A maximum likelihood approach to multi-objective learning using generalized gaussian distributions for DNN-based speech enhancement

Shu-tong Niu[1], Jun Du[1], Li Chai[1], Chin-Hui Lee[2]

[1]University of Science and Technology of China, Hefei, Anhui, P.R.China

[2]Georgia Institute of Technology, Atlanta, GA, USA

ICASSP 2020

CONTENTS

CONTENTS

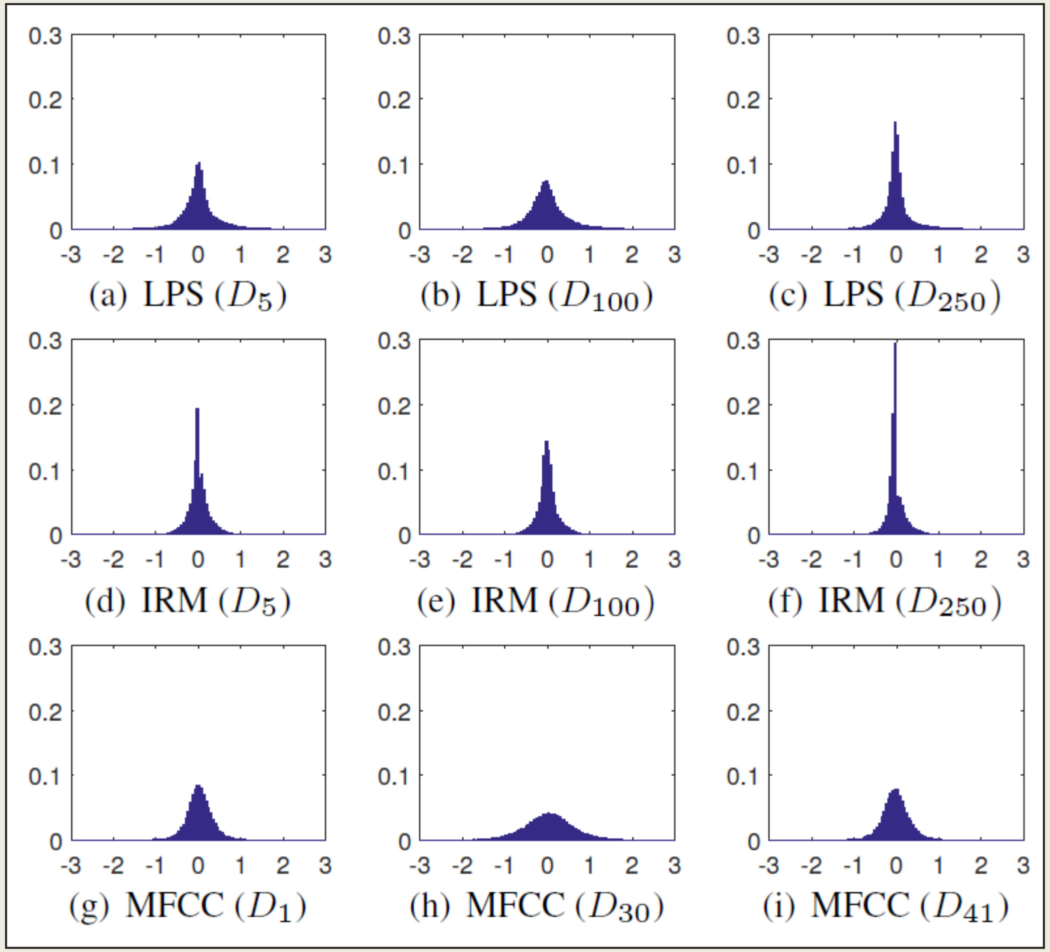
Background & Motivation

- The multi-objective learning framework based on MMSE criterion (MMSE-MOL-DNN) has been found one problem.
  - the values of the prediction errors vary greatly among different target feature types.

- Extend the maximum likelihood approach based on the generalized gaussian distributions (GGD) model to the multi-objective learning for DNN-based speech enhancement (ML-MOL-DNN).
  - Adjust the dynamic range of prediction error values on different targets automatically.
  - Update the shape factors of objective function in GGD model automatically.

Background & Motivation

The values of the prediction errors in MOL framework based on MMSE criterion vary greatly among different target feature types.



