The Toshiba Entry to the CHiME 2018 Challenge

TOSHIBA
Leading Innovation >>>

Rama Doddipatla*, Takehiko Kagoshima†, Cong-Thanh Do*, Petko N. Petkov*, Cătălin-Tudor Zorilă*, Euihyun Kim †, Daichi Hayakawa†, Hiroshi Fujimura† and Yannis Stylianou*

*firstname.lastname@crl.toshiba.co.uk, †firstname.lastname@toshiba.co.jp

*Toshiba Cambridge Research Laboratory, Cambridge, United Kingdom

†Toshiba Corporation Corporate R&D Center, Kawasaki, Japan

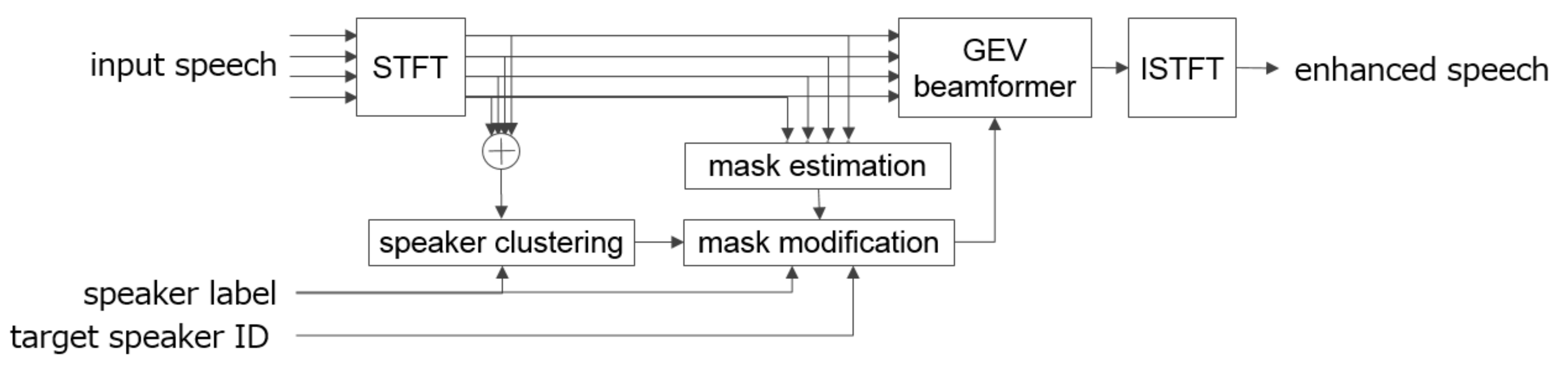
Overview

- The Toshiba entry focuses on single array track and uses the reference array during recognition (**Category-A**).
- The system explores:
  - **Speech Enhancement**: GEV beamforming, WPE based de-reverberation enhancement and Speaker suppression (SS).
  - **Front-ends**: Log Mel filter-bank (FBANK) and subband temporal envelopes (STE) features.
  - **Speaker adaptation**: Vocal tract length normalisation.
- The system achieves a performance of **56.5% WER** on the *eval* set.

Speech Enhancement

A) NN supported *GEV-beamforming* approach is explored.

- The objective is to enhance the target speaker while suppressing background interference (noise and competing speakers).
- Mask training:
  - The worn microphone data is used as clean speech data.
  - Non speech portions are used as noise.
  - Noisy speech is simulated using the clean speech and noise.
- Mask modifications:
  - A GMM is used to estimate the dominant speaker at each frame.
  - The speech mask is set to zero where the dominant speaker (GMM) is different from the target speaker (transcription).
  - The noise mask is set to one where competing speaker and target speaker overlap.
- The modified masks are used as input for GEV beamformer.



B) NN supported linear prediction for *de-reverberation* has been applied following the GEV beamforming.

C) Automatic gain control (AGC) based *speaker suppression* (SS) has also been explored, where an AGC system is used to suppress the interfering speaker. This is applied following the GEV beamforming.

Front-end and Acoustic Model

- 40 dim. Log Mel filter-bank (FBANK) and subband temporal envelope (STE) features are explored.

  - STE's are computed from slowly-varying temporal envelopes in the frequency bands, extracted by filtering speech with Gammatone filters, followed by full-wave rectification and LP filtering.
- The acoustic model (AM) for the presented system uses a combination of 2 convolutional (CNN) and 3 bi-directional long short-term memory (BLSTM) networks.
  - The 2 CNN layers have 256 and 128 filters having 3x3 kernels.
  - Each BLSTM has a cell dimension of 1024 and a recurrent projection of 256. A context of 40 frames (both left and right) is used for the BLSTM layers.
  - *i-vectors* are by-passed from the CNN processing and are append to the output of CNN's as input to the BLSTM layers.

Speaker adaptation: VTLN

- VTLN scales the frequency axis to normalise speaker variability.
- A grid search in the range of 0.85 - 1.25 (steps of 0.01) is performed.
- Applied on top of the GEV beamformed data.
- During recognition, a two-pass approach is performed.

System description

- Multiple AMs are trained, one for each array as well as using data from the all the arrays, with the intention to combine ASR outputs.
- All AMs include worn (**W**) and the corresponding array data.
- The performance of various systems using FBANK and STE features and the CNN-BLSTM AM are presented below.

Track	Data	System	FBANK	STE
	W + U01		67.4	66.6
	W + U02		67.0	65.8
	W + U04		66.1	66.0
	W + U05		66.9	65.6
Single	W + U06		66.3	66.7
	W + Uall	C	64.9	-
	W + Uall - SS	D	64.8	-
	W + Uall - VTLN	E	64.1	-
	W + Uall - WPE	F	**63.3**	-

- **W+U01** refers to data from worn and array-1, **W+Uall** refers to data from all the arrays including worn microphone data.
- The performance of individual arrays are similar.
- **W+Uall - WPE** is the single best performing system.

Lattice Combination

- ASR outputs of various systems are merged using lattice combination (uniform weights) for the final submission system.
- The table below summarises the results:

Systems combined	System	WER
W + U[1-6] - FBANK	A	63.0
W + U[1-6] - STE	B	62.8
A + B		62.0
A + B + D + E + F		**60.8**

- Lattice combination on the individual arrays either FBANK (A) or STE (B) perform better than systems trained using all the data.
- The best performance is achieved with the combination of ASR outputs from A, B, D, E and F (*C in the excluded as D is included*).
- The breakdown over sessions are shown below:

Test Set	Session	Kitchen	Dining	Living	Overall
Dev	S02	70.3	59.7	53.6	**60.8**
	S09	60.9	64.4	57.6	
Eval	S01	69.7	50.2	65.8	**56.5**
	S21	59.2	47.1	54.5	

Summary

- The Toshiba system explored various enhancement methods, multiple front-ends and VTLN for speaker adaptation.
- The system achieved a performance of **60.8% WER** on the *dev* and **56.5% WER** on the *eval* sets respectively.
- The system in **ranked 4th** in the category.