TOSHIBA

Toshiba's speech recognition system for the CHiME 2020 Challenge

Cătălin Zorilă[1], Mohan Li[1], Daichi Hayakawa[2], Min Liu[3], Ning Ding[2], Rama Doddipatla[1]

[1]Toshiba Cambridge Research Laboratory, UK
[2]Toshiba Corporate R&D Center, Japan
[3]Toshiba China R&D Center, China

May 5th, 2020

CHiME 2020 Virtual Workshop

Presentation outline

- System overview
- Front-end
- Acoustic model
- Language model
- Results
- Conclusions

System overview

- Toshiba entry for the multi-array speech recognition task
  - speaker diarization information is provided (**Track 1**)

- **Conventional** (not end-to-end) **HMM-DNN ASR**

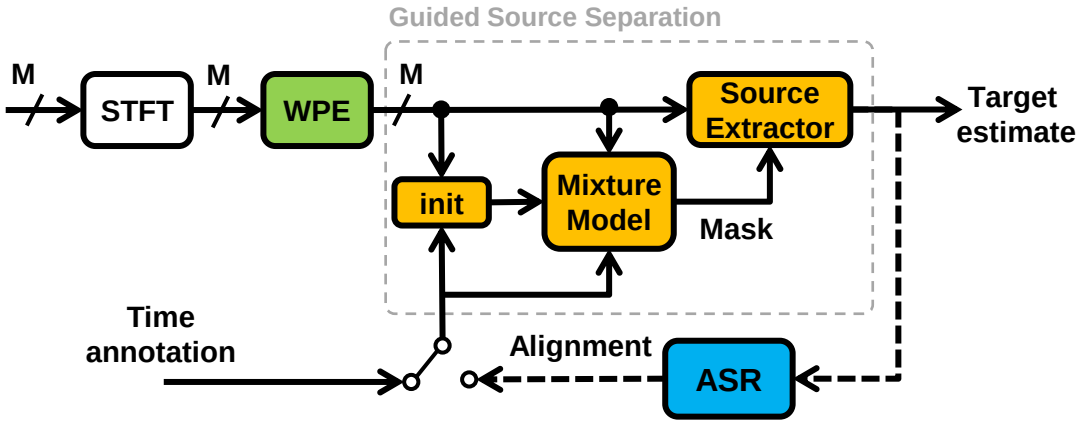
- **Front-end**
  - enhancement: dereverberation (WPE), source separation
  - acoustic features: FBANK, excitation based features, i-vectors

- **Acoustic model**
  - 4 distinct AM topologies, LF-MMI
  - combination of CNN (w/ or w/o residual connections) and TDNNF layers
  - speed perturbation, discriminative training, spk. normalization (VTLN)

- **Language model**
  - baseline 3-gram (**Cat. A**)
  - neural network based (**Cat. B**)

Front-end (1/2)

- **Enhancement**
  - followed enhancement strategy for training and test in [4]
  - unprocessed worn data (in training)
  - dereverberation: Weighted Prediction Error (WPE) [1, 2]
  - source separation: Guided Source Separation (GSS) [3, 4]
    - M=12 (train, test)
    - M=24 (test)



Front-end (2/2)

- **Acoustic features**

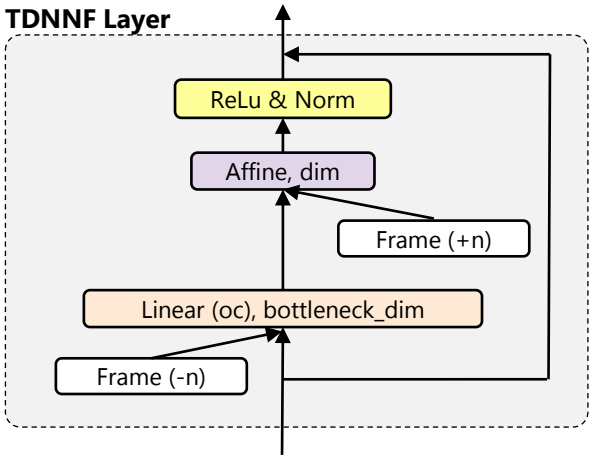
  - Mel filter-bank features (64-dim)

