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 Cepstrum (NMCC) [Mitra’2012]
  • Tracks amplitude modulation (AM) of subband speech signals
  • Uses Discrete Energy Separation Algorithm [Maraga’93] to obtain instantaneous AM estimates
  • Performs non-linear root compression

• Modulation of Medium Duration Speech Amplitudes (MMeDuSA) [Mitra’2014]
  • Uses directly the Tegner energy operator [Tegner’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 -500 to 200 Hz
  • More noise-robust to obtain info. about the overall modulation.
  • Captures vowel stress and prominence information.

• Gammatone Ceptra (GCC)
  • Uses perceptually motivated gammatone banks to analyze speech
  • Performs root compression
  • All of these features had their $\Delta$, $\Delta^1$, and $\Delta^2$ 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

• Lip Aperture (LA)
  • Lip-Opening (LS)
  • Tongue Tip
  • Tongue Body
  • Lip Closure
  • Glottis

• MFC(39) 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 MFCs.
• Resulting feature was PCA transformed to retain 30 dimensions (MFCc+ModTV_pca30)

Evaluation 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-lexicalized pronunciation, closed vocabulary set language model (LM)

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 Reverberation

Conclusion

• Robust features motivated by speech perception and production can improve reverberation robustness
• Long term (temporal) modeling through $\Delta$, $\Delta^1$, and $\Delta^2$ coefcs. 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

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