Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

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

THE UNIVERSITY of EDINBURGH

CSTR

Introduction

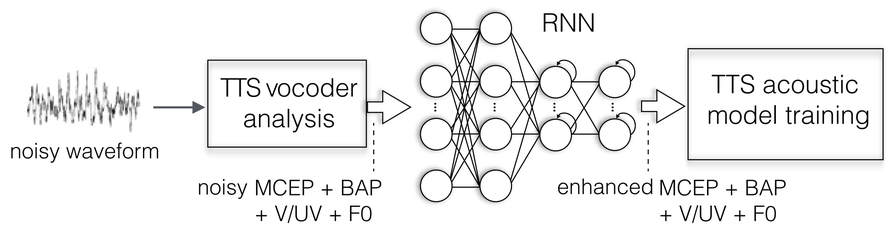
In this work we train a recurrent neural network with BLSTM layers to enhance noisy speech prior to TTS training.

Most speech enhancement methods operate on the magnitude spectrum or some parameterisation of it. Phase is obtained directly from the noisy speech waveform.

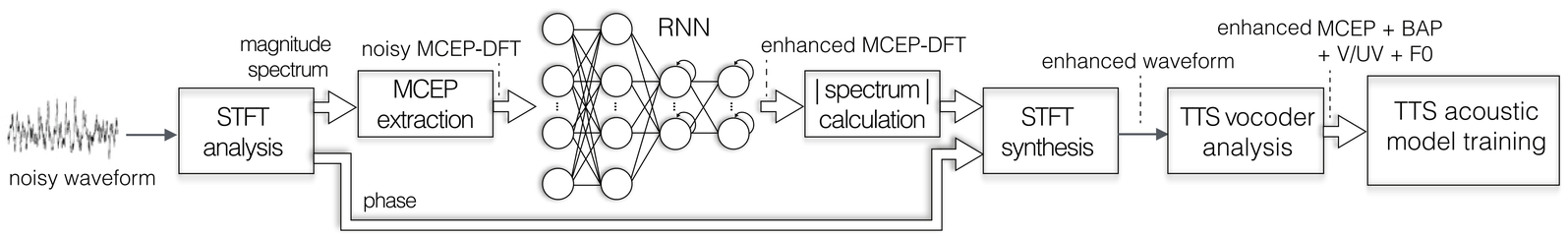
We compare this methodology with one where the neural network enhances the acoustic parameters extracted from a TTS style vocoder.

Proposed TTS training framework

RNN-V

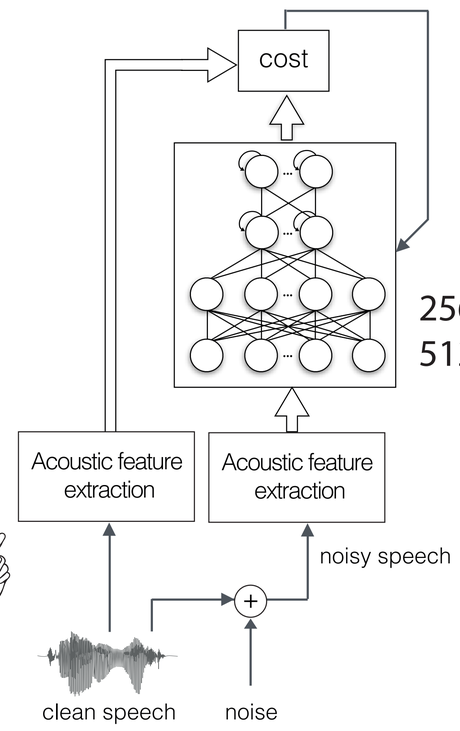


RNN-DFT



Speech enhancement methods

Training using database of clean and noisy speech



RNN-V uses STRAIGHT and SPTK:
- 60 Mel cepstral (MCEP) coefficients
- 25 band aperiodicity (BAP) values
- F0 value per frame (RAPT)
- voiced/unvoiced (V/UV) decision

RNN-DFT uses short-term Fourier transform (STFT):
- 87 MCEP extracted from magnitude spectrum

Dataset

We created a noisy speech database** using:

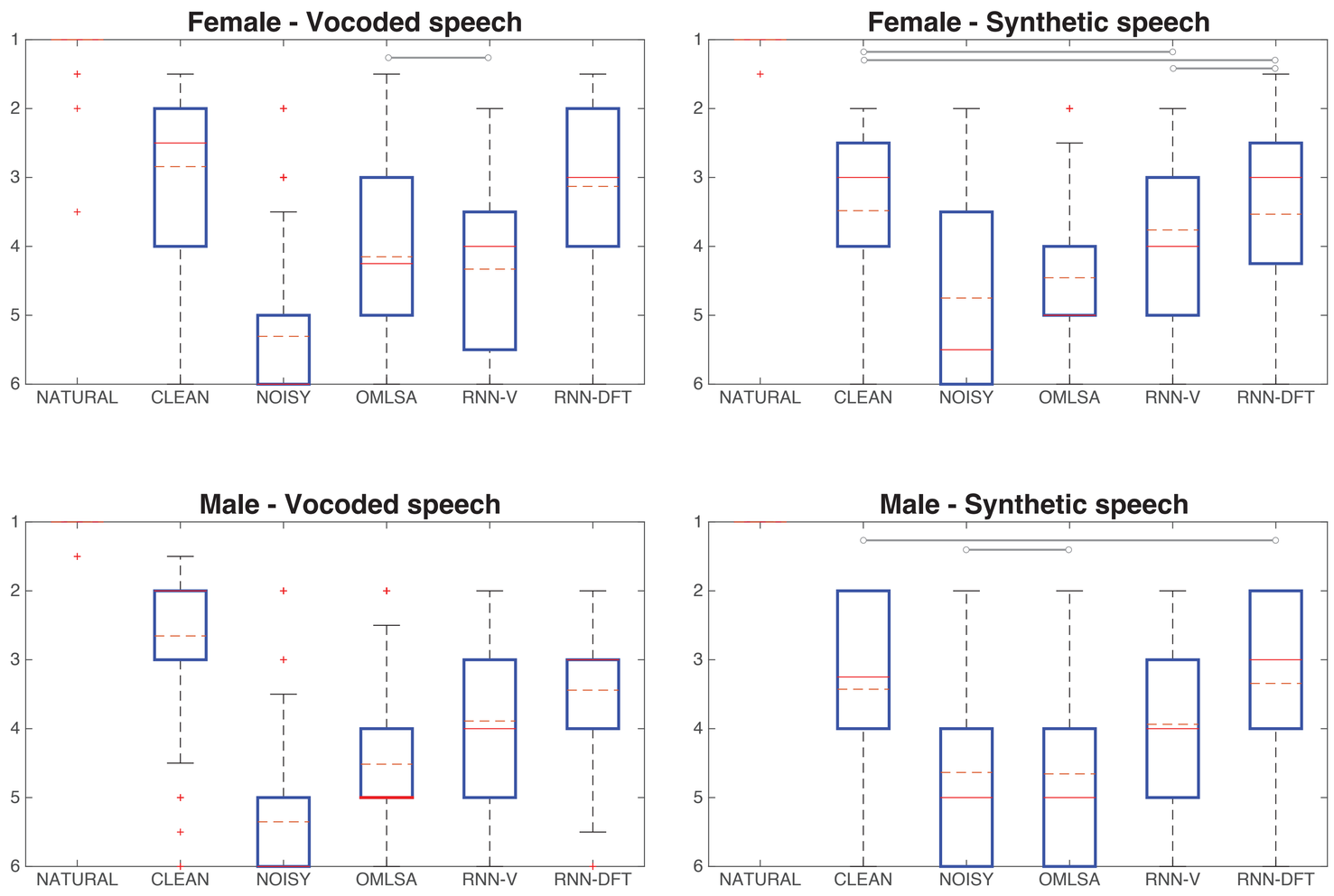
- the VCTK speech corpus (400 sentences / speaker)
train set : 56 English speakers (~32hrs)
test set: : 2 English speakers (~1hr)

- 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

Rank scores obtained in listening test



Subjective rank scores were obtained via a MUSHRA test with 24 native English speakers.

	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
CLEAN*	1.84 / 1.61	1.24 / 1.10	0.58 / 0.62	17.14 / 1.84
NOISY*	9.41 / 10.13	2.75 / 2.50	10.39 / 8.49	41.17 / 4.70
OMLSA	8.19 / 8.36	3.15 / 2.77	8.73 / 8.28	34.03 / 6.31
RNN-V	**4.59** / **5.05**	**1.86** / **1.72**	2.46 / **2.15**	24.90 / 8.43
RNN-DFT	4.90 / 5.22	2.44 / 2.32	**2.06** / 2.44	**22.59** / **3.31**

*Distortion measures calculated from the vocoded parameters of the female / male voice. * are STFT resynthesised clean and noisy signals.*

We have found that although MCEP distortion is higher, the RNN-DFT method was rated of a higher quality for both vocoded and synthetic speech for all speakers. The reconstruction process required in the RNN-DFT method does not negatively impact results.

The synthetic voices trained using data enhanced with this method were rated similar to voices trained with clean speech.

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

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