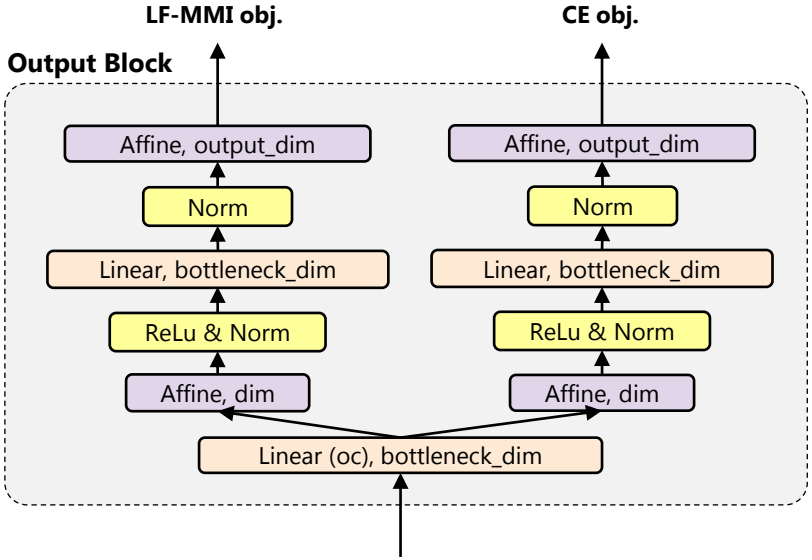
  - excitation-based features (EXF, 10-dim) [5]

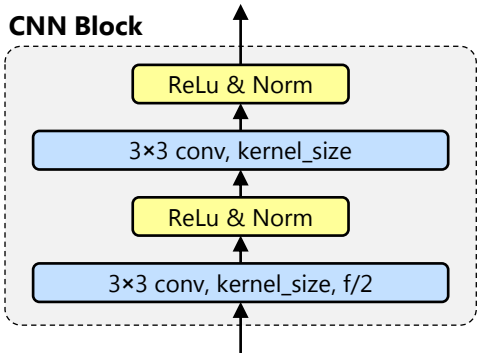
    - zero-crossing rate, height of the auto-correlation function, depth of average magnitude difference function, normalized linear prediction error, residual skewness and kurtosis, maximum of the harmonic product spectrum, 2 features based on the summation of the residual harmonics, cepstral peak prominence

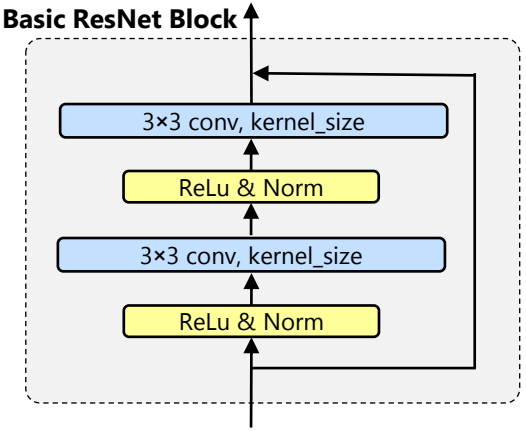
  - i-vectors (100-dim)

Acoustic model (1/4)

