Improving Automatic Speech Recognition Containing Additive Noise Using Deep Denoising Autoencoders of LSTM Networks

Marvin Coto, John Goddard, Fabiola Martínez

Metropolitan Autonomous University (México)

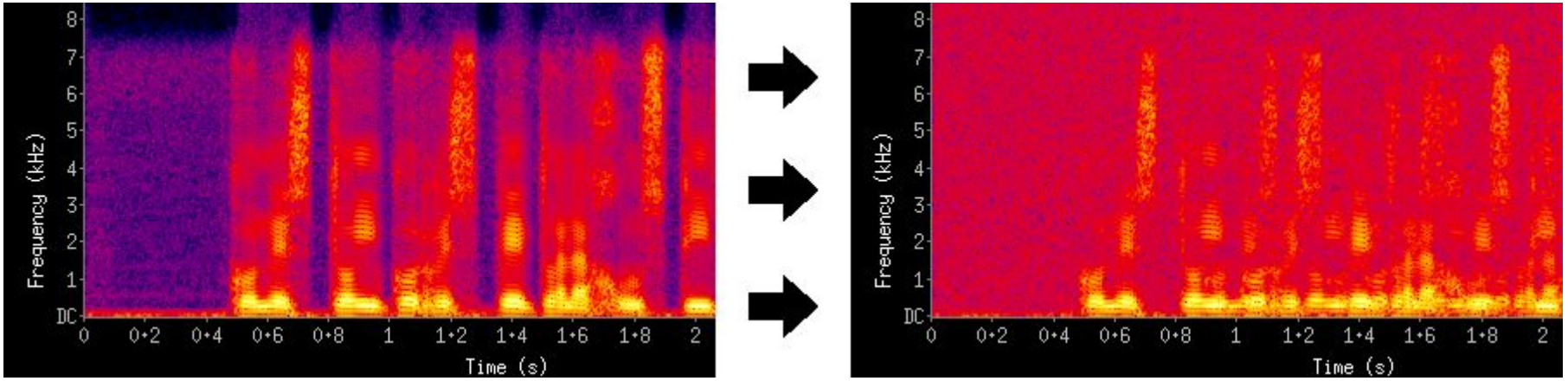
August 2016

Contents

- Introduction
- Long Short-Term Memory Networks
- Description of the system
- Experiments
- Results
- Conclusions

1. Introduction

- Real world environments often adversely affect speech signals through the introduction of contaminants such as noise and reverberation.



Clean speech signal

Noisy speech signal

- In this case, Automatic Speech Recognition systems (ASR) may experience a degradation in their recognition performance

Introduction

- Would like to achieve recognition accuracy similar to that of quiet, controlled environments.
- Two ways to attack this problem of noise robustness are:

Feature-based methods:

Attempt to remove the corrupting noise from the observations prior to recognition

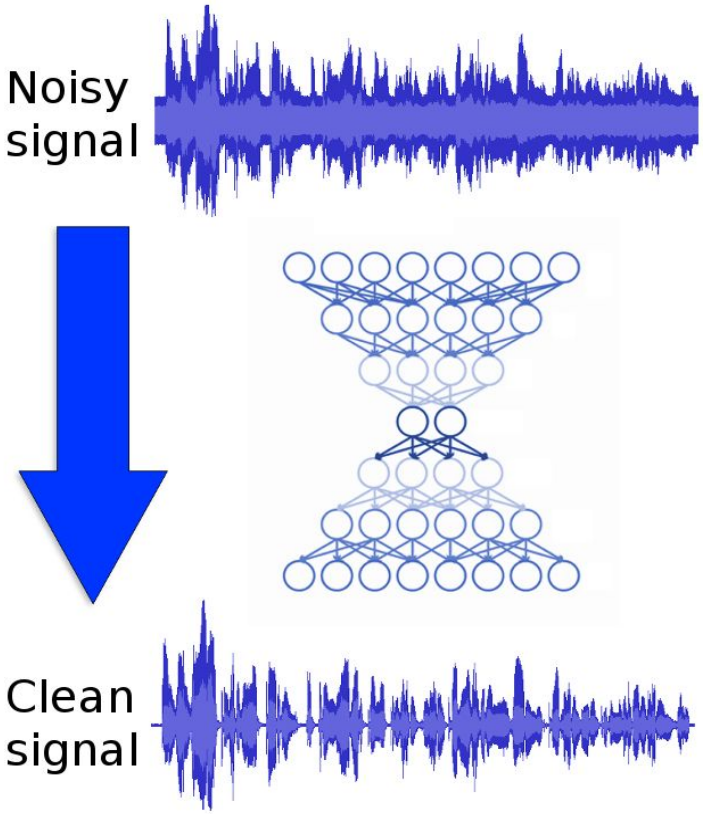
Model-based methods:

Leave the observations unchanged and instead update the model parameters of the recognizer

Introduction

- Recently, deep neural networks have been shown to be effective for a variety of speech research tasks, for example:

- Robust speech recognition
- Speech enhancement systems



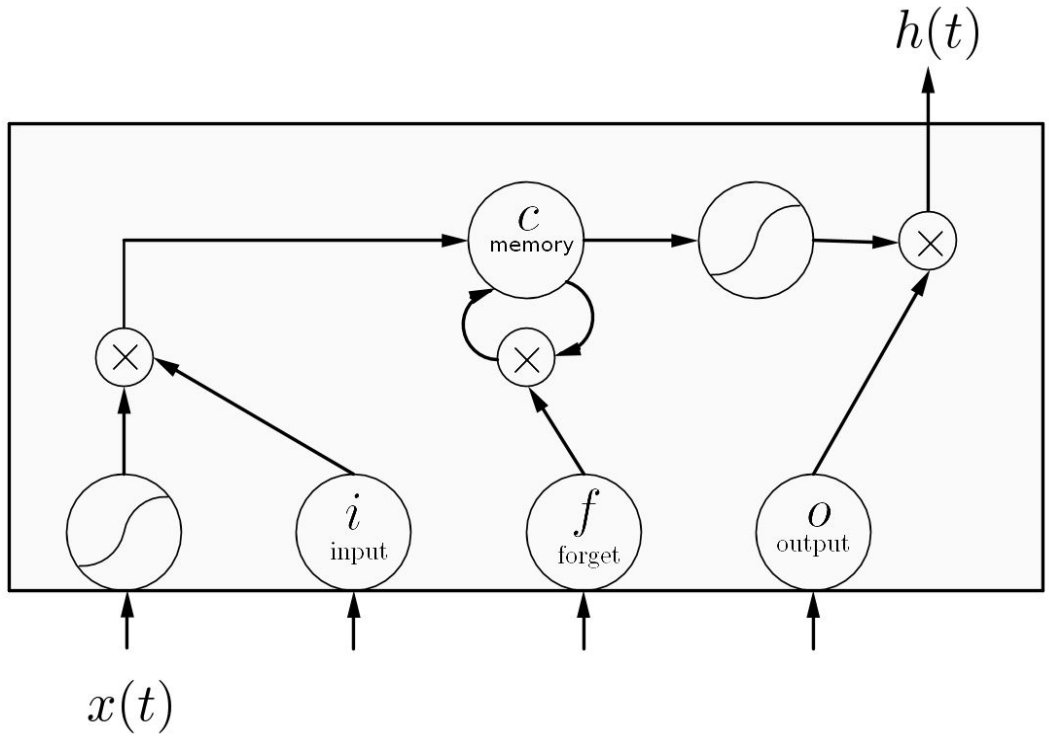
Introduction

Our proposal:

- Extend a previous approach for feature enhancement of ASR with additive noise by considering not only mfcc mapping from noisy to clean speech, but also include other acoustic features together with collections of deep neural networks.
- Use the audio signal components of: spectral, fundamental frequency, and energy.
- Incorporate Long Short-Term Memory Recurrent Networks

Long Short-Term Memory Networks (LSTM)

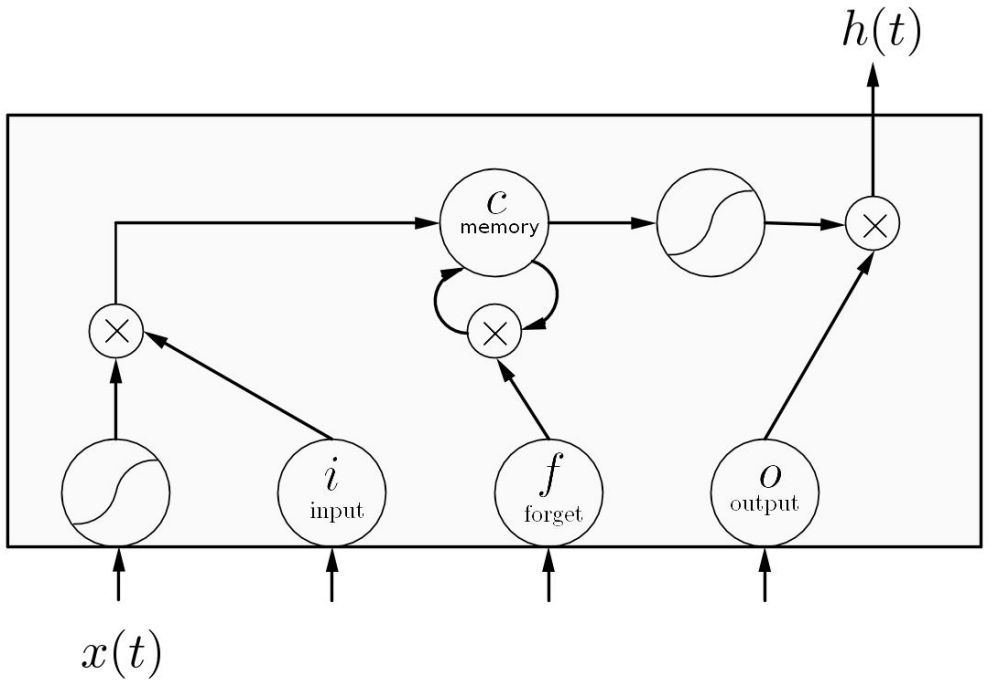
- The basic unit of the network is a Long Short-Term Memory block



Structure of a LSTM

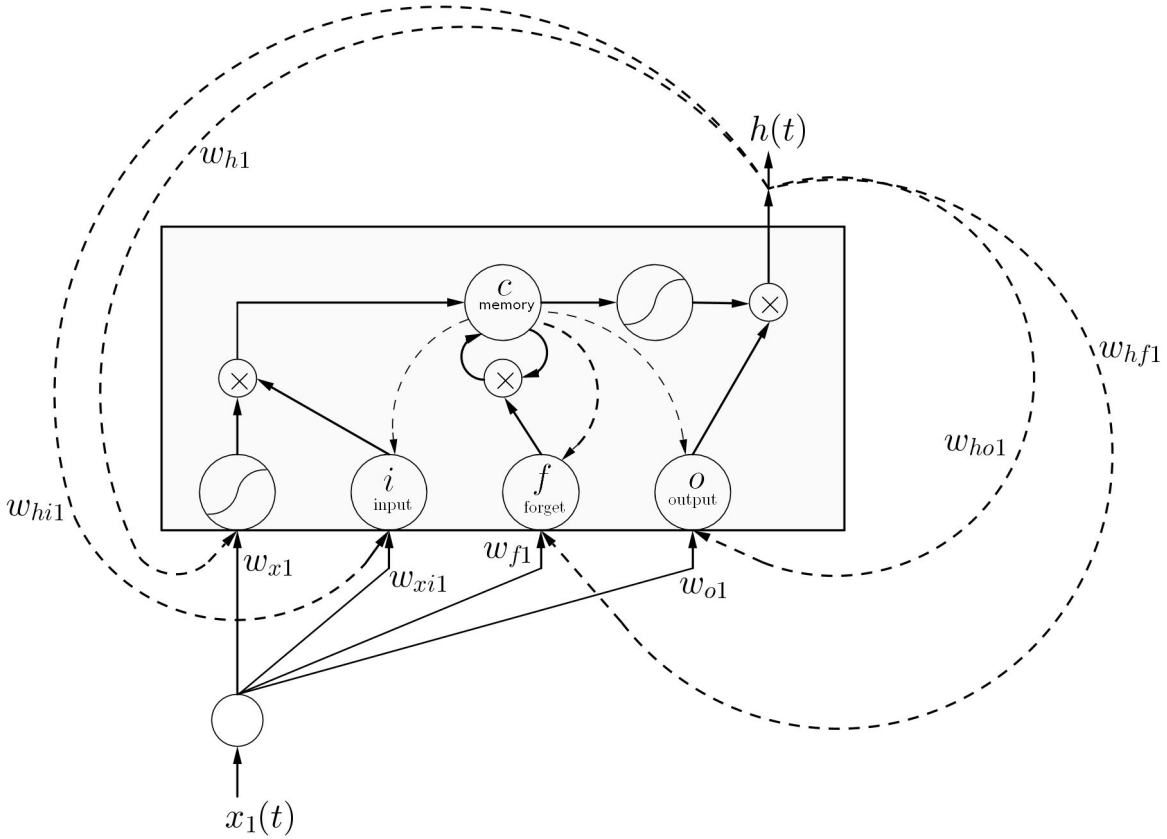
Each memory block has:

- Three gates: input, output, forget
- One input
- One output
- One memory cell



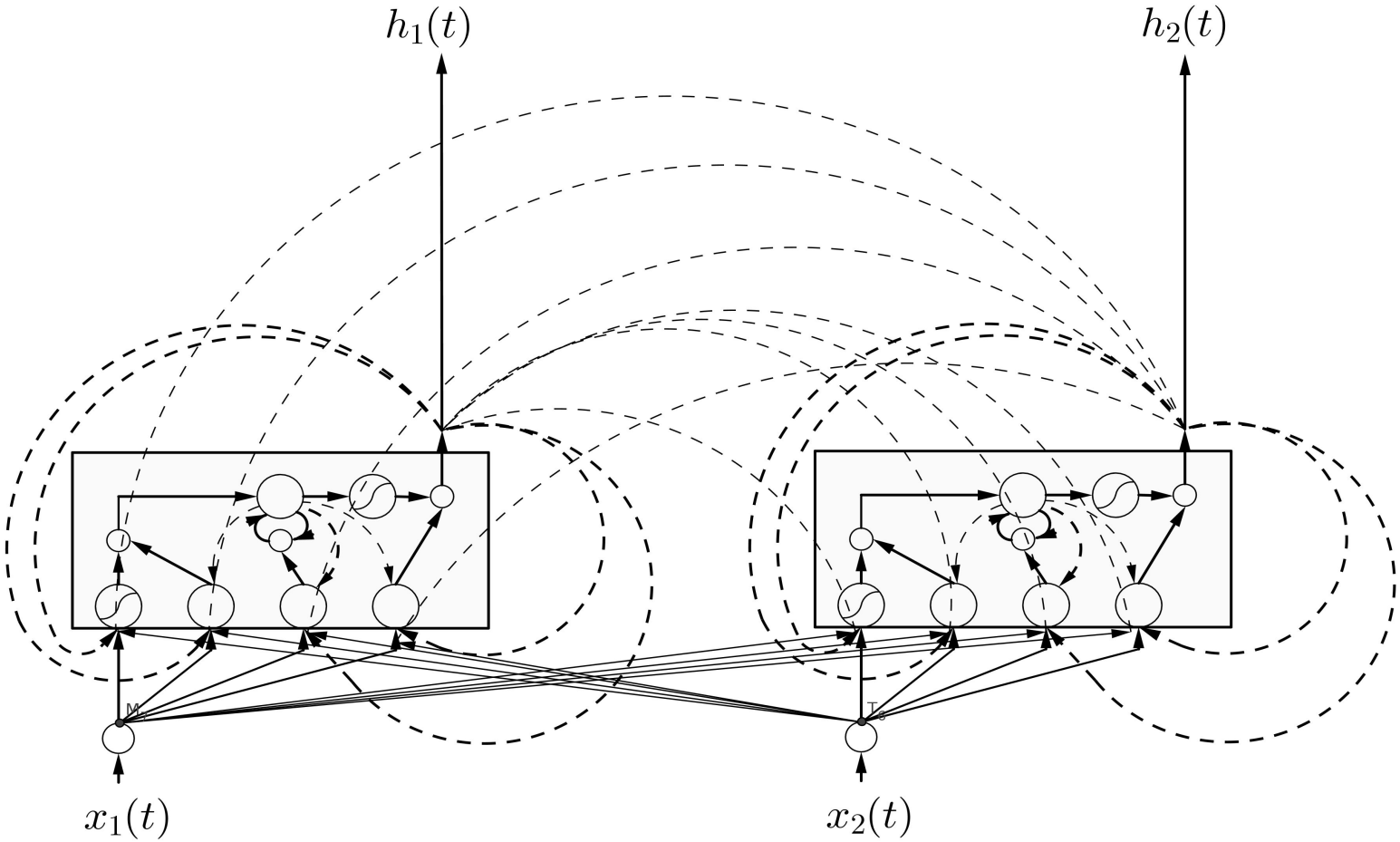
Structure of a LSTM

- There are recurrences from the outputs



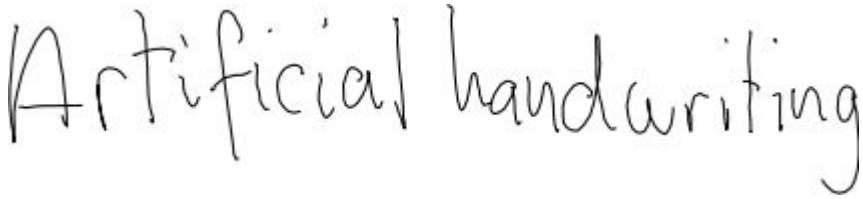
Structure of a LSTM

- There are recurrences from the outputs



Successful applications of LSTM

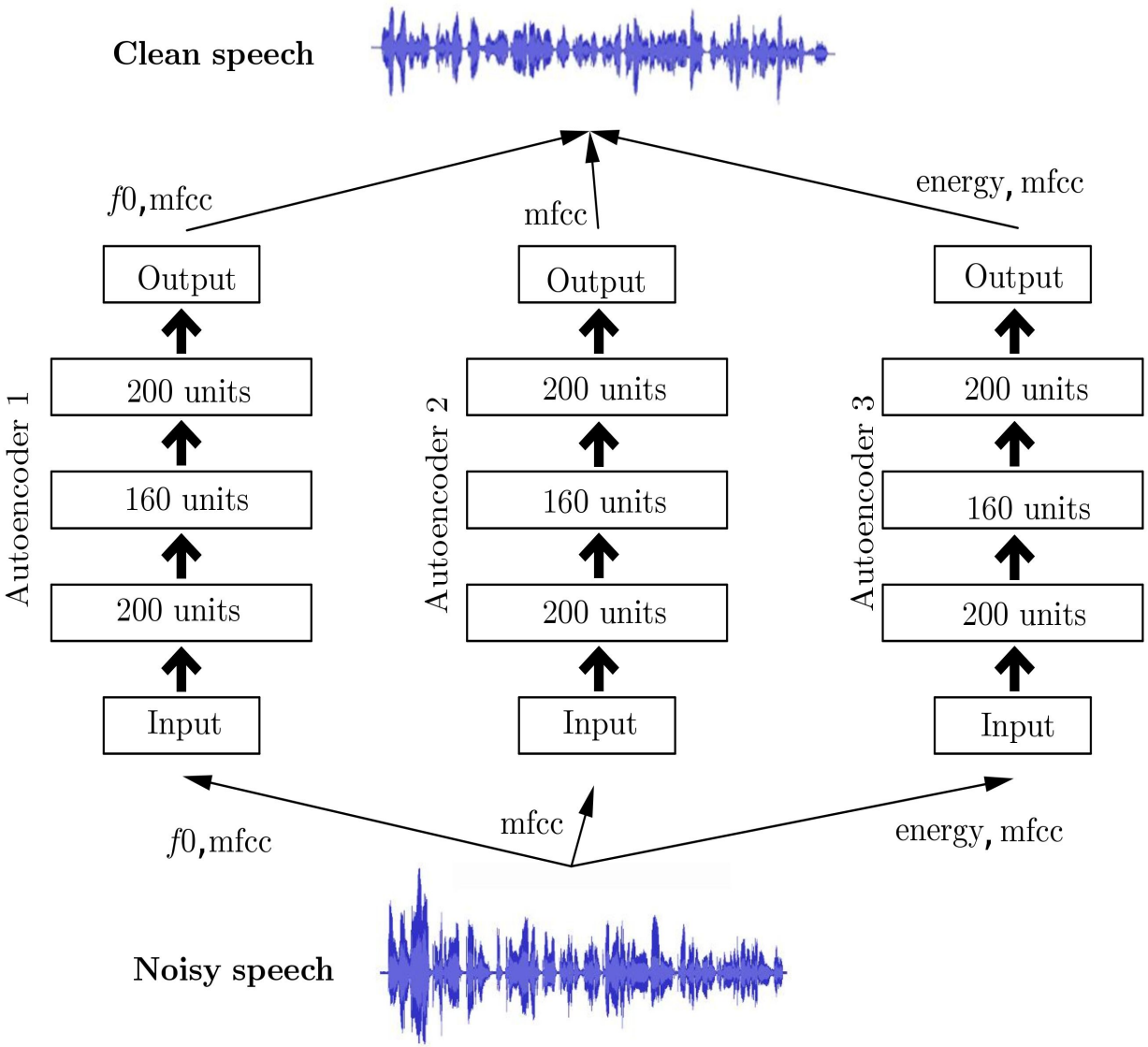
- Handwritten text recognition
- Automatic speech recognition
- Time series predictions
- Robot control
- Music composition