Three types of target features:

LPS: 257-dims

IRM: 257-dims

MFCC: 41-dims

Average prediction error:

	LPS error	IRM error	MFCC error	max/min
MMSEMOL	0.1637	0.0476	0.3283	6.897

CONTENTS

The proposed ML-MOL-DNN

- The derivation for ML-MOL-DNN
  - The conditional likelihood function of all types of target features can be seen as a product of the likelihood functions of single target feature (formula (a)):

$$p(\boldsymbol{x}_n|\boldsymbol{y}_n, \boldsymbol{W}, \boldsymbol{\Theta}) = \prod_{s=1}^{S} [p(\boldsymbol{x}_{n,s}|\boldsymbol{y}_{n,s}, \boldsymbol{W}, \boldsymbol{\Theta}_s)]^{\gamma_s}$$

  - In GGD model, the joint distribution for all dimensions at sample index n of s-th feature (formula (b)):

$$\begin{aligned} & p(\boldsymbol{x}_{n,s}|\boldsymbol{y}_{n,s}, \boldsymbol{W}, \boldsymbol{\alpha}_s, \boldsymbol{\beta}_s) \\ & = \prod_{d_s=1}^{D_s} \frac{\beta_{s,d_s}}{2\alpha_{s,d_s}\Gamma(\frac{1}{\beta_{s,d_s}})} \exp\left(-\left(\frac{|x_{n,s,d_s} - \hat{x}_{n,s,d_s}|}{\alpha_{s,d_s}}\right)^{\beta_{s,d_s}}\right) \end{aligned}$$

The proposed ML-MOL-DNN

- The derivation for ML-MOL-DNN
  - Assuming the distributions in formula (b) are drawn independently in different samples, the corresponding likelihood function is (formula (c)):

$$p(\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{W},\boldsymbol{\alpha},\boldsymbol{\beta}) = \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{y}_n,\boldsymbol{W},\boldsymbol{\alpha},\boldsymbol{\beta})$$
$$= \prod_{n=1}^{N}\prod_{s=1}^{S}\left[\prod_{d_s=1}^{D_s}\frac{\beta_{s,d_s}}{2\alpha_{s,d_s}\Gamma(\frac{1}{\beta_{s,d_s}})}\exp\left(-\left(\frac{|x_{n,s,d_s}-\hat{x}_{n,s,d_s}|}{\alpha_{s,d_s}}\right)^{\beta_{s,d_s}}\right)\right]^{\gamma_s}$$

  - Accordingly, the log-likelihood function can be written as (formula (d)):

$$\ln p(\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{W},\boldsymbol{\alpha},\boldsymbol{\beta}) = \sum_{n=1}^{N}\ln p(\boldsymbol{x}_n|\boldsymbol{y}_n,\boldsymbol{W},\boldsymbol{\alpha},\boldsymbol{\beta}) = \sum_{n=1}^{N}\sum_{s=1}^{S}\sum_{d_s=1}^{D_s}\gamma_s\times\ln\left(\frac{\beta_{s,d_s}}{2\alpha_{s,d_s}\Gamma(\frac{1}{\beta_{s,d_s}})}\right)$$
$$-\sum_{n=1}^{N}\sum_{s=1}^{S}\sum_{d_s=1}^{D_s}\gamma_s\times\left(\frac{|x_{n,s,d_s}-\hat{x}_{n,s,d_s}|}{\alpha_{s,d_s}}\right)^{\beta_{s,d_s}}$$

The proposed ML-MOL-DNN

- The derivation for ML-MOL-DNN
  - The corresponding loss function (formula (e)):

$$E(\mathbf{W}) = \sum_{n=1}^{N} \sum_{s=1}^{S} \sum_{d_s=1}^{D_s} \gamma_s \times \left(\frac{\left| x_{n,s,d_s} - \hat{x}_{n,s,d_s} \right|}{\alpha_{s,d_s}} \right)^{\beta_{s,d_s}}$$

  - The update formula of scale factor can be derived as (formula (f)):

$$\alpha_{s,d_s} = \left(\frac{\beta_{s,d_s}}{N} \sum_{n=1}^{N} \left| x_{n,s,d_s} - \hat{x}_{n,s,d_s} \right|^{\beta_{s,d_s}} \right)^{\frac{1}{\beta_{s,d_s}}}$$

ML-MOL-DNN can naturally reduce the difference in prediction error values among different types of target features by scale factor.

