ROBUST AUTOMATIC RECOGNITION OF SPEECH WITH BACKGROUND MUSIC

ICASSP 2017
New Orleans
Louisiana
USA

Jiri Malek, Jindrich Zdansky and Petr Cerva
Technical University of Liberec, Czech Republic

Introduction

- Robust recognition of speech with background music
- Two approaches:
  1. Multi-condition training of the acoustic models
  2. Denoising autoencoders followed by acoustic model training on the pre-processed data
- Both technique improve robustness of ASR significantly
  - Artificial mixture, Signal-to-Noise Ratio (SNR) of 0 dB: absolute improvement of accuracy 35.8%
  - Real-world mixture, SNR about 10 dB: absolute improvement of accuracy 2.4%
- Studied approaches do not deteriorate clean speech recognition: about 1% decrease of accuracy

Motivation

Introduction:

- **ASR:** current research focused on robustness to environmental conditions
  1. Distant microphones
  2. Concurrent speech
  3. Background interference

Our specific task:

- Robust recognition of speech
- Background interference: Music
- Application: online $24/7$ monitoring of broadcast media

Considered Techniques for Robust ASR

Approaches:

- **Multi-condition training** of acoustic models (MCT)
  - *Architecture:* Hybrid Hidden Markov Model - Deep Neural Network
  - *Neural network topology:* Fully-connected feed-forward
- **Denoising autoencoders** for feature enhancement + training of acoustic model on enhanced features (DAE)
  - *Architecture:* Deep Neural Network
  - *Topology:* Fully-connected and convolutional

Training data

- Generated artificially by augmentation of clean speech
- **Clean speech dataset:**
  - Language: Czech
  - Duration: 132 hours
- **Music dataset:**
  - Genres: Piano tracks and electronic music
  - Duration: 11 hours 40 minutes

General acoustic model structure

Hybrid HMM-DNN:

- **Underlying GMM:** Context dependent, speaker independent, 2219 physical states
- **Features:**
  - Filter bank coefficients (frames 25 ms long with 10 ms shift)
  - Applied Cepstral Mean Subtraction (window 1 s)
  - Input of DNN: 11 concatenated frames
- **DNN:**
  - Fully-connected feed-forward
  - 5 hidden layers, 768 neurons each
- **Baseline:** Single-style training on undistorted instance of speech dataset

Multi-condition training of acoustic model

- **Training dataset:**
  - *Artificially created:* Summation of clean speech with music
  - Training database split into N parts
  - *Noise levels:* Each part distorted with specific average SNR level
- **Considered models:**
  - *Piano 1:* High SNR levels of piano music only
  - *Piano 2:* Broad range of SNR levels with piano music
  - *Electronic:* Electronic music resembles broadcast jingles

Table 1: Setup of the training set for multi-style acoustic models and respective autoencoders

Dataset (genre)	N	SNR levels	Music styles included
Piano 1	3	clean, $10, 5, 0$	Classical piano
Piano 2	7	clean, $10, 5, 0, -5, -10, -15, -20$	Classical piano
Elect. 1	3	clean, $10, 5, 0$	Ambient, dance, down-tempo, chillout or idm

Fully connected denoising autoencoder

- **Feed-forward DNN**
  - *Input:* 11 frames of 39 distorted filter bank coefficients
  - *Target:* Signal frame of clean speech filter bank coefficients
  - *Training set:* The same as for multi-condition training
  - *Criterion:* Mean square distance
  - *Normalization:* Zero mean and unitary variance of inputs and targets
  - *Topology:* 3 hidden layers, 1024 neurons each

Convolutional denoising autoencoder

- **Feed-forward DNN**
  - *Input:* 11 feature maps of 39 distorted filter bank coefficients
  - *Target:* Signal frame of clean speech, 39 filter bank coefficients
  - *Training set:* The same as for multi-condition training
  - *Criterion:* Mean square distance
  - *Normalization:* Zero mean and unitary variance of inputs and targets
  - *Topology:* 2 convolutional + max-pooling (factor of 3) + 2 full layers
  - *Convolutional kernel:* covers 5 $\times$ 1 coefficients
  - *Feature maps:* 13 $\times$ 39 and 39 $\times$ 13 elements

Experiments

Test sets:

- **Generated test set**
  - *Speech duration:* 2 hours 44 minutes (close-talk mic)
  - *Seen music genre:* piano (8 minutes), electronic (40 minutes)
  - *Unseen music genre:* piano and violin (144 minutes)
  - Dataset replicated for each SNR level in Table 2
- **Real-world dataset**
  - *Distorted speech:* 18 minutes of radio broadcasts
  - Electronic music jingle is present at the background (approximate SNR 10dB).