Description of the system

- In order to improve the accuracy of the ASR system on noisy utterances, we train a collection of LSTM networks, which map features of a noisy utterance x to a clean utterance y.
- The special kind of neural network we chose is called a denoising autoencoder:
  - First, an encoder transforms a n-dimensional input vector x into a hidden representation z.
  - Then, hidden representation z is mapped back to a reconstructed n-dimensional vector y in input space.

Description of the system



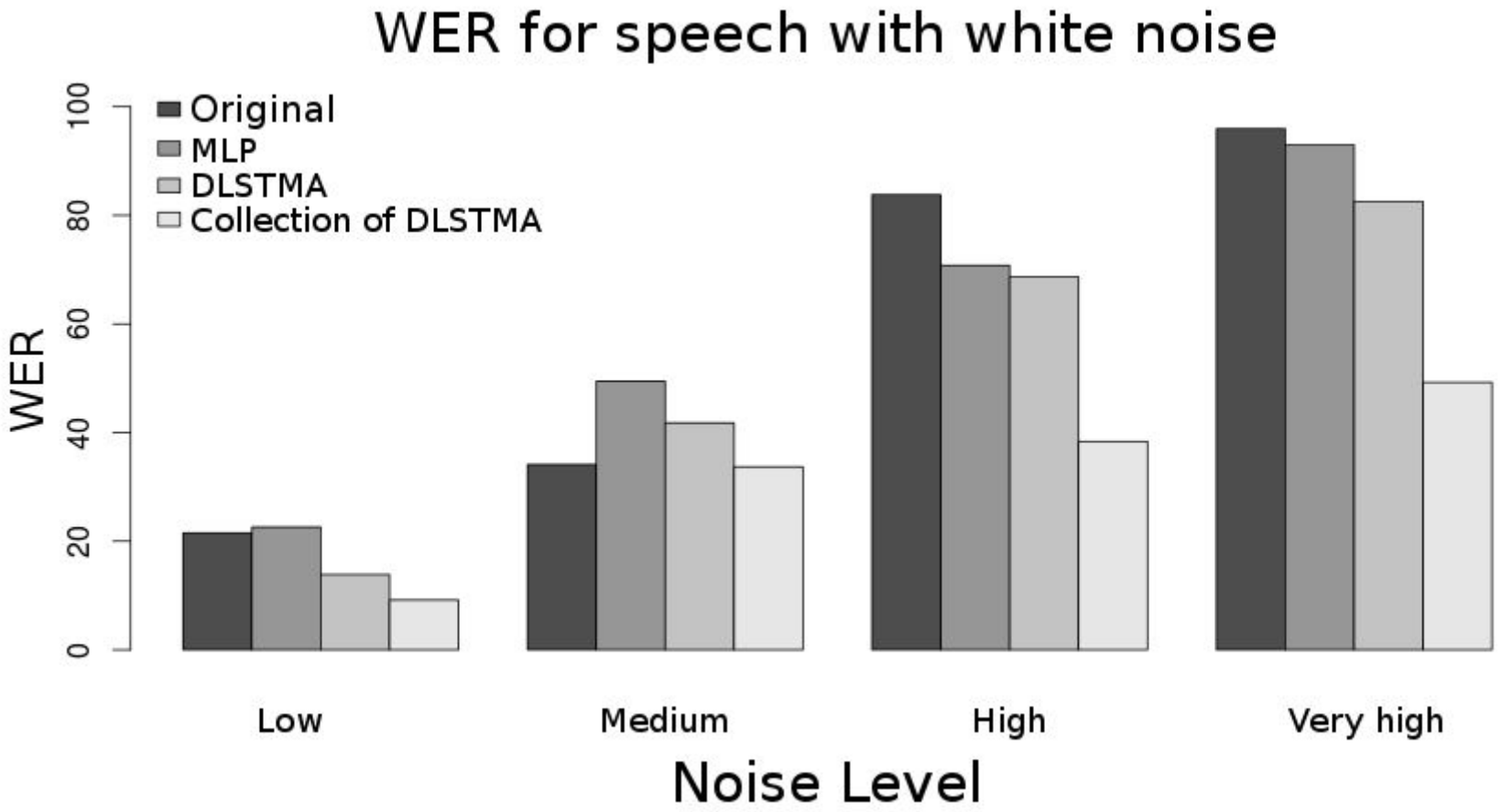
Experiments

- Database: 600 sentences from CMU ARTIC databases.
- We added several kinds of noise: white gaussian noise, brown noise, pink noise.
- Four levels of noise for noise type were added to the sentences, progressively affecting the ASR accuracy.
- 50 randomly selected sentences were tested in a state-of-the art online ASR, Speechmatics.

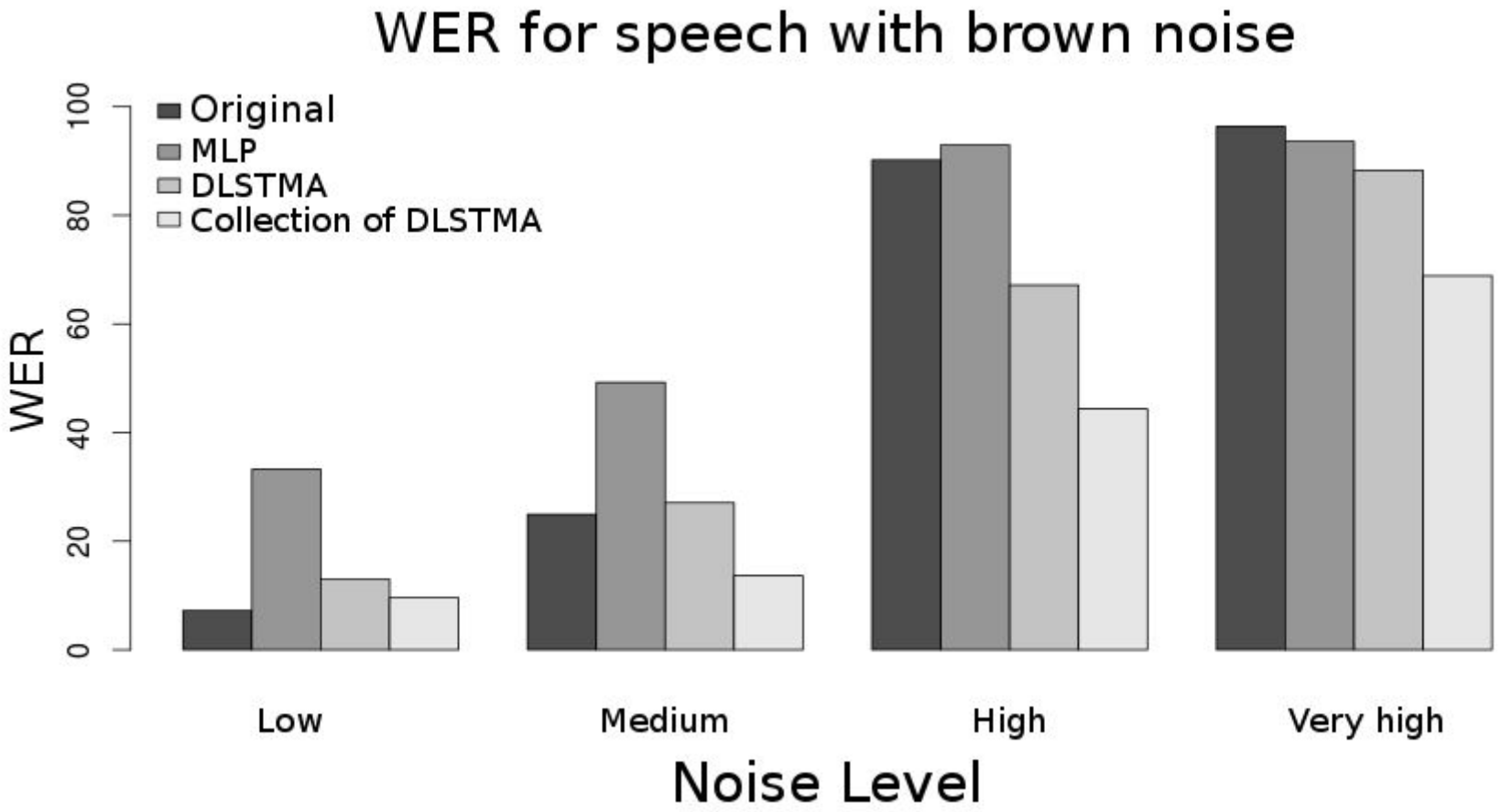
Experiments

- For comparison purposes, three different experiments were analyzed:
  1. transforming only the mfcc spectral features with one DLSTMA (base system)
  2. our proposal for transforming three acoustic features with separate LSTM autoencoders
  3. changing the LSTM units with simple sigmoid functions (Multi-Layer Perceptron)