The proposed ML-MOL-DNN

- Update of shape factors (MLkurtosis-MOL-DNN)
  - The kurtosis of each dimension of s-th feature prediction error vector is defined as (formula (g)):

$$\mathrm{Kurt}[e_{s,d_s}] = \mathrm{E}\left[\left(\frac{e_{s,d_s} - \mu_{s,d_s}}{\sigma_{s,d_s}}\right)^4\right] = \frac{\mathrm{E}[(e_{s,d_s} - \mu_{s,d_s})^4]}{(\mathrm{E}[(e_{s,d_s} - \mu_{s,d_s})^2])^2}$$

  - Meanwhile, the kurtosis above can also be calculated as follows in GGD (formula (h)):

$$\mathrm{Kurt}[e_{s,d_s}] = \frac{\Gamma(5/\beta_{s,d_s})\Gamma(1/\beta_{s,d_s})}{\Gamma(3/\beta_{s,d_s})^2} - 3$$

CONTENTS

Experiments and Results

- Experimental setup
  - Data set: WSJ0 corpus
  - DNN: 3 hidden layers with 2048 nodes for each layer
  - All models:

    MMSE-MOL-DNN

    ML-MOL-DNN

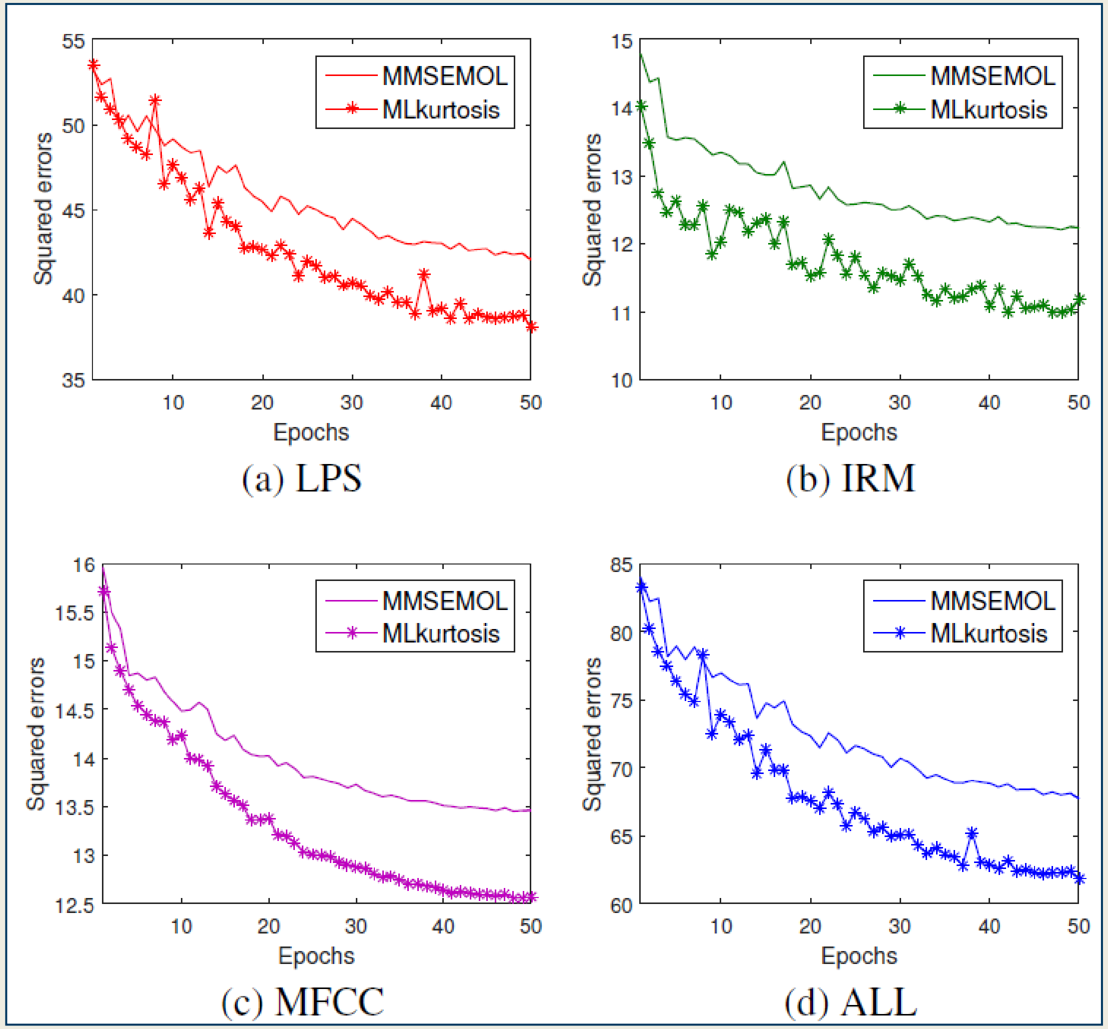
    ML-MOL-DNN with shape factors 1 (ML111-MOL-DNN)

