ROBUST RECOGNITION OF SPEECH WITH BACKGROUND MUSIC IN ACOUSTICALLY UNDER-RESOURCED SCENARIOS

Jiří Málek, Jindřich Žďánský a Petr Červa
Technical University of Liberec, Czech Republic

ICASSP 2018
Calgary, Canada

INTRODUCTION

Task: Robust automatic speech recognition of speech with background music
• Applications like online 24/7 monitoring of broadcast media

Two scenarios, where we aim to achieve robust recognition:
1) Acoustically under-resourced: Small amount of labeled training utterances
   (only 1 hour) + additional amount of non-labeled training utterances (20 hours)
2) Standard: Large amount of labeled training utterances (132 hours)

Three techniques to achieve the goal:
1) Multi-condition training of acoustic models
2) Denoising autoencoders for feature enhancement
3) Joint training of both above mentioned techniques

For both scenarios all three techniques achieve improved performance compared to baseline acoustic models trained on clean speech.

Improvements in under-resourced scenario:
• Using non-labeled data, autoencoder is trained to provide robust feature enhancement
• Using the small amount of available labeled data, the autoencoder is fine-tuned along with acoustic model to provide more robust recognition.

TRAINING DATASETS

Training datasets:
1) Large - 132 hours of labeled czech speech
2) Small - 1 hour of labeled subset of Large, under-resourced scenario
   Additional 20 hours of non-labeled data, easier to obtain than labeled

All distorted training sets created by augmentation:
• Partitioning of available speech dataset into four parts
• First part left undistorted
• Other parts: summation of speech and music; SNR 0.5 and 10 dB

Music dataset: 667 minutes of Electronic music
• Resembles background music in TV shows

TEST DATASETS

Generated dataset: 13622 words, dictated in silence on colose-talk mic
• Using augmentation with electronic music with SNR levels 10, 5, 0, -5 dB

In total five instances for different SNR levels

Real-world dataset: 2222 words from local radio news
• Electronic music with approximated SNR 10 dB on the background

GENERAL ACOUSTIC MODEL ARCHITECTURE

HMM-DNN architecture
• Underlying GMM context dependent, speaker independent
• Small dataset – 619 states, Large – 2219 states

Features
• 39 filter bank coefficients, 25 ms frames, 10 ms shift
• Input vector: 11 consecutive frames, 5 preceding, 5 following current
• Normalization: Mean subtraction; floating window of 1 s.

RECOGNITION ENGINE

One-pass speech decoder with time-synchronous Viterbi search
• We do not investigate the under-resourced scenario from linguistic point of view

Linguistic part: Lexicon: 550k entries (words and collocations)
• Newspaper language model: For simulated datasets
• Broadcast language model: For real-world datasets
• Bigram language model structure

INVESTIGATED TECHNIQUES

Multi-condition training
• Acoustic models have HMM-DNN architecture
  1) FAM - Fully-connected deep neural network Acoustic Model
  2) CAM - Convolutional deep neural network Acoustic Model
• Autoencoder for removal of music from features
  1) FAE - Fully connected autoencoder
  2) CAE - Convolutional autoencoder
• Followed by FAM training on the processed data
• Joint training of cascade FAE + FAM
• Multi-condition training using noisy data
• Baseline acoustic model
  • Single-style training (SCT) using undistorted speech data

Multi-condition training

FAM – Fully-connected deep neural network Acoustic Models
• 5 feedforward fully-connected hidden layers, 768 units.

CAM – Convolutional deep neural network Acoustic Models
• 2 convolutional, 3 fully-connected layers (768 units)
  • Input: 11 feature maps, 39 x 1 in size, 11 consecutive feature vectors
  • First conv. layer: 105 maps 39 x 1, second conv. layer: 157 maps 13 x 1
  • Target: Senones (619 small dataset model, 2219 large dataset model)

Training criterion:
• Negative log-likelihood criterion

Fully-connected denoising autoencoder (FAE)
• Input: 11 distorted feature frames
• Architecture: Feedforward, four hidden layers, 768 units each
• Target: True undistorted speech feature frame
• Training criterion: Mean square error
  • Sensitive to scaling, feature normalization to zero mean and unit variance

Convolutional denoising autoencoder (CAE)
• Input: 11 feature maps 39 x 1, i.e., 11 consecutive feature vectors
• Architecture: Two conv. layers (105 maps 39x1 and 157 maps 13x1)
  • 3 fully-connected layers (768 units)
• Convolutional kernel: 5 x 1
• Target: True undistorted speech feature frame
• Training criterion: Mean square error

Joint training of CAE and FAM (JCMT)
1) CAE is trained as described above, but:
  • Target: 11 consecutive frames of true clean speech
  • Architecture change: Single fully connected layer only
  • FAM is trained using data processed by CAE.
  • Architecture change: Two fully connected layers only
2) FAM is trained using data processed by CAE.
  • Architecture change: Two fully connected layers only
3) Concatenation of CAE and FAM into single network
4) Fine-tuning of joined network using negative log-likelihood criterion; target: senones

JCMT acoustic model is of the same size and topology as CAM.

EXPERIMENTS: MODELS TRAINED ON SMALL DATASET

Results stated as absolute improvements of accuracy

Undistorted dataset: SCT baseline: 76.8% accuracy
• MCT and JCMT achieve comparable performance to SCT

Distorted generated datasets: Performance of SCT baseline deteriorates to 20.5% at 0 dB
• Most of the robust techniques achieve considerably higher accuracy
• FAE: Not beneficial when applied to the small dataset
• CAE: Significantly better results than FAE, improves over SCT by 5-14%
• MCT: Significantly improves over SCT by 14-23%, CAM/FAM comparable
• JCMT: Comparable in topology to CAM, better results, especially for low SNR
• Additional non-labeled data (20 hours): Improves performance of all applicable techniques, e.g., 1-4% for JCMT

Real-world dataset: Comparable to 10dB generated case, SCT performance deteriorates less significantly, otherwise consistent with results above

EXPERIMENTS: MODELS TRAINED ON LARGE DATASET

Undistorted dataset: SCT baseline: 84.9% accuracy
• All compared techniques achieve comparable performance

Distorted generated datasets: Performance of SCT baseline deteriorates to 38.7% at 0 dB
• All of the robust techniques achieve considerably higher accuracy
• FAE: The least beneficial technique, improves over SCT by 3-28%
• CAE: Better results than FAE, improves over SCT by 6-31%
• MCT: Significantly improves over SCT by 5-37%, CAM/FAM comparable on high SNR
• JCMT: Comparable in topology to CAM, improves over CAM by about 1%

Real-world set: Comparable to 10dB generated scenario, consistent with results above

CONCLUSIONS

Both training dataset sizes: All techniques improve accuracy compared to SCT
• Autoencoders: CAE is more beneficial than FAE
• Multi-condition training: CAM achieves higher accuracy compared to FAM
• Joint training: Topology comparable to CAM, better results (especially for small dataset)

Small dataset: Smaller accuracy compared to large training dataset
• Additional non-labeled data: improve significantly autoencoder and JCMT performance