TOSHIBA

On Reducing the Effect of Speaker Overlap for CHiME-5

Tudor-Cătălin Zorilă and Rama Doddipatla
Toshiba Cambridge Research Laboratory, United Kingdom
{catalin.zorila,rama.doddipatla}@crl.toshiba.co.uk

Introduction

Problem statement

- multi-talker distant conversational speech recognition
- competing speakers, reverberation and background noise pose serious challenges for ASR systems

Proposed approach

Improve multi-talker distant ASR performance by **suppressing interfering speakers** using a **neural network supported automatic gain control** (AGC) mechanism.

Context

CHiME-5 challenge: distant multi-microphone conversational speech recognition challenge in everyday home environments [1].

Corpus description:

- 20 dinner party recordings (aprox. 2 hours each)
- 4 participants and 3 locations (kitchen, dining and living room)
- 6 x 4-channels Microsoft Kinect recording devices (array set)
- in-ear binaural microphones (worn set)
- recording devices were not synchronized

- single (reference) device track and multiple device track
- speaker overlap is a major issue for CHiME-5
  - amount of speech frames with more than one active speaker at the same time: 24% (train), 42% (dev)
  - traditional source separation methods were ineffective (moving speakers, reverberation and background noise)
  - speaker-dependent systems exploited the speaker diarisation information provided in the challenge [2]

Methods

Hard overlap suppression (HOS)

- has used the baseline speaker diarisation to detect the segments where only the target speaker is active
- binary masks were computed every 16-ms

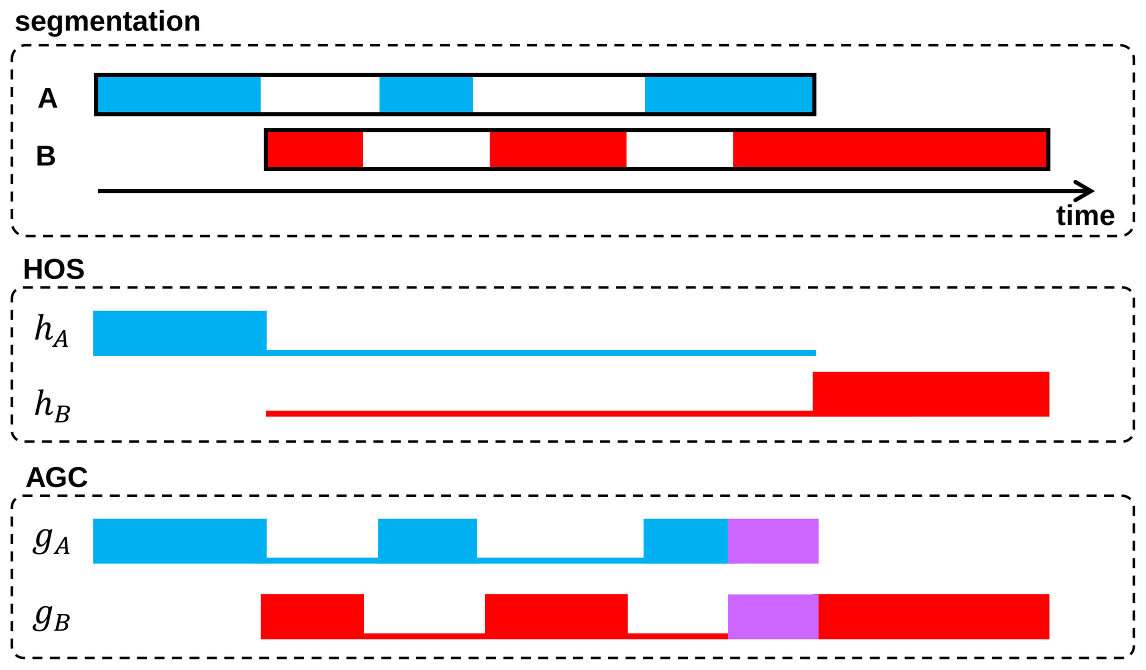

Figure 1: Example of suppressing interfering speakers using HOS and AGC.

