BRNO UNIVERSITY OF TECHNOLOGY
speechfit

BUT “GOT-HIM” ASR system for Indian Evaluation 2021

Martin Karafiat, Murali Karthick Baskar, Karel Vesely
….

Agenda

- Title Introduction
  - **GOT-HIM: G**ujarat, **O**dia, **T**amil, Telugu, **Hi**ndi, **M**arathi
- System description
- LM data
- Baseline system analysis
- Conclusion

System

- Kaldi based hybrid DNN-HMM LF-MMI models
- DNN feature extraction was based on 2 concatenated feature streams resulting in 240 dimension vector:
  1. 40dim f-bank output
  2. Adaptation vector -> Affine transform 200dim
- Architecture:
  - 6CNN+19TDNNF layers
  - CNN: 64, 64, 128, 128, 256, 256 filters
  - TDNN-F - 19x (1536-160-1536 neurons)

System II

- Training style
  - Multilingual training on all challenge languages
  - Multilingual NN -> fine-tuning to the target language

- Adding Wikipedia data
- Language ID based on utterance based MBR confidences

System adaptation

- NN adaptation based on appending auxiliary vector to input features
- Two approaches:
  - Standard online i-vectors - 100 dimensions
  - x-vectors - 512 dimensions
- x-vectors
  - Training: extended TDNN architecture, SRE 2016 data used for training
  - Application: x-vectors are extracted on utterance basis for all challenge data and cloned according feature frames.

LM data

- Found significant amount duplicate sentences in the training data
  - 350k -> 10k
  - Odia: 60k -> 0.8k
  - training data uniq
- Download Wikipedia data
  - 7.5M of new sentences: (0.7M BE, 0.4M GU, 0.7M HI, 0.66M MA, 0.1 OR, 2M TA, 2.9M TE)
- Interpolated LM based on test set perplexity.
- Language ID based on utterance based MBR confidences

Multilingual systems

- AM + LM + test data pooled together

System	test [%WER]
Multilingual	30.6
+ WB	30.5
++ xvec	30.2
+++ wikiLM	20.3
++++ suniqLM	**18.2**

Monolingual systems

- MBR - LID: decoding with all lang.spec. models and pickup the most confidence one.

AM	LM	GU	HI	MR	OR	TA	TE
Multilingual	Mult	14.7	18.8	10.5	30.3	17.4	18.1
Monolinguual	Mono (oracle LID)	12	19.2	11.3	30.4	16.7	17.8
Multilingual	Mono (oracle LID)	13.7	18.3	9.3	28.1	17.2	17.6
Fine-Tuned from multilingual NN	Mono (oracle LID)	12.1	17.3	8.4	27.8	16.8	17.2
	Mono (MBR LID) - Submitted	13.2	17.8	11.0	30.5	19.2	20.2

- Completely multilingual systems are doing good LID job
- Monolingual system are better but LID needed (11% Err)
- No gain from system fusion

Leaderboard

#	Team Name	Hindi (% WER)	Marathi (% WER)	Oriya (% WER)	Tamil (% WER)	Telugu (% WER)	Gujarati (% WER)	Average (% WER)
1	CSTR	14.33	15.79	25.34	23.16	21.88	20.59	20.18
2	Bytedance-SA	16.59	15.65	17.81	28.59	25.37	21.3	20.89
3	EthereumMiner	17.54	20.15	19.99	28.52	26.08	20.11	22.06
4	Uniphore	22.79	14.9	29.55	18.8	28.69	22.79	22.92
5	GOT-HIM	17.18	18.48	29.99	29.66	28.74	21.79	24.31
6	GoVivace	21.77	25.73	29.05	28.92	26.5	21.22	25.53
7	Ekstep	12.24	39.74	27.1	27.2	22.43	30.65	26.56
8	TUTU	19.93	26.52	34.18	27.69	30.25	25.34	27.32
9	TCS-SpeechNLP	19.77	37.45	35.21	26.26	26.82	28.53	29.0
10	Lottery	17.81	58.78	17.74	30.69	27.67	23.62	29.39
11	IIITHSPL	31.11	33.8	37.19	35.03	17.0	26.94	30.18
12	ScribeTech	27.78	33.05	34.57	33.01	30.08	28.22	31.12
13	Dialpad	21.49	46.41	32.13	28.6	28.03	34.57	31.87
14	Baseline	37.2	29.04	38.46	34.09	31.44	26.15	32.73

Conclusion

- Upsampling is slightly better than downsampling
- x-vectors overcome i-vectors
- Significant boost from external wikipedia LM data
- Significant degradation by confidence based LID
  - proper LID experiments
- Missing RNN-LM (submitted after deadline)