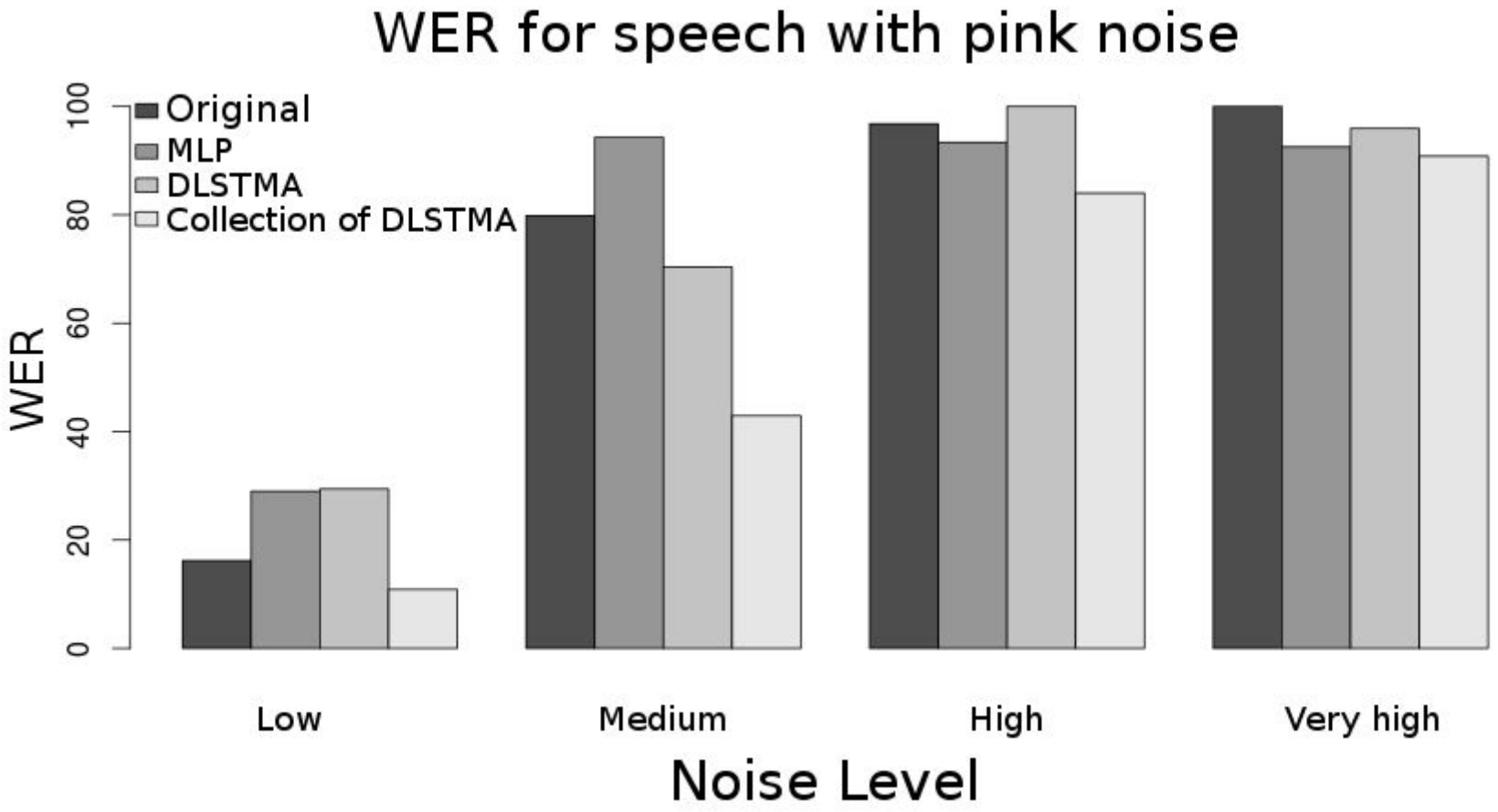
Results


WER for speech with white noise
Original
MLP
DLSTMA
Collection of DLSTMA
WER
0
20
40
60
80
100
Low
Medium
High
Very high
Noise Level

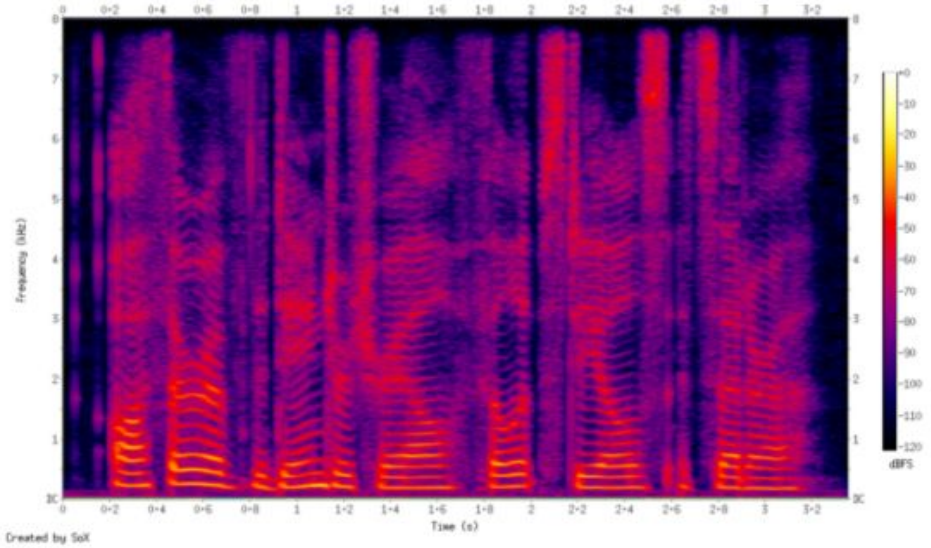
Results


WER for speech with brown noise
Original
MLP
DLSTMA
Collection of DLSTMA
WER
0
20
40
60
80
100
Low
Medium
High
Very high
Noise Level