Soft overlap suppression (AGC)

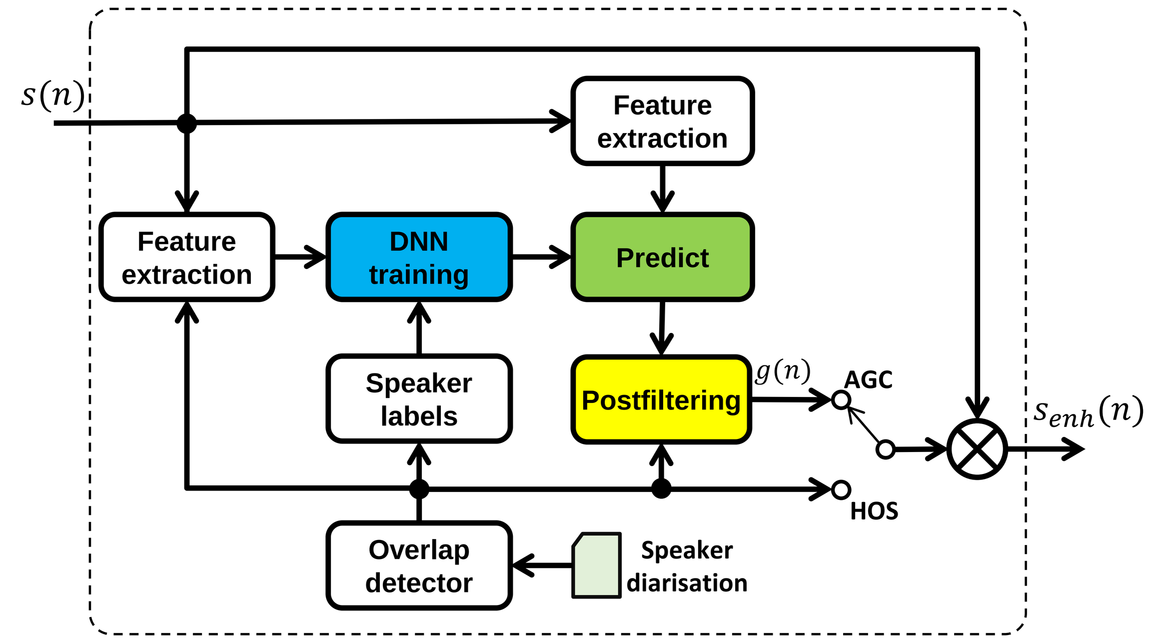

Figure 2: Block diagram of AGC approach for suppressing interfering speakers.

- DNN
  - frame-wise speaker classifier, 3 hidden layers, 4 output nodes
  - training data: single speaker data according to transcription
  - dominant speaker was chosen using maximum likelihood criterion
- postfiltering: 11 taps median filter, double exp. moving average
- single-channel enhancement only

Data selection (DTS)

- random selection of array devices for data extraction is not optimal
- carefully selected devices may reduce interfering speakers' effect
- solution: choose the array device whose data has the **strongest correlation with the in-ear recording** of target speaker
- metric/criterion: normalized cross-correlation/max

Speaker-dependent GEV (SDGEV) [2]

- speaker adaptive maximum SNR beamforming (generalized eigenvalue beamformer, GEV), 4 channels
- neural network to estimate speech and noise statistics (masks)
- GMM based speaker-dependent mask adaptation to alleviate the effect of interfering speakers

Evaluation

- Data
  - ASR training: worn + 100k (randomly chosen) array segments
  - ASR testing: development set of CHiME-5, pre-enhanced using a weighted delay-and-sum beamformer (BeamformIt, BF)
  - enhancement: baseline (unprocessed), HOS, AGC or DTS
- Front-end: 40-dims MFCCs for acoustic model training
- Acoustic model
  - TDNN: 8 layers, lattice-free MMI
  - CNN-BLSTM: 2 layers 2D CNN, 3 layers BLSTM
    - ∗ data cleaning, i-vectors (100), speed perturbation (3-folds)

Results

Table 1: WER(%) using TDNN AM (single device track).

Train data	Enhancement		
	BF	+HOS	+AGC
Baseline	88.3	98.5	88.2
AGC	87.9	97.1	86.6
HOS	87.2	85.0	85.6

Table 2: WER(%) using CNN-BSTM AM trained with unprocessed data.

Track	Enhancement		
	BF (A)	+AGC (B)	A+B
Single	74.0	74.3	71.8
DTS	71.6	71.1	68.9

Table 3: WER(%) using CNN-BSTM AM trained with SDGEV enhanced data (single array).

Track	Enhancement		
	SDGEV (C)	+AGC (D)	C+D
Single	64.9	65.0	63.7

Conclusions

- DNN-based AGC enhancement was proposed for reducing the effect of speaker overlap in CHiME-5
- Experiments have shown that the proposed approach yields WER reductions between 2% and 3% absolute on the *dev* set of CHiME-5.

References

[1] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'CHiME' speech separation and recognition challenge: dataset, task and baselines," in *Proc. Interspeech*, 2018, pp. 1561–1565.

[2] R. Doddipatla, T. Kagoshima, C. Do, P. Petkov, C. Zorila, E. Kim, H. Hayakawa, H. Fujimura, and Y. Stylianou, "The Toshiba entry to the CHiME 2018 challenge," in *Proc. CHiME-5 Workshop*, 2018.