Effectiveness of Single-Channel BLSTM Enhancement for Language Identification

Sibbern Frederiksen, Peter; Villalba, Jesus; Watanabe, Shinji; Tan, Zheng-Hua; Dehak, Najim

Published in:
Interspeech 2018

DOI (link to publication from Publisher):
10.21437/Interspeech.2018-2458

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

Interspeech 2018
2-6 September 2018, Hyderabad

Effectiveness of Single-Channel BLSTM Enhancement for Language Identification

Peter Sibbern Frederiksen[1], *Jesús Villalba*[2], *Shinji Watanabe*[2], *Zheng-Hua Tan*[1], *Najim Dehak*[2]

[1]Department of Electronic Systems, Aalborg University, Denmark
[2]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD

psf@ieee.org, jvillal7@jhu.edu, shinjiw@jhu.edu, zt@es.aau.dk, ndehak3@jhu.edu

Abstract

This paper proposes to apply deep neural network (DNN)-based single-channel speech enhancement (SE) to language identification. The 2017 language recognition evaluation (LRE17) introduced noisy audios from videos, in addition to the telephone conversation from past challenges. Because of that, adapting models from telephone speech to noisy speech from the video domain was required to obtain optimum performance. However, such adaptation requires knowledge of the audio domain and availability of in-domain data. Instead of adaptation, we propose to use a speech enhancement step to clean up the noisy audio as preprocessing for language identification. We used a bi-directional long short-term memory (BLSTM) neural network, which given log-Mel noisy features predicts a spectral mask indicating how clean each time-frequency bin is. The noisy spectrogram is multiplied by this predicted mask to obtain the enhanced magnitude spectrogram, and it is transformed back into the time domain by using the unaltered noisy speech phase. The experiments show significant improvement to language identification of noisy speech, for systems with and without domain adaptation, while preserving the identification performance in the telephone audio domain. In the best adapted state-of-the-art bottleneck i-vector system the relative improvement is 11.3% for noisy speech.

Index Terms: speech enhancement, BLSTM, language recognition, NIST LRE17.

1. Introduction

Language recognition refers to the process of automatically detecting the language spoken in a speech utterance. Its applications range across customized speech recognition, multilanguage translation, service customization and forensics [1]. The focus of research in the field has been on developing recognition methods to improve the performance of general systems, while little attention has been given to improving the noise-robustness of language recognition systems.

NIST Language recognition evaluations (LRE) has played an instrumental role in driving language recognition research over the years and LRE constantly increases the challenge level of its evaluations. The most recent LRE 2017 evaluation [2] presents a new scenario with a significant mismatch between training and evaluation data. The training dataset consists of a large amount of narrow-band telephone speech, which is in line with previous evaluations. However, the evaluation dataset consists of a combination of narrow-band telephone data and wide-band data from Internet videos. Furthermore, the LRE 2017 organizers provide a limited amount of in-domain development data for model adaptation and calibration purposes. While telephone speech contains low levels of noise and reverberation, we observed that the video data are severely degraded by babble noise, music and reverberation.

Single-channel speech enhancement (SE) can be used as preprocessing to mitigate the aforementioned degradation and reduce the mismatch between training and evaluation data. SE has been widely used as preprocessing for speech applications, such as automatic speech recognition (ASR) [3], speaker verification [4], mobile communications and hearing aids [5]. In this study we investigate the effectiveness of utilizing single-channel SE to improve the noise-robustness of a language recognition system.

It has been experimentally shown that applying ideal binary mask in the time-frequency domain is able to improve speech intelligibility of noisy speech signals for both normal hearing and hearing impaired listeners with various noise types [6]. Various ideal ratio masks have become preferable over ideal binary mask in recent studies [5, 7, 8]. In [9, 10] a DNN is trained to predict clean speech from noisy speech without the use of a mask by casting it as a regression problem. A long short-term memory (LSTM) network has shown to outperform feed-forward DNN methods, when used as preprocessing for noise robust ASR [3], and the bidirectional extension of LSTM (BLSTM) achieves further improvement [8]. This paper follows the success of the BLSTM SE method, and applies it to a language recognition system. The BLSTM SE is processed in the time-frequency domain, but only deals with the magnitude while the phase component remains corrupted, similar to the other DNN-based SE methods. The method internally predicts a mask from BLSTM, and the predicted mask is multiplied by the noisy speech magnitude, which yields the enhanced magnitude. The network is trained with the mean square error criterion between the clean and enhanced magnitudes. In BLSTM SE (and other DNN-based enhancement), only additive noise is considered, where the noise source is extracted from in-domain data with limited size in our setup. The effectiveness of BLSTM SE on the language identification is evaluated by a state-of-the-art bottleneck i-vector LRE system, where BLSTM SE is used as preprocessing of the LRE system [11].

To validate the effectiveness of BLSTM SE methods, we also compare our SE with the optimally-modified log-spectral amplitude (OM-LSA) speech estimator with the improved minima controlled recursive averaging (IMCRA) noise estimator [12], [13]. OM-LSA is a well-known signal processing method that does not require data-driven training and adaption stages.

2. Speech Enhancement system

2.1. Speech enhancement system evaluation

To verify if the SE model itself works, it should be evaluated with listening test to fully evaluate the performance. However,

10.21437/Interspeech.2018-2458

to quickly and cheaply evaluate development work, a number of objective algorithms are used instead. These algorithms are designed to emulate human evaluation of SE, with a higher score being better. The first is perceptual evaluation of speech quality (PESQ) which is meant to emulate human evaluation of the pleasantness of listening to the speech audio [14, 15]. The PESQ score is defined in the interval $[-.5, 4.5]$. Another is the short-time objective intelligibility measure (STOI) [16] and the extended STOI (eSTOI) [17] meant to emulate human word comprehension, i.e. a human word error rate if you will. They are defined in the interval $[0, 1]$. Compared with the above measures, signal-to-distortion ratio (SDR) aims to evaluate the audio source separation quality, but it is still used as a speech enhancement measure by regarding enhanced data and subtracted noise data as sources [18], which is defined in the interval $(-\infty, \infty)$. The enhancement algorithms in this paper are evaluated with these measures by comparing their enhanced signals to the original uncorrupted signals. The need for uncorrupted signals restricts this evaluation form to simulated data.

2.2. Speech enhancement dataset

This section describes our speech enhancement dataset, which is generated for the purpose of speech enhