AUTOMATIC SPEECH RECOGNITION FOR LOW-RESOURCE LANGUAGES: THE THUEE SYSTEMS FOR THE IARPA OPENASR20 EVALUATION

*Jing Zhao, Guixin Shi, Guan-Bo Wang, Wei-Qiang Zhang**

Beijing National Research Center for Information Science and Technology
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

清華大学
Tsinghua University

CONTENTES

Introduction

IARPA Open Automatic Speech Recognition Challenge (OpenASR20)

To assess the state-of-the-art ASR technologies for low-resource languages.

- Amharic
- Cantonese
- Guarani
- Javanese
- Kurmanji-Kurdish
- Mongolian
- Pashto
- Somali
- Tamil
- Vietnamese

- The second open challenge created out of the Intelligence Advanced Research Projects Activity **(IARPA)**
- Machine Translation for English Retrieval of Information in Any Language **(MATERIAL)** program
- A track of **NIST**'s OpenSAT (Open Speech Analytic Technologies) evaluation series

- **Metrics**: Word Error Rate (WER) (Formats: STM; CTM)
- **Training Conditions**: Constrained & Unconstrained
- **Data Resources**:

Modality	Build (training), Constrained	Build (training), Unconstrained
Audio	10h	unlimited
Text	unlimited	unlimited

Introduction

IARPA Open Automatic Speech Recognition Challenge (OpenASR20)

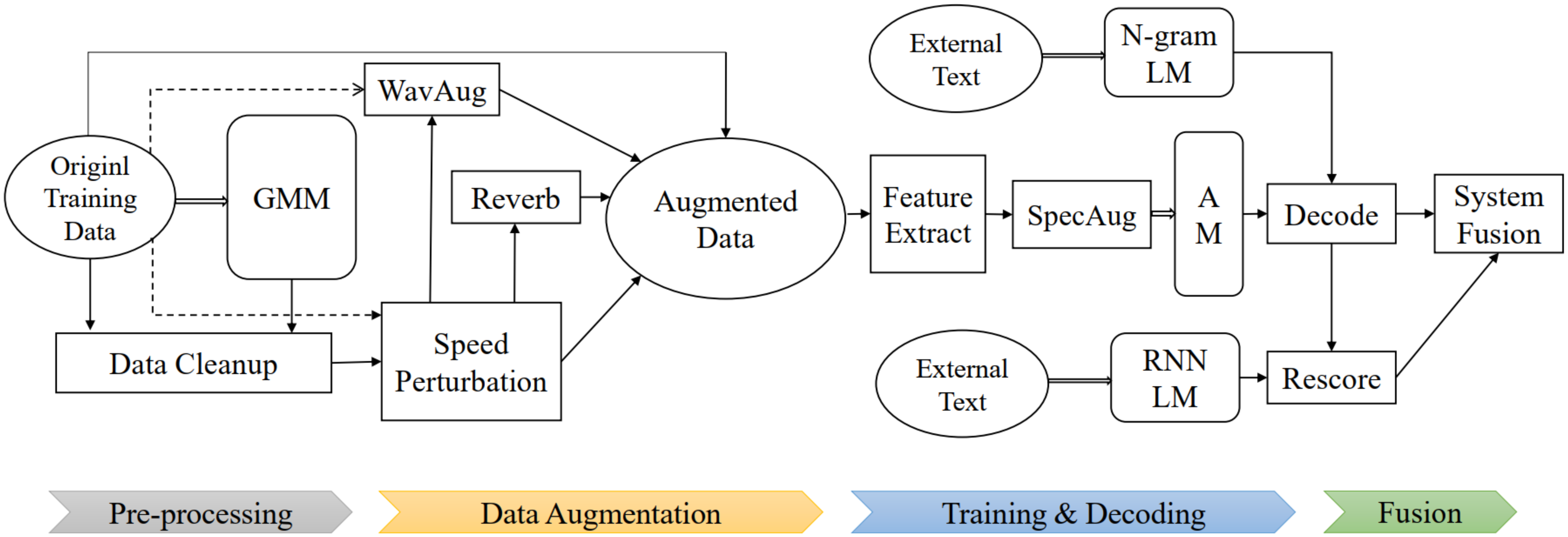
To assess the state-of-the-art ASR technologies for low-resource languages.

- Amharic
- Cantonese
- Guarani
- Javanese
- Kurmanji-Kurdish
- Mongolian
- Pashto
- Somali
- Tamil
- Vietnamese

Constrained training condition

- **Data usage:**
  - Train: **10 hour** audio
  - Train: unlimited text (provided or publicly available)
  - DEV (for system development only): 10 hour
  - EVAL (for system evaluation): 5 hour
- **Datasets:**
  - Somali → IARPA **MATERIAL** program
  - The others→ IARPA **Babel** program
  - Conversational telephone speech, in separate channels for each speaker
  - Sampled at **8kHz, 44.1kHz, or 48kHz**

Workflow


Originl Training Data
GMM
WavAug
Reverb
Data Cleanup
Speed Perturbation
Augmented Data
External Text
N-gram LM
Feature Extract
SpecAug
A M
Decode
System Fusion
External Text
RNN LM
Rescore
Pre-processing
Data Augmentation
Training & Decoding
Fusion

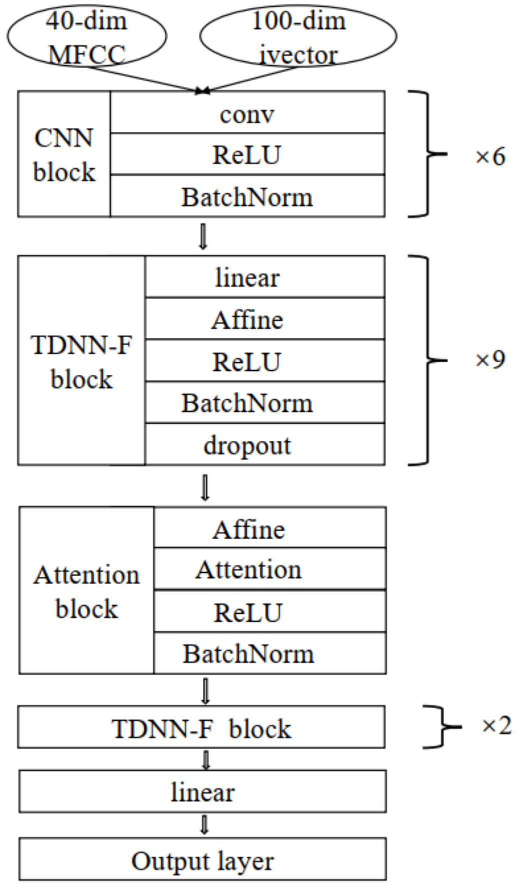
Acoustic model: CNN-TDNNF-A architecture

CNN-TDNNF-A

- **CNN:**
  The numbers of filters:
  48, 48, 64, 64, 64, 128
- **TDNN-F:**
  11 blocks; dimension 768;
  bottleneck dimension 160
- **Self-attention mechanism:**
  Attention heads: 20
  Key-dimension: 8
  Value-dimension: 16

Table 1. Word Error Rate (WER%) on DEV for different contexts in self-attention layer. *Non* means the self-attention layer is NOT added. *Fusion* are the fused results from the first four systems by ROVER

Language	Non	[-15;6]	[-12;9]	[-9;9]	Fusion
Guarani	53.3	53.5	53.3	53.1	51.8
Pashto	54.9	54.8	54.7	54.9	52.2
Vietnamese	51.7	51.7	51.8	51.6	49.8
Average	53.3	53.3	53.3	53.2	51.3



Acoustic Model and Language Model

Tsinghua University

Language model: N-gram

RNNLM rescore: TDNN-LSTM

Text data from web crawling don't help

Table 2. The extra used texts in IARPA Babel program.

Language	Dataset ID	#words
Amharic	IARPA-babel307b-v1.0b-build	281k
Cantonese	IARPA-babel101b-v0.4c-build	892k
Guarani	IARPA-babel305b-v1.0c-build	311k
Javanese	IARPA-babel402b-v1.0b-build	309k
Kurmanji	IARPA-babel205b-v1.0a-build	346k
Mongolian	IARPA-babel401b-v2.0b-build	403k
Pashto	IARPA-babel104b-v0.bY-build	888k
Tamil	IARPA-babel204b-v1.1b-build	486k
Vietnamese	IARPA-babel107b-v0.7-build	923k

Table 3. WERs on DEV set with different LMs.

Language	LM	Babel LM	Rescore
Amharic	51.3	49.7	53.7
Cantonese	51.5	49.3	48.1
Guarani	52.4	50.5	49.8
Javanese	59.6	58.2	58.0
Kurmanji-Kurdish	69.6	67.5	67.4
Mongolian	55.6	52.5	43.6
Pashto	53.0	49.9	48.8
Somali	59.6	-	58.7
Tamil	71.5	70.1	69.0
Vietnamese	51.8	48.8	48.2
Average	**57.4**	**55.2**	**54.5**

Data processing:

- Cleanup
- Augmentation
  - Speed and Volume Perturbation (SVP)
  - SpecAugment (SA)
  - Reverberation (Reverb)

Speech activity detection (SAD): CRNN & RNN

System fusion: ROVER

Results filtering: filter the word lists by the corresponding degree of confidence

Table 4. WERs on DEV set wi/wo data cleanup.

Language	Non-cleanup	Cleanup
Amharic	49.0	48.7
Cantonese	49.2	48.3
Somali	59.6	59.2
Average	**52.6**	**52.1**

Table 5. WERs on DEV set with different data augmentations.

Language	Non	+SA	+SVP+SA
Guarani	57.9	53.6	50.2
Javanese	65.6	61.0	57.8
Pashto	60.1	54.8	49.8
Average	**61.2**	**56.5**	**52.6**

Challenge results:

For EVAL, the results are released by NIST OpenASR scoring server

For DEV, the WERs are obtained by sclite with Reference File Format (STM) generated locally (take some non-scored words into account)

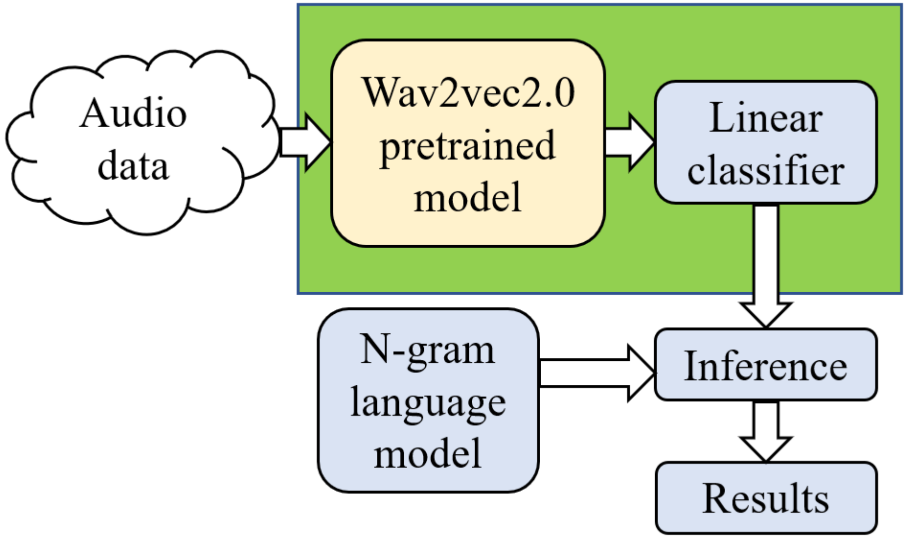
Table 6. Results of the ASR systems on 10 languages. The numbers in the brackets mark our rankings for the Evaluation.

Language	DEV	EVAL	
	Single	Single-best	Fusion-best
Amharic(2)	48.7	46.2	45.8
Cantonese(2)	48.3	44.9	43.6
Guarani(1)	49.9	48.0	46.1
Javanese(1)	57.1	54.5	52.1
Kurmanji -Kurdish(2)	67.1	68.6	66.9
Mongolian(1)	52.4	48.1	45.4
Pashto(2)	49.6	50.3	48.6
Somali(2)	59.2	60.0	59.6
Tamil(2)	69.1	67.5	66.0
Vietnamese(2)	48.8	47.8	46.0
Average	**55.0**	**53.6**	**52.0**