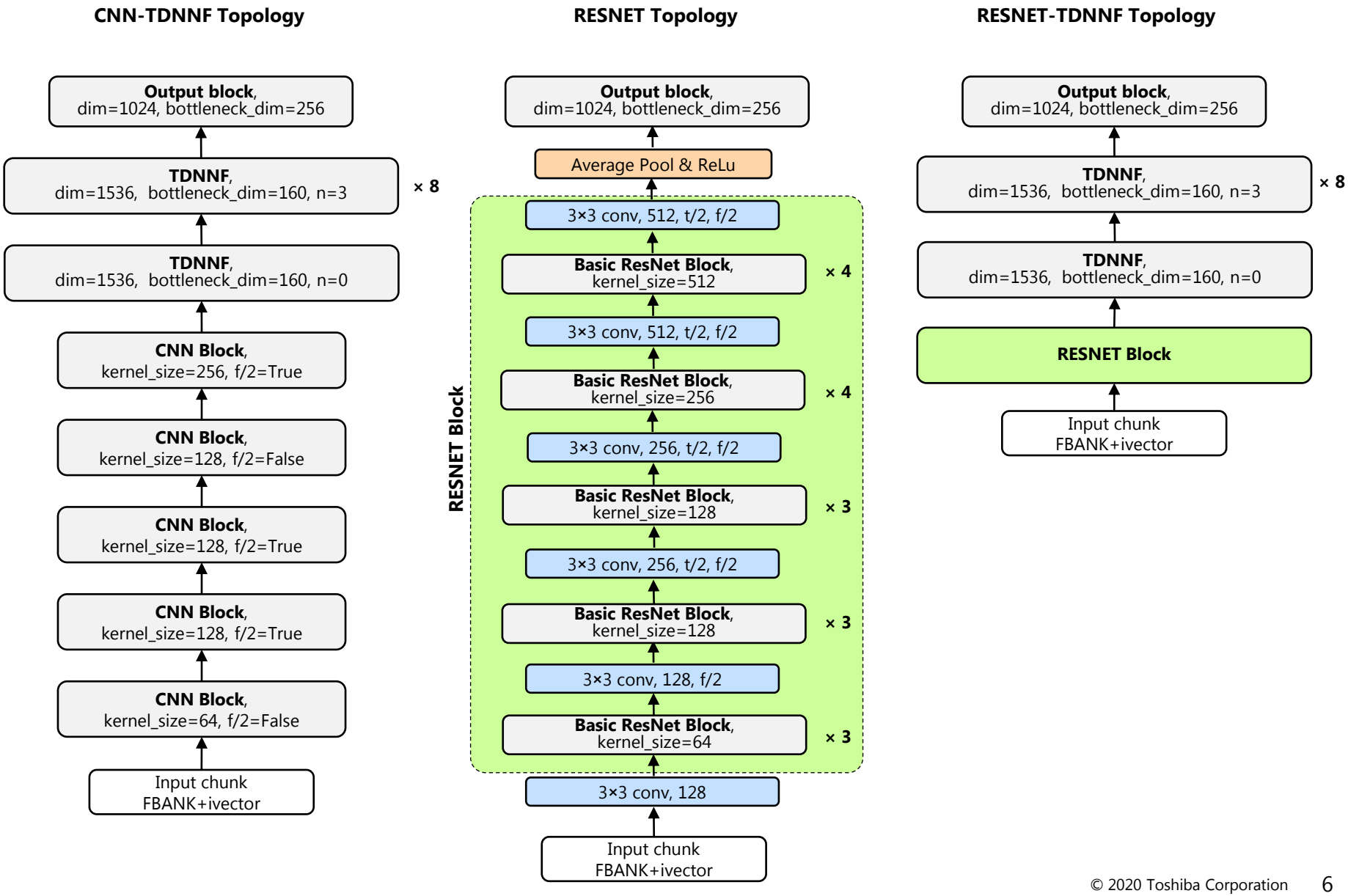








Acoustic model (2/4)



Acoustic model (3/4)

- **Training**
  - cleaned up training data (WPE, GSS) [4]
  - data augmentation: 3-fold speed perturbation
  - training criterion: LF-MMI
  - discriminative training (DT) on top of LF-MMI models
  - Vocal Tract Length Normalization (VTLN)
    - speaker normalization
    - warp factors are estimated for each speaker and room condition
  - 2-pass decoding with refined i-vector refinement during testing

Acoustic model (4/4)