Table 2: Setup of the artificially generated test sets

Dataset (genre)	SNR levels	Music styles included
Clean	clean	None
Test:Piano	$10, 0, -10, -20$	Classical piano
Test:Violin	$10, 0, -10, -20$	Piano and violin compositions
Test:Electro	$10, 5, 0, -5$	Ambient, dance, down-tempo, chillout or idm

Recognition engine

- One-pass speech decoder with time-synchronous Viterbi search
- Linguistic part:
  - *Newspaper language model*: For simulated datasets
  - *Broadcast language model*: For real-world datasets
  - *Lexicon:* 550k entries (words and collocations)
  - Bigram language model

Matched training-test conditions

Undistorted data: (Figure 1)

- *Baseline model:* 85.0% accuracy
- *Robust techniques:* Comparable (degradation 0.1 - 1.1%)

Piano dataset: (Figure 1)

- *Baseline model:* Decrease by 16.9% for SNR level 0 dB
- *Robust techniques:* Much lower degradation (1.3-2.2%)
- Comparable results of MCT and autoencoders

Electronic dataset: (Figure 1)

- *Baseline model:* Decrease by 46.1% for SNR level 0 dB
- *Robust techniques:* Improvement over baseline by up to 35.8%
- MCT achieves higher performance than autoencoders

Figure 1: Dataset Test:Piano (numbers in braces: unseen SNR level)

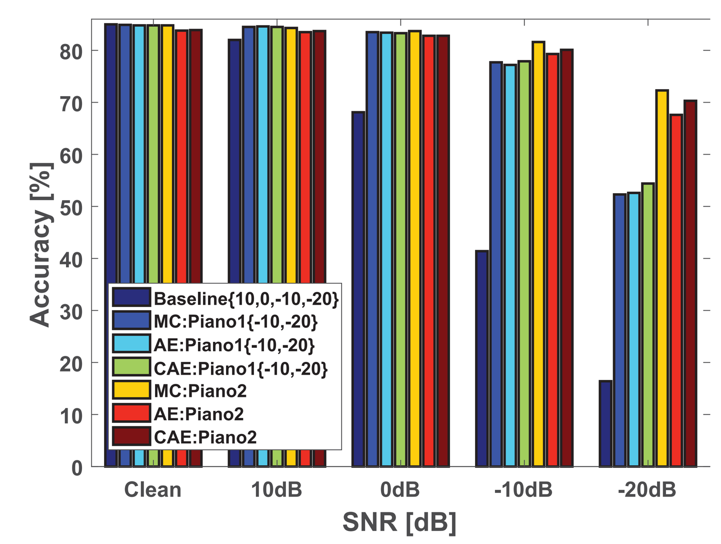
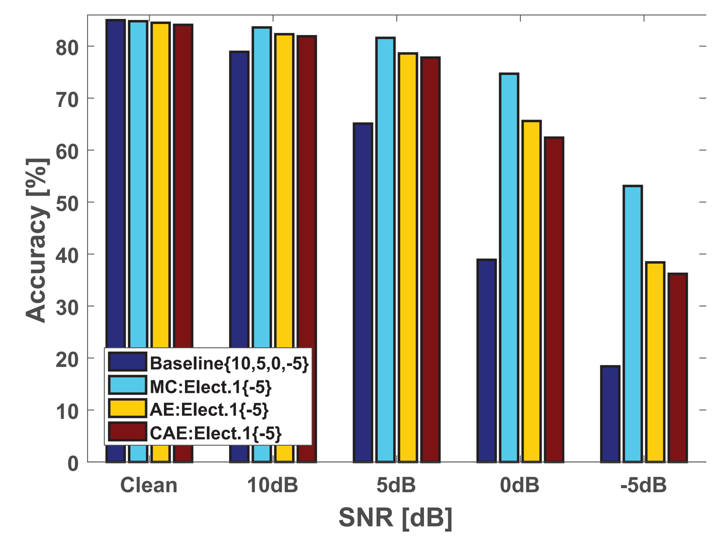