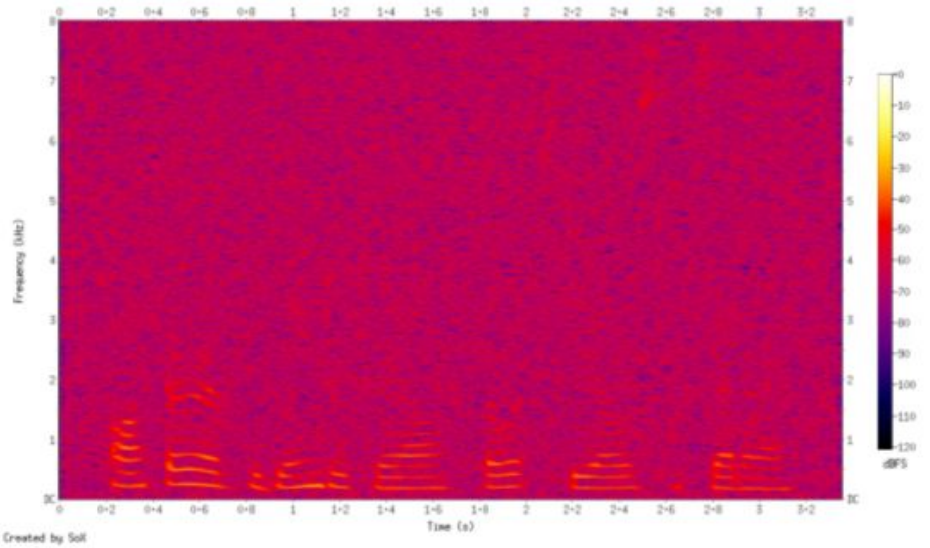
Results


WER for speech with pink noise
Original
MLP
DLSTMA
Collection of DLSTMA
WER
0
20
40
60
80
100
Low
Medium
High
Very high
Noise Level

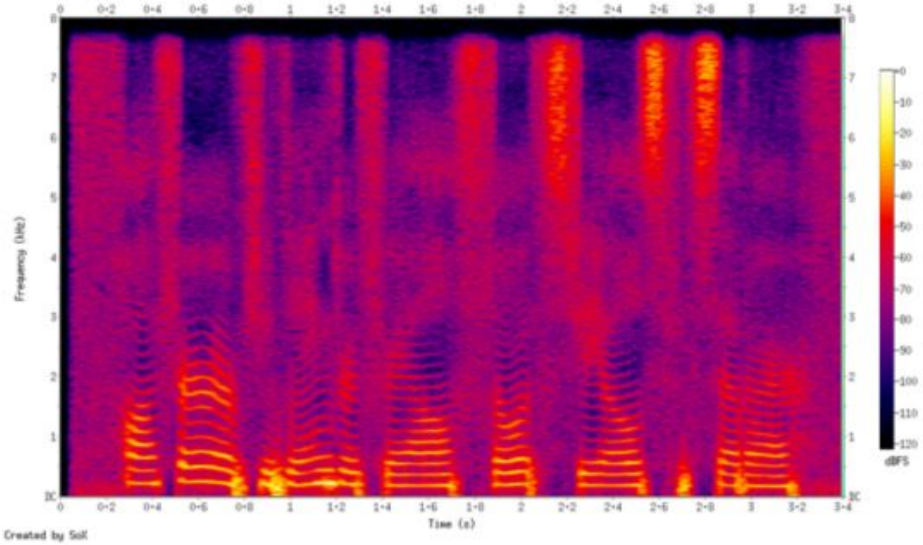
Results: Spectrograms



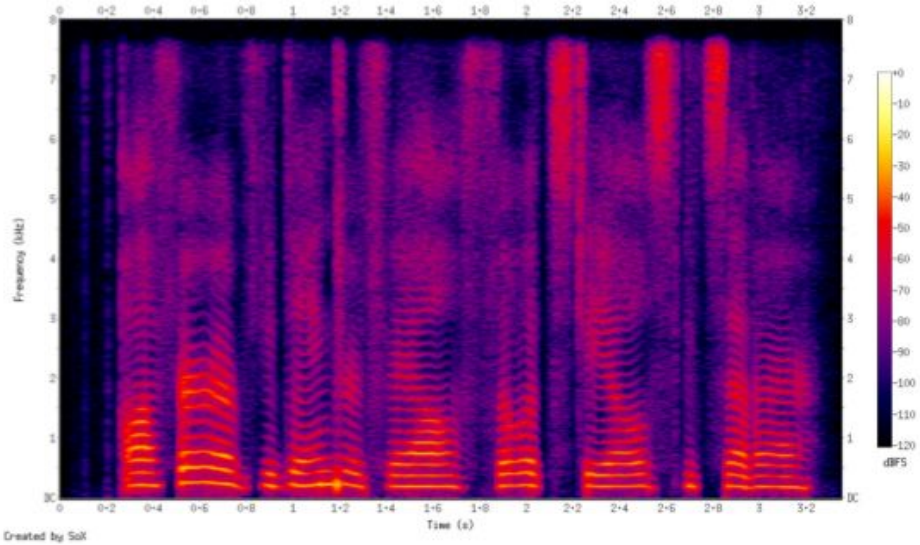
(a) Original



(b) Noisy



(c) DLSTMA



(d) Collection of DLSTMA

Examples

White noise

Pink noise

Conclusions

- In this paper, we have presented an extension of the DLSTMA, from mapping only mfcc features of a speech utterance, to a collection of DLSTMAs, each one composed of LSTM units.

- We evaluated the proposed system on data containing three different noise types, each at four levels, using a commercial ASR. The results showed that in almost all cases the collection of DLSTMAs improved the WER performance in comparison with the DLSTMA trained using only mfccs.

- Even for the case where the noise level is high enough to degrade WER to almost 100%, the collection of DLSTMAs lowers the WER signicantly. These are encouraging initial results.

Conclusions

- While this paper presents the preliminary results of our research using DLSTMAs for denoising, there are a number of questions left unanswered:
  1. How effective are other combinations of features, such as mfcc and energy only.
  2. Do other architectures for the autoencoder change the results we obtain. In particular, how can we reduce the training time.
  3. Do other types of noise affect the results we obtain with the proposed approach.
- We hope to provide some answers to these questions in future work.

Thank you for your attention