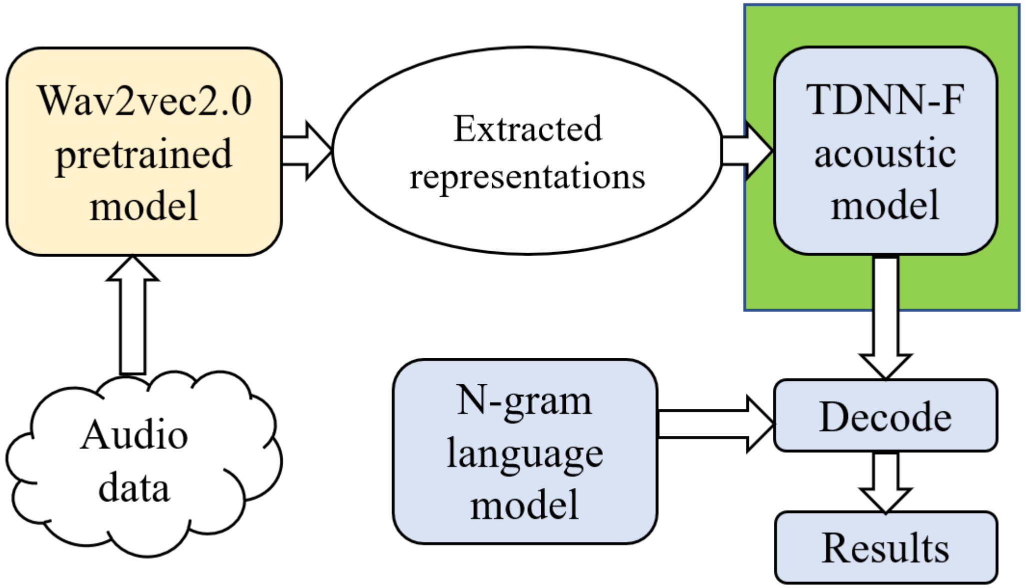
https://www.nist.gov/itl/iad/mig/openasr20-challenge-results

Unsupervised pretrained model: Wav2vec2.0

A Framework for Self-Supervised Learning of Speech Representations



Pipeline of E2E ASR with wav2vec2.0.



Pipeline of hybrid ASR with wav2vec2.0 representations.

Improvements:

XLSR-53 (wav2vec2.0 pretrained model):
A multilingual model trained with 56k hours audio data in 53 different languages

Unlabeled audio from target language
Train the pretrained model first

Frame shift mismatch (pretrained model 20ms & hybrid system 10ms)
Copy features from last frame / directly change the stride in pretrained model

Feature extraction & augmentation
Extract features from different layers
SpecAugment & speed perturbation

Experiments:

Pashto: 10h labeled data, 68h unlabeled audio

Table 7. WERs on DEV set of Pashto with pretrained models. *CTC* in Downstream stands for directly inferring with fairseq while hybrid means training AM in hybrid system with the extracted features by Kaldi. *FeaturePro (cessing)* lists the two ways to solve the frame-shift problem.

Fine-tune	Downstream	Feature Pro	WER
FT1	CTC	-	48.0
FT1	hybrid	chang stride	52.0
FT1	hybrid	copy	48.4
FT2	CTC	-	**44.5**
FT2	hybrid	copy	45.9

Table 8. WERs on DEV set of Pashto with pre-trained models. *Extracted* explains which layer of the Transformer in the pretrained model the presentations are extracted from. *SA* means SpecAugment and *SP* means Speed Perturbation.

NO.	Extracted	AM Layer	Data Aug	WER
1	layer 6	11	SA	49.5
2	layer 9	11	SA	47.5
3	layer 12	11	SA	46.3
4	layer 18	11	SA	45.1
5	layer 21	11	SA	45.2
6	layer 24	11	-	45.9
7	layer 24	11	SA	45.2
8	layer 24	11	SP	44.8
9	layer 24	11	SA;SP	**44.1**
10	layer 24	9	SA;SP	45.3
11	layer 24	13	SA;SP	44.3
12	System Fusion			**41.7**

IARPA Open Automatic Speech Recognition Challenge (OpenASR20)

- CNN-TDNNF-A acoustic model
- Domain-matched text V.S. mismatched text data.
- Data augmentation methods are especially necessary and effective in low-resource conditions.

- Hybrid ASR system with pretrained model: wav2vec2.0-to-Kaldi pipeline
- Target domain: unsupervised & supervised
- Frame-shift mismatch
- Extracting layer position and data augmentation

清華大學
Tsinghua University

Q & A

THANK YOU FOR YOUR WATCHING

清华语音与音频技术实验室

更多精彩

欢迎扫码关注THUsatlab