ROBUST FEATURES AND SYSTEM FUSION FOR REVERBERATION-ROBUST SPEECH RECOGNITION

Vikramjit Mitra[1], Wen Wang[1], Yun Lei[1], Andreas Kathol[1], Ganesh Sivaraman[2], Carol Espy-Wilson[2]

[1]SRI International, Menlo Park, CA {vmitra, wwang,yunlei,kathol} @speech.sri.com

[1]University of Maryland, College Park, MD {ganesa90, espy}@umd.edu

Introduction

- Reverberation in speech degrades the performance of speech recognition systems.
- Human listeners can often ignore reverberation,
  - auditory system somehow compensates for reverberation degradations
- We present robust acoustic features motivated by human speech perception and production

Robust Features

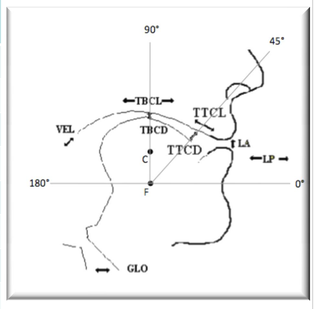
Perceptual Features:

- **Damped Oscillator Cepstra** (DOCC) [Mitra'2013]
  - Models the dynamics of the hair cells within the cochlea using forced damped oscillators (FDO)
  - Uses the response of the FDOs as acoustic features.
  - Performs non-linear root compression
- **Normalized Modulation Cepstra** (NMCC) [Mitra'2012]
  - Tracks amplitude modulation (AM) of subband speech signals
  - Uses Discrete Energy Separation Algorithm [Maragos'93] to obtain instantaneous AM estimates
  - Performs non-linear root compression
- **Modulation of Medium Duration Speech Amplitudes** (MMeDuSA) [Mitra'2014]
  - Uses directly the Teager energy operator [Teager'80] to estimate the AM signals.
  - Computes the cumulative AM modulation feature.
    - Cumulative info. obtained by summing the AM signals across frequency, keeping the modulation info. between 5 to 200 Hz
  - More noise-robust to obtain info. about the overall modulation.
  - Geared to capture voicing information.
  - Captures vowel stress and prominence information.
- **Gammatone Cepstra** (GCC)
  - Uses perceptually motivated gammatone banks to analyze speech
  - Performs root compression
- All of these features had their Δ, Δ^2, and Δ^3 coefficients appended and were HLDA transformed to 39 coeffs.

Production Features:

- **Tract-variable trajectories** (TVs), are articulatory features that captures the dynamics of the vocal tract shape.
- Used a thin deep neural network (DNN) with 150, 200, 100, 80, 60, and 40 neurons to predict the TVs from speech

Constriction organs	Tract Variables
Lip	Lip Aperture (LA)
	Lip Protrusion (LP)
Tongue Tip	Tongue tip constriction degree (TTCD)
	Tongue tip constriction location (TTCL)
Tongue Body	Tongue body constriction degree (TBCD)
	Tongue body constriction location (TBCL)
Velum	Velum (VEL)
Glottis	Glottis (GLO)



- DNN was trained using an artificially generated synthetic clean word corpus
- No information regarding noise or reverberation was used while training the DNN
- Modulation of the TVs (over a 200 ms window) was computed and combined with the MFCCs.
- Resulting feature was PCA transformed to retain 30 dimensions (MFCC+ModTV_pca30)

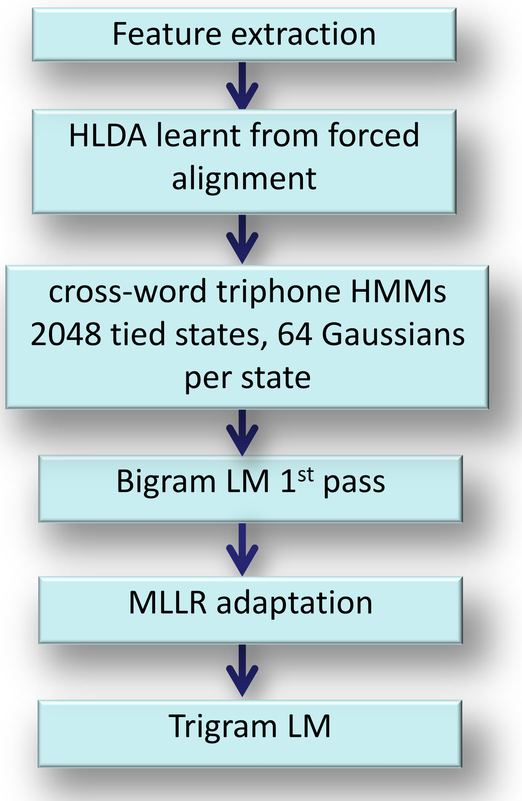
Data

Evaluations performed using

- REVERB 2014 challenge speech dataset.
- Single-speaker utterances recorded with 1-, 2- or 8-channel circular microphone arrays
  - We used only the 1-channel training condition
  - 7861 utterances (5699 unique utterances)
- Dataset includes a training set, a development set, and an evaluation set
  - Eval. and Dev. data contain both real and simulated reverberation data.