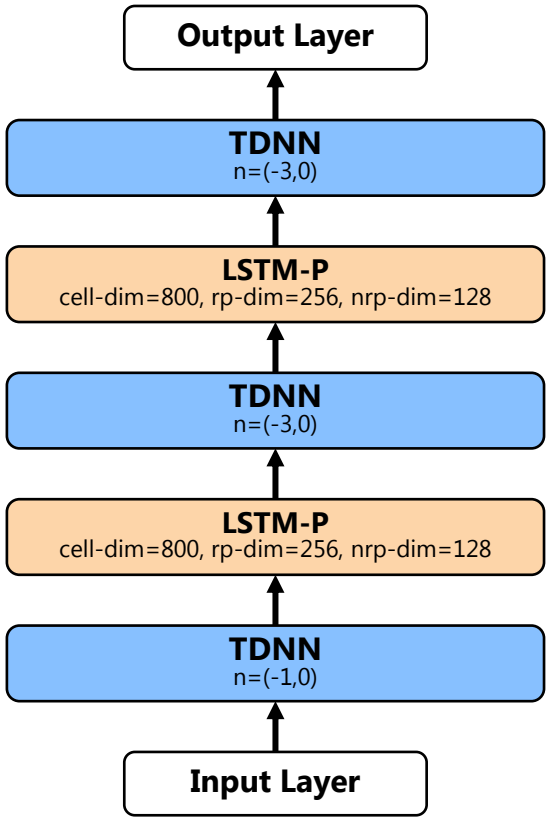
- **Overview of acoustic model configuration**

Enh. in train	Topology	Sys ID	DT	VTLN	2-pass
W+U+U.rvb	TDNNF (15)	Base			yes
W+U.WPE	CNN-TDNNF (19)	A			no
W+U.GSS12	CNN-TDNNF (18)	B			no
	CNN-TDNNF (19)	C	✔		no
		D	✔	✔	no
		E			yes
		F	✔		yes
		G		✔	yes
		H	✔	✔	yes
	RESNET (40)	I			yes
		J	✔		yes
		K		✔	no
	RESNET-TDNNF (49)	L			no
		M	✔		no

2-pass= 2-pass decoding with i-vec refinement
DT = Discriminative Training
U = array data
VTLN = Vocal Tract Length Normalization
W = worn unprocessed data

Language model

- Neural network based language model



- word embedding dim is 800
- rp = recurrent projection
- nrp = non-recurrent projection

Model	Perplexity
3-gram	154.2
TDNN-LSTM	140.5

Results Category A (3G LM)

% WER for DEV (EVAL)

Enh. in train	Topology	Sys ID	DT	VTLN	2-pass	hrs
W+U+U.rvb	TDNNF (15)	Base			yes	1407
W+U.WPE	CNN-TDNNF (19)	A			no	802
W+U.GSS12	CNN-TDNNF (18)	B			no	102
	CNN-TDNNF (19)	C	✔		no	309
		D	✔	✔	no	309
		E			yes	309
		F	✔		yes	309
		G		✔	yes	309
		H	✔	✔	yes	309
	RESNET (40)	I			yes	309
		J	✔		yes	309
		K		✔	no	309
	RESNET-TDNNF (49)	L			no	309
		M	✔		no	309

Sys ID	GSS12+ASR	GSS24+ASR
Base	51.39 (51.38)	-
A	45.81 (46.09)	44.79 (46.78)
B	44.88 (49.48)	-
C	42.47 (44.28)	42.15 (45.03)
D	41.74 (43.84)	41.73 (44.92)
E	42.67 (44.78)	42.28 (45.11)
F	41.78 (44.21)	41.49 (44.72)
G	41.66 (44.11)	41.13 (44.66)
H	41.27 (43.71)	41.34 (44.80)
I	41.34 (42.42)	41.06 (43.09)
J	41.07 (41.84)	40.55 (42.94)
K	40.86 (42.41)	40.53 (43.12)
L	42.03 (42.78)	-
M	41.71 (42.86)	-

A-M	**35.89 (37.54)**

* Lattice Combination

Results Category B (NN LM)