    ML-MOL-DNN with shape factors 2 (ML222-MOL-DNN)

    ML-MOL-DNN with updating shape factors (MLkurtosis- MOL-DNN)

Experiments and Results

➢ Evaluation on ML-MOL-DNN



- The comparison of learning curves between MMSE-MOL-DNN (denoted as MMSEMOL) and MLkurtosis-MOL-DNN (denoted as MLkurtosis).
- The MLkurtosis-MOL-DNN can achieve better convergence than MMSE-MOL-DNN in all types of target features.

Experiments and Results

Evaluation on ML-MOL-DNN

	LPS error	IRM error	MFCC error	max/min
MMSEMOL	0.1637	0.0476	0.3283	6.897
ML222	0.4724	0.4767	0.4980	1.054
ML111	0.9365	0.9509	0.9990	1.067
MLkurtosis	1.1111	1.1060	0.7621	1.458

- The comparison of average prediction error values among MMSE-MOL-DNN (denoted as MMSEMOL), ML222-MOL-DNN (denoted as ML222), ML111-MOL-DNN (denoted as ML111) and MLkurtosis-MOL-DNN (denoted as MLkurtosis).
- ML-MOL-DNNs can better control the dynamic range of prediction error values among different types of target features.

Experiments and Results

➢ Shape factors update results

Epoch	1 (Init-MMSE)	20	40	50
LPS	0.9365	0.8744	0.8553	0.8542
IRM	1.1854	0.9105	0.8847	0.8732
MFCC	1.3673	1.3404	1.3268	1.3245

- The change of shape factors after dimension averaging in MLkurtosis-MOL-DNN under different types of features (LPS, IRM and MFCC) with respect to the epoch.
- The shape factors are gradually becoming smaller during the training process, this means the kurtosis becomes larger.

Experiments and Results

➢ Overall comparison

- PESQ for measuring speech quality.
- STOI for measuring speech intelligibility.
- SSNR (dB) and LSD (dB) for evaluating signal differences in the time domain and the frequency domain.

SNR(dB)		-5	5	15	Ave
PESQ	MMSEMOL	1.650	2.540	3.136	2.463
	ML222	1.762	2.616	3.192	2.544
	ML111	1.834	2.672	3.242	2.603
	MLkurtosis	1.838	2.680	3.253	**2.611**
STOI	MMSEMOL	0.680	0.891	0.968	0.857
	ML222	0.689	0.903	0.974	0.867
	ML111	0.682	0.904	0.976	0.866
	MLkurtosis	0.686	0.905	0.976	**0.868**
SSNR	MMSEMOL	-3.222	0.661	4.966	0.767
	ML222	-3.385	0.689	5.274	0.819
	ML111	-3.341	1.020	5.888	1.149
	MLkurtosis	-3.125	1.179	6.017	**1.316**
LSD	MMSEMOL	6.588	3.380	1.856	3.803
	ML222	6.676	3.570	1.860	3.924
	ML111	6.457	3.432	1.847	3.793
	MLkurtosis	6.087	3.228	1.737	**3.572**

CONTENTS

Conclusion and Future Work

- Conclusion
  - Compared with the MMSE-MOL-DNN, the ML-MOL-DNNs can adjust the prediction error values under different types of features automatically and achieve performance improvement.
  - Compared with the ML-MOL-DNNs with fixed shape factors, the proposed MLkurtosis-MOL-DNN can achieve consistent improvements on four evaluation metrics.

- Future work
  - Introducing more types of complementary features and apply the ML-MOL-DNN approach to multi-objective learning and ensembling models with compact neural network architectures.

CONTENTS

Thanks for listening!

If you have any questions about this paper, welcome to ask and I will answer them.