Figure 2: Dataset Test:Electro (numbers in braces: unseen SNR level)



Mismatched training-test conditions

Piano dataset: (**unseen low SNR level**, Figure 1)

- *Baseline model:* Decrease by 68.6% for SNR level -20 dB
- *Robust, mismatched train-test SNR:* improvement by 38%
- *Robust, matched train-test SNR:* improvement by 55.9%

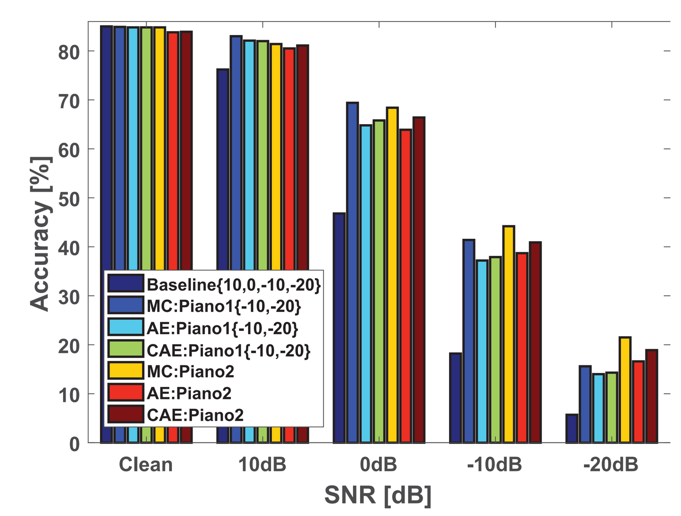
Electronic dataset: (**unseen low SNR level**, Figure 2)

- *Baseline model:* Decrease by 66.6% for SNR level -5 dB
- *Robust techniques:* improvement by 34.7%
- MCT performs better than AEs by up to 14.7%

Piano and violin: (**unseen music and low SNR level**, Figure 3)

- *Baseline model:* Decrease to 38.2% for SNR level 0 dB
- *Robust techniques:* improvement over baseline by 24.3%
- MCT more robust to unseen condition than AEs

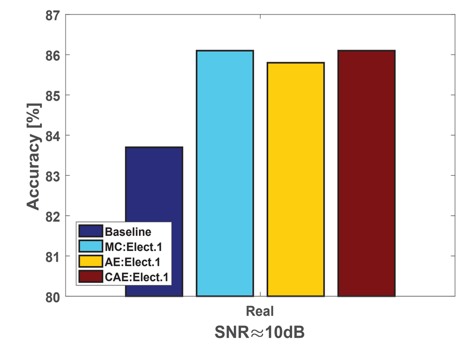
Figure 3: Dataset Test:Violin (unseen music genre, numbers in braces: unseen SNR level)



Real-world dataset

Radio broadcast: (**unseen music, SNR level 10dB**)

- Robust techniques improve by 2.4% over baseline
- Comparable results to Test:Piano at SNR level 10dB



Conclusions

1. The considered techniques are robust to music interference
2. MCT and autoencoders:
   - comparable for matched conditions and simpler music
   - MCT superior for mismatched conditions and complex music
3. Autoencoder topologies (equal number of hidden units):
   - AE performs better in more complex scenarios
   - CAE performs better in simpler scenarios and for lower SNR
   - See **Addendum** for more details
4. Broader range of music during training results in robustness vs unseen genre
5. Broader range of SNR levels during training improves performance
6. *MCT advantage:* Simpler training procedure; single network
7. *AE advantage:* training data do not need to be labeled; easier training set compilation

Addendum - autoencoder topologies

- CAE benefits from
  - deeper network (more than AE)
  - broader convolutional layers
- Using these fact, CAE outperforms AE in general
- MCT is still superior to autoencoders

INSTITUTE OF INFORMATION TECHNOLOGY AND ELECTRONICS
Technical University of Liberec

SpeechLab
Speechlab
Technical University of Liberec

ACOUSTIC SIGNAL ANALYSIS AND PROCESSING GROUP
http://asap.ite.tul.cz