Speech Recognition System

- Primary submission system used SRI's DECIPHER™ speech recognition system
- Speaker info. was not used.
- 5K non-verbalized punctuation, closed vocabulary set language model (LM)



- HLDA transform to reduce features to 39D before training the acoustic model.
- HLDA not performed on 30D MFCC+ModTV_pca30 features.
- HLDA learnt from forced-alignment of the chan-1 training data

FEATS	Before HLDA	After HLDA
MFCC	52	39
NMCC	52	39
DOCC	52	39
GCC	52	39
MMeDuSA	55	39
MFCC+ModTV_pca30	30	30

Results

We tried two baseline systems

1. MFCC-HTK system distributed through the REVERB 2014 challenge website
2. DECIPHER™-MFCC system

Observations:

- MLLR and HLDA helps to reduce WER
- Decipher™ baseline better than REVER baseline

Acoustic model	Feature	Adapt.	WER (%) far	WER (%) near	WER (%) avg.
HTK [REVERB2014]	MFCC	none	51.5	53.7	52.6
	MFCC	cMLLR	46.0	47.3	46.7
DECIPHER [SRI]	MFCC	none	50.1	52.3	51.2
	MFCC	MLLR	43.7	45.3	44.5
	MFCC+HLDA	none	46.2	49.1	47.7
	MFCC+HLDA	MLLR	**41.2**	**39.9**	**40.5**

Conclusion

- Robust features motivated by speech perception and production can improve reverberation robustness
- Long term (temporal) modeling through Δ, Δ^2, and Δ^3 coefs. helps .
- System fusion helps to lower error rates
- CNN system provided a substantial gain.

WERs on the evaluation set from the different systems (1-channel training and full-batch processing) submitted to the REVERB 2014 challenge

FEATURES	WER (%) sim data Room 1 Near	Room 1 Far	Room 2 Near	Room 2 Far	Room 3 Near	Room 3 Far	Ave.	real data Room 1 Near	real data Room 1 Far	Ave.
MFCC(39)-HTK	14.21	17.45	21.07	37.19	22.73	40.28	25.49	46.97	47.37	47.17
MFCC	12.83	12.10	13.99	25.49	16.81	32.61	18.97	41.90	39.87	40.89
DOCC	8.64	9.88	12.85	23.43	14.08	30.32	16.53	40.85	40.75	40.80
MFCC-TV_pca30	9.79	11.22	14.20	29.02	18.36	40.44	20.51	43.53	44.16	43.85
2-way ROVER (DOCC+MFCC-TV_pca30)	7.42	8.98	11.83	22.87	14.06	30.98	16.02	38.61	38.45	38.53
3-way Rover (MFCC+DOCC+MFCC-TV_pa30)	7.83	8.71	11.14	21.34	13.27	28.51	15.14	36.44	35.75	36.10

WERs on the evaluation set from the different systems (1-channel train. and full-batch processing) from post-sub. to the REVERB 2014 challenge

FEATURES	WER (%) sim data Room 1 Near	Room 1 Far	Room 2 Near	Room 2 Far	Room 3 Near	Room 3 Far	Ave.	real data Room 1 Near	real data Room 1 Far	Ave.
MFCC(39)-HTK [REVERB 2014 baseline]	14.21	17.45	21.07	37.19	22.73	40.28	25.49	46.97	47.37	47.17
MFCC [SRI's baseline]	12.83	12.10	13.99	25.49	16.81	32.61	18.97	41.90	39.87	40.89
MFCC-TV_pca30	9.79	11.22	14.20	29.02	18.36	40.44	20.51	43.53	44.16	43.85
DOCC	8.64	9.88	12.85	23.43	14.08	30.32	16.53	40.85	40.75	40.80
NMCC	10.03	11.32	14.29	29.78	17.62	39.57	20.44	42.03	40.95	41.49
MMeDuSA	9.62	11.06	13.11	26.40	16.31	34.53	18.51	46.92	45.17	46.05
GCC	9.78	11.78	13.12	27.09	16.77	36.34	19.15	39.12	41.22	40.17
DOCC (CNN-system)	9.73	10.73	10.58	**18.43**	13.38	**23.46**	14.39	37.85	37.54	37.70
2-way-ROVER (opt: GCC+MFCC)	8.91	9.61	11.43	22.17	13.89	30.17	16.03	35.84	36.36	36.10
3-way-ROVER (opt: GCC+MFCC+NMCC)	8.56	9.62	11.64	23.40	14.04	32.09	16.56	36.09	36.87	36.48
4-way-ROVER (opt: DOCC+GCC+MFCC+NMCC)	7.52	8.69	10.74	21.11	12.92	29.45	15.07	34.78	35.18	34.98
5-way-ROVER (opt: DOCC+GCC+MFCC+MFCCTV+NMCC)	7.25	8.34	10.66	21.51	12.92	29.53	15.04	34.40	35.01	34.71
6-way-ROVER (all subsystems)	7.22	8.40	10.55	21.24	12.92	29.62	14.99	35.20	35.45	35.33
ROVER with DOCC CNN system	**6.27**	**7.13**	**9.09**	18.83	**11.55**	26.19	**13.18**	**32.64**	**33.25**	**32.95**

Acknowledgement: Research was partially supported by NSF Grant # IIS-1162046