Speech Enhancement for a Noise-Robust Text-to-Speech Synthesis System using Deep Recurrent Neural Networks

Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki and Junichi Yamagishi
The Centre for Speech Technology Research (CSTR), University of Edinburgh, UK
National Institute of Informatics, Japan

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

THE UNIVERSITY of EDINBURGH

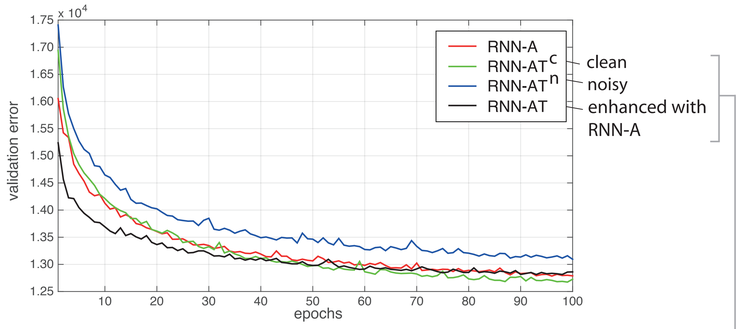
Introduction

It has been found that HMM-adapted voices built using clean speech are significantly better than voices built using noisy speech and speech enhanced using a conventional enhancement method.

Speech enhanced using neural networks has been found to be of high quality. For this reason we propose a noise-robust framework that uses a deep neural network to enhance data prior to training.

In this work, we enhance the vocoded parameters used to train the TTS acoustic model directly and evaluate the use of text-derived features as additional input of the network.

Speech enhancement methods



Methods differ in terms of the data used as input to the network :
- RNN-A : acoustic parameters from noisy speech
- RNN-AT : acoustic parameters from noisy speech + text features derived from aligned clean/noisy/enhanced speech

Target output:
- acoustic parameters derived from clean speech

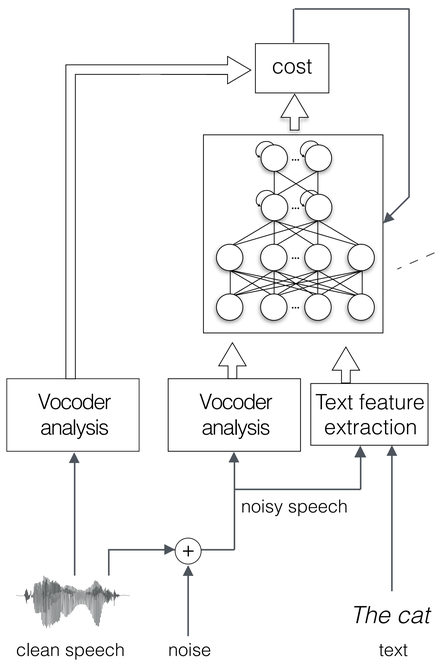
Acoustic parameters:
- 60 MCEP + 25 BAP + F0 + V/UV

Text-derived parameters:
- 327 binary + 37 integer + 3 continuous values

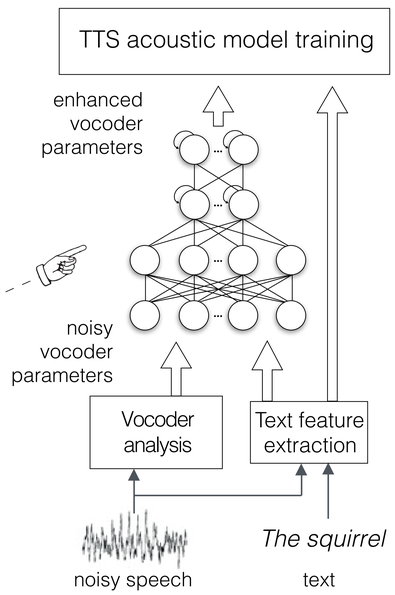
Architecture:
2 recurrent layers with 256 BLSTM units
2 feed-foward layers with 512 logistic units

Proposed framework

Training speech enhancement



Training Text-To-Speech system



Dataset

We created a noisy speech database* using:

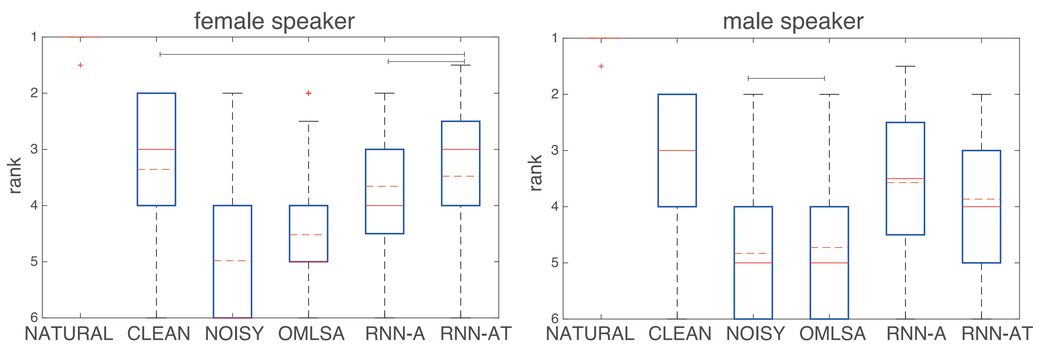
- the VCTK speech corpus (400 sentences / speaker)
train set 1 : 28 English speakers (~16hrs)
train set 2 : 56 English speakers (~32hrs)
test set: : 2 English speakers (~1hr)

- the Demand noise database
train set : 8 noises from Demand
2 artificially created noises
4 SNRs (15, 10, 5, 0 dB)
test set : 5 noises from Demand
4 SNRs (17.5, 12.5, 7.5, 2.5 dB)

In total, 40 (20) noisy conditions for training (testing), so that every 10 (20) sentences are from a different noisy condition.

Evaluation

We evaluated the quality of synthetic voices trained with the proposed framework against synthetic voices built with CLEAN, NOISY and data enhanced with the OMLSA method.



Subjective rank scores for the various synthetic voices were obtained via a MUSHRA test with 30 native English speakers, while objective distortion measures were calculated using the acoustic parameters that were used to train the synthetic voices.

We have found that the quality of synthesised speech produced with models that have been trained with enhanced data is significantly better.
For the female voice, these results were not significantly different from the models trained with clean data.

	MCEP (dB)	BAP (dB)	V/UV (%)	F_0 (Hz)
NOISY	9.86 / 10.68	2.62 / 2.41	9.55 / 7.88	40.27 / **4.38**
OMLSA	8.19 / 8.36	3.15 / 2.77	8.73 / 8.28	34.03 / 6.31
RNN-A	**4.59 / 5.05**	**1.86 / 1.72**	**2.46 / 2.15**	**24.90** / 8.43
RNN-AT	4.87 / 5.41	1.86 / 1.77	2.61 / 2.25	25.50 / 10.30

Table 1: Distortion measures calculated from the vocoded parameters extracted from the female / male test speaker.

This work was supported by EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology).

* The full NST research data collection may be accessed at http://hdl.handle.net/10283/786