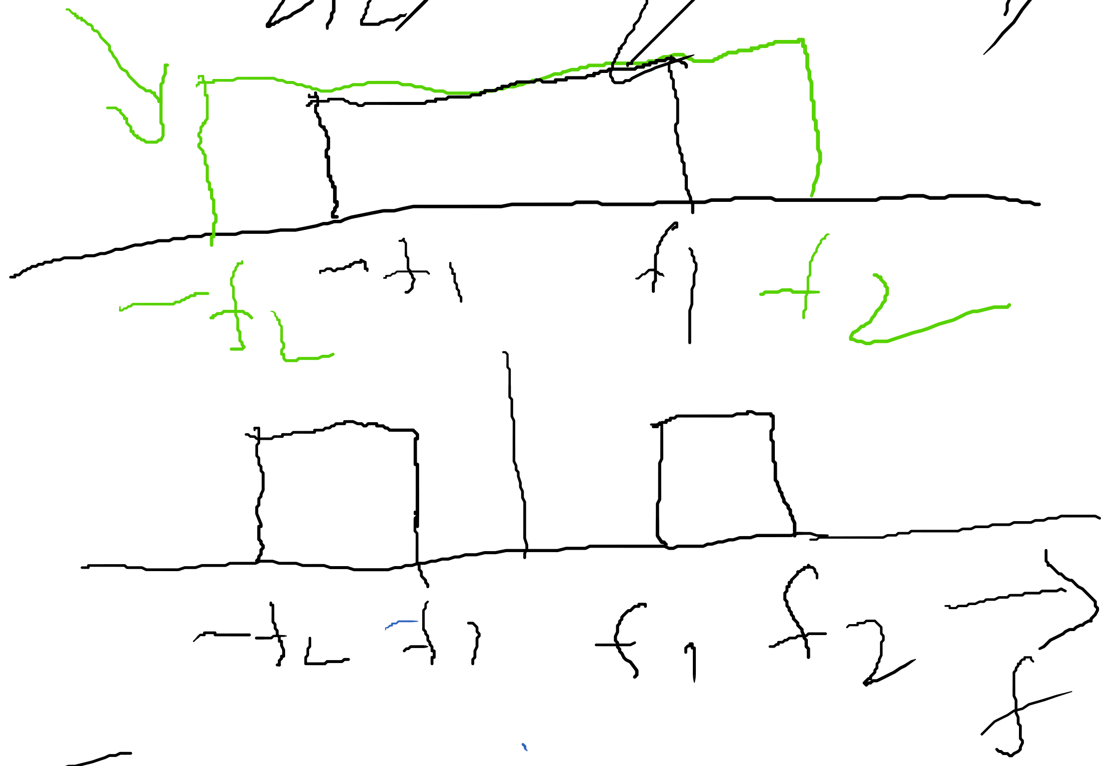
- Objective:- capture task specific info. from data.
- See fig-1 of ref. paper.
- 1st layer:- used as a data driven F.E.
  - ✓ Public contains conv filters only learns higher and lower (both frequency)
- See En-1 to 4 (ref paper)

continue !

eqn 1 !-

$$G(f; f_1, f_2) = rect\left(\frac{f}{2f_2}\right) - rect\left(\frac{f}{2f_1}\right)$$

=



- ⊗ will this feasible to set (finite) (answer)
- ⊗ $f_2 > f_1$ (how)
- ⊗ initialization ??

IFT

$$g(n; f_1, f_2) = 2f_2 \, sinc(2\pi f_2 n) - 2f_1 \, sinc(2\pi f_1 n)$$

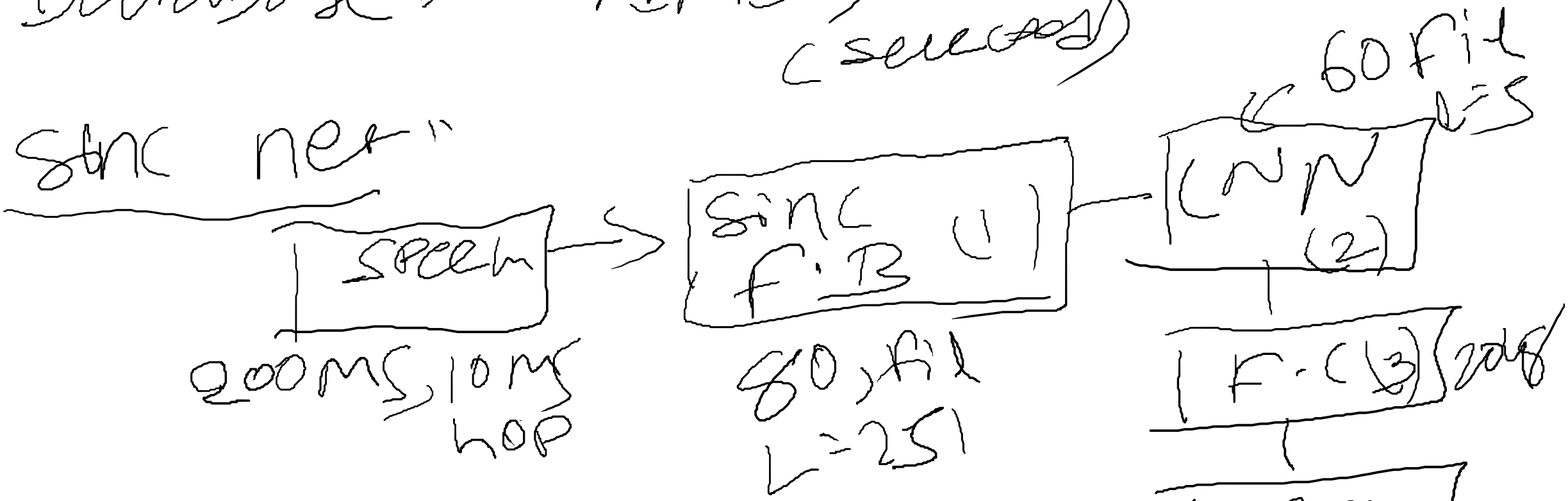
$-\infty < n < \infty$

Continue -

- $f_2 > f_1$ → eqn 5 2 6
- finite length :- eqn 7 2 8
- Initialization :-
  - Random (0 - $f_s/2$)
  - mel iten
- Expected outcome :-
  - fast convergence
  - few Parameter (CFL vs 2F)
  - → Interpretable

5. Experimental setup :

- Database :- TIMIT, LibriSpeech (selected)
- Structure :-



- Initial :- Glorot
- Optimizer :- RMSPROP, $\eta = 0.001$, $\alpha = 0.95$, $\epsilon = 10^{-7}$, 128 mini batch size

6. Result & Discussion :-

- see fig 2, 3 & 4 (titer analysis)
- S.I & S.V. Result & discussion
  - ↳ section - 5
- Future direction :-
  - → can take other titers (ingredients)
  - → improve models after F.E
  - → other application (screen based)


Thank you

② Reference! -

Speaker Recognition from raw waveform with SincNet

by

M. Ravanelli, Yoshua Bengio

pioneer at DL, along with Prof. Hinton