An Investigation into the Effectiveness of Enhancement in ASR Training and Test for CHiME-5 Dinner Party Transcription

Tudor-Cătălin Zorilă[1], *Christoph Boeddeker*[2], *Rama Doddipatla*[1] *and Reinhold Haeb-Umbach*[2]
[1]Toshiba Cambridge Research Laboratory, United Kingdom [2] Paderborn University, Department of Communications Engineering, Germany

Introduction

Motivation

- accuracy of multi-talker distant conversational ASR is still poor
- problems: competing speakers, reverberation, background noise, speech disfluency etc.

Context

- speech enhancement improves word error rate (WER), but is typically applied on the test data only
- it is generally agreed upon that enhancement in ASR training would reduce the acoustic variability
- training data is often artificially increased by adding more degraded speech to it

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-channel distant recording devices (‘U’ set)
  in-ear binaural microphones (‘W’ set)
  recording devices not time synchronized
- single (reference) U device track and multiple U device track
- baseline CHiME-5 system achieved roughly 80% WER

Contributions of this work

- study on the effectiveness of acoustic enhancement in ASR training and test for CHiME-5
- state-of-the-art single-system for CHiME-5

Guided Source Separation (GSS)

- blind source separation method adapted to CHiME-5 [2]
- spatial mixture model:
  complex Angular Central Gaussian Mixture Model (cACGMM)
- cACGMM parameters and posterior probabilities of each speaker being active estimated by EM algorithm
- mask based beamforming (Fig. 1)

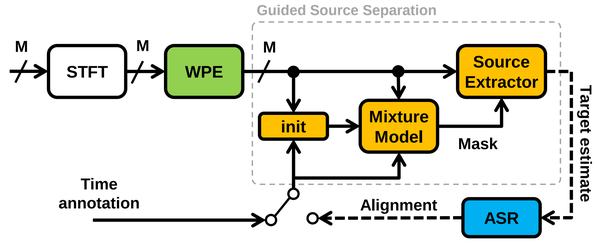

Figure 1: Overview of Guided Source Separation enhancement method.

Experiments & Results

- CHiME-5 corpus was used for ASR training and test (Table 1)
- GMM-HMM alignment model
- acoustic model topology: 6 x CNNs + 9 x TDNNFs
- speed perturbation (3x), 40-dim MFCCs + 100-dim i-vectors
- Lattice-Free Maximum Mutual Information criterion, 3-G LM

Table 1: Naming of the speech enhancement methods.

Enhancement	Array	Label
Unprocessed	Single/Multi	None
BeamformIt	Single	BFIt
WPE + GSS1 + BF w/o Context [2]	Single	GSS1
WPE + GSS6 + BF w/o Context [2]	Multi	GSS6

Effect of acoustic enhancement in ASR training and test

Table 2: WER results on the DEV (EVAL) set and various combinations of speech enhancement for ASR training and test. Amount of training data (hrs) is also specified.

Enh. in trng (hrs)	Enhancement in test			
	None	BFIt	GSS1	GSS6
None (2046)	69.3 (59.9)	69.1 (59.7)	62.2 (58.2)	51.8 (51.6)
BFIt (680)	68.9 (59.1)	68.5 (58.5)	59.9 (57.3)	48.8 (49.9)
GSS1 (791)	74.3 (67.5)	73.7 (66.4)	53.0 (49.6)	48.0 (47.5)
GSS6 (308)	78.5 (73.1)	76.9 (69.2)	58.0 (56.1)	**45.4** (**45.7**)

State-of-the-art single-system for CHiME-5

Table 3: Comparison of the reference [3] and proposed systems in terms of amount of training data.

Track	System	Amount trng data (hrs)	WER in %
Single	H/UPB [3]	4500	58.3 (53.1)
	Proposed	791	**48.6** (**46.7**)
Multiple	H/UPB [3]	4500	45.1 (47.3)
	Proposed	308	**41.6** (**43.2**)

Table 4: Comparison of reference [3] and proposed (single) systems in terms of WER for the DEV (EVAL) set. Test data enhancement was refined using ASR alignments or oracle alignments.

Track	System	Enh. in trng	Enh. in test	DT	RNN-LM	WER in %
Single	H/UPB [3]	None	GSS1 w/ ASR	✓		58.3 (53.1)
	Proposed	GSS1	GSS1 w/ ASR			50.2 (48.4)
	Proposed	GSS1	GSS1 w/ ASR	✓		49.1 (47.3)
	Proposed	GSS1	GSS1 w/ ASR	✓	✓	**48.6** (**46.7**)
	Proposed	GSS1	GSS1 w/ oracle	✓	✓	47.3 (46.1)
Multiple	H/UPB [3]	None	GSS6 w/ ASR	✓		45.1 (47.3)
	Proposed	GSS6	GSS6 w/ ASR			43.2 (44.2)
	Proposed	GSS6	GSS6 w/ ASR	✓		42.3 (43.9)
	Proposed	GSS6	GSS6 w/ ASR	✓	✓	**41.6** (**43.2**)
	Proposed	GSS6	GSS6 w/ oracle	✓	✓	39.9 (42.0)

- best CHiME-5 system (multiple device track, unconstrained LM): USTC-iFlytek; 5-system combination; 45.0 (46.1)% WER

Speaker overlap vs. WER accuracy analysis

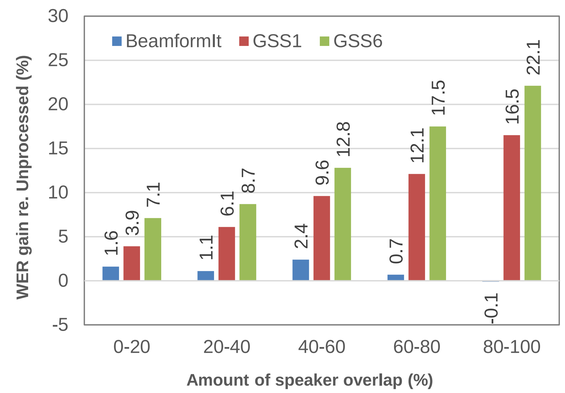

Figure 2: Relative WER gain for the matched case vs unprocessed (EVAL set).

Conclusions

- cleaning up training data can lead to substantial WER reduction
- enhancement in training is advisable as long as enhancement in test is at least as strong as in training
- top *single-system* performance for CHiME-5: 41.6 (43.2)% WER

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] 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.

[3] N. Kanda, C. Boeddeker, J. Heitkaemper, Y. Fujita, S. Horiguchi, K. Nagamatsu, and R. Haeb-Umbach, “Guided source separation meets a strong ASR backend: Hitachi/Paderborn University joint investigation for dinner party ASR,” in *Proc. Interspeech*, 2019, pp. 1248–1252.