% WER for DEV (EVAL)

3G LM

Sys ID	GSS12+ASR	GSS24+ASR
A	45.81 (46.09)	44.79 (46.78)
B	44.88 (49.48)	-
C	42.47 (44.28)	42.15 (45.03)
D	41.74 (43.84)	41.73 (44.92)
E	42.67 (44.78)	42.28 (45.11)
F	41.78 (44.21)	41.49 (44.72)
G	41.66 (44.11)	41.13 (44.66)
H	41.27 (43.71)	41.34 (44.80)
I	41.34 (42.42)	41.06 (43.09)
J	41.07 (41.84)	40.55 (42.94)
K	40.86 (42.41)	40.53 (43.12)
L	42.03 (42.78)	-
M	41.71 (42.86)	-

A-M	**35.89 (37.54)**

TDNN-LSTM LM

Sys ID	GSS12+ASR	GSS24+ASR
A	44.64 (44.62)	43.42 (45.38)
B	43.39 (45.66)	-
C	41.01 (43.07)	40.71 (43.92)
D	40.53 (42.96)	40.56 (43.74)
E	41.42 (42.85)	40.93 (43.74)
F	40.74 (42.84)	40.38 (43.73)
G	40.57 (42.74)	39.99 (43.43)
H	40.10 (42.70)	40.02 (43.99)
I	40.29 (41.84)	40.04 (42.29)
J	40.11 (41.00)	39.76 (42.19)
K	39.94 (41.42)	39.62 (42.27)
L	41.14 (42.05)	-
M	40.85 (42.16)	-

A-M	**34.83 (36.83)**

* Lattice Combination

Performance analysis

% WER for 3G LM (NN LM)

DEV

Session	Room	% WER
S02	DINING	39.84 (38.46)
	KITCHEN	41.67 (40.65)
	LIVING	32.65 (31.83)
S09	DINING	36.03 (34.47)
	KITCHEN	33.63 (32.62)
	LIVING	31.14 (30.14)
Overall		**35.89 (34.83)**

EVAL

Session	Room	% WER
S01	DINING	31.31 (30.56)
	KITCHEN	53.03 (52.80)
	LIVING	43.38 (42.80)
S21	DINING	29.91 (28.64)
	KITCHEN	45.45 (44.99)
	LIVING	30.16 (29.25)
Overall		**37.54 (36.83)**

Conclusions

- Final system explored variability in:
  - training and test data enhancement
  - acoustic model topology
  - acoustic features and speaker adaptation

- Achieved % WER accuracy DEV (EVAL) for Track 1:
  - 3-gram LM **35.89 (37.54)**
  - NN based LM **34.83 (36.83)**

- Future work
  - address the drop in accuracy in some test conditions (e.g., kitchen)

- There is still enough room for improvement

References

1. T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 20, no. 10, pp. 2707–2720, 2012
2. L. Drude, J. Heymann, C. Boeddeker, and R. Haeb-Umbach, "NARA-WPE: A Python package for weighted prediction error dereverberation in Numpy and Tensorflow for online and offline processing," in Proc. of ITG Fachtagung Sprachkommunikation, Oct 2018
3. C. Boeddecker, J. Heitkaemper, J. Schmalenstroeer, L. Drude, J. Heymann, and R. Haeb-Umbach, "Front-end processing for the CHiME-5 dinner party scenario," in Proc. of CHiME-5 Workshop, 2018
4. C. Zorilă, C. Boeddeker, R. Doddipatla, and R. Haeb-Umbach, "An investigation into the effectiveness of enhancement in ASR training and test for CHiME-5 dinner party transcription," in Proc. ASRU, 2019, pp. 47–53
5. T. Drugman, Y. Stylianou, L. Chen, X. Chen, and M. Gales, "Robust excitation-based features for automatic speech recognition," in Proc. ICASSP, 2015, pp. 4664–4668

TOSHIBA

Thank you!

Contact: catalin.zorila@crl.toshiba.co.uk