FUSION OF DIVERSE DENOISING SYSTEMS FOR ROBUST AUTOMATIC SPEECH RECOGNITION SYSTEM

Naveen Kumar, Maarten Van Segbroeck, Kartik Audhkhasi, Peter Drotár, Shrikanth Narayanan
Signal Analysis and Interpretation Laboratory (SAIL), University of Southern California, Los Angeles, CA-900089
{komathnk,audhkhas}@usc.edu, drotarp@feec.vutbr.cz, {maarten,shri}@sipi.usc.edu

Motivation

Denoising is essential for robust ASR

- Wiener Filtering (WF)
- Spectral Subtraction (SS)
- Cepstral Mean Normalization (CMN)
- Harmonic Decomposition based noise estimation (HD)

Issues

- Sensitive to parameter settings
- Not robust across multiple noise conditions
- Might require noise specific models

Fusion of diverse denoising front ends

Fusion

- **Inter-system** Denoising algorithms (CMN, WF, SS, HD)
- **Intra-system** Diverse HD parameter settings

Recognizer Output Voting Error Reduction (ROVER)

- Unsupervised fusion of 1-best ASR hypotheses
- Align hypotheses to get sausage network
- Combine by majority voting

Harmonic Decomposition

Algorithm:

- Estimate aperiodic part by least-squares estimation
- Noise estimation from residual using minimum statistics
- Spectral subtraction denoising

Parameter Settings

- Size of running window for NE
- Noise floor level of aperiodic signal
- Noise reduction factor during speech
- Noise factor to compensate for non linear artifact

23 parameter settings $\{p_1, \ldots, p_{23}\}$ chosen.
Objective: Find the 3 subset $\{p_x, p_y, p_z\}$ of parameter settings that is most diverse.

Experimental Setup

Dataset

- Aurora4 dataset
- 6 noise types from 5db to 15db, clean train
- 7138 train, 330 test utterances

ASR

- *KALDI* Speech Recongition toolkit
- Triphone Models
- CMN + LDA + MLLT + SAT

ROVER

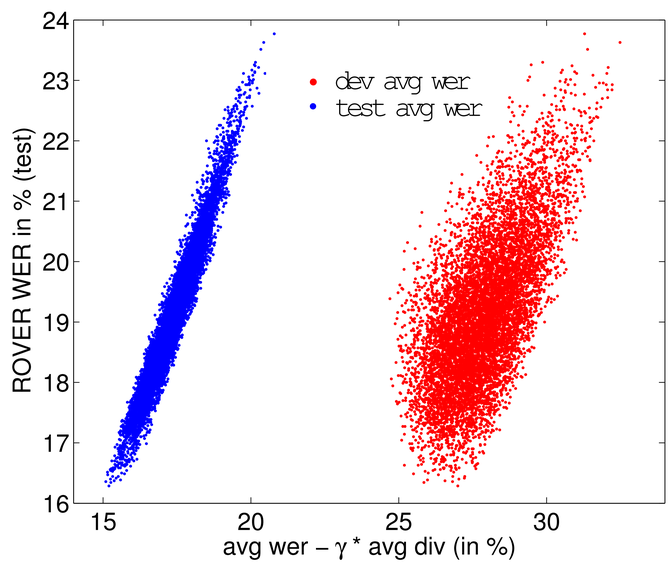
- Word Frequency based (wfR)
- Confidence based (cfR)

Diverse Decomposition Approximation

We use the decomposition proposed in [1]

$$\underbrace{E(r,h^*)}_{\text{ROVER WER}} \approx \underbrace{\frac{1}{K}\sum_{k=1}^{K} E(r,h_k)}_{\text{Average WER}} - \gamma \underbrace{\frac{1}{K}\sum_{k=1}^{K} E(h^*,h_k)}_{\text{Diversity}}$$

h_k Hypothesis by kth system, h^* ROVER system

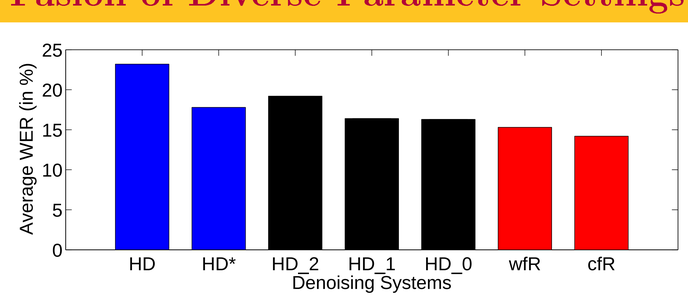


InterSystem Rover

Aurora4, 16kHz, clean condition training.								
	Close Talk							
Test	01	02	03	04	05	06	07	Avg.
CMN	**4.2**	**9.4**	24.1	31.2	25.9	**21.4**	28.1	20.6
WF	4.4	9.7	20.9	26.6	23.0	22.4	23.6	18.7
SS	4.5	10.8	22.9	24.9	22.9	23.0	24.0	19.0
HD*	6.6	12.6	**17.0**	**24.1**	**22.7**	22.7	**19.2**	**17.8**
wfR	4.1	7.9	16.9	22.1	19.0	18.0	18.8	15.3
cfR	3.6	6.9	16.1	21.2	18.2	16.4	18.6	14.2

*n*o noise, car, babble, restaurant, street, airport and train
Parameter setting for HD tuned on a small subset of test

Fusion of Diverse Parameter Settings



- HD_0 best acheivable WER by intra-system fusion
- HD_1 dev average WER
- HD_2 test average WER

Discussion

✓

- Noise type agnostic models
- Tuned on average WER
- Parameters complementary to noise types

?

- Sensitive to γ
- Dev test mismatch
- Search for diverse parameter settings

References

[1] Kartik Audhkhasi, A Zavou, P Georgiou, and S Narayanan, "Theoretical analysis of diversity in an ensemble of automatic speech recognition systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 3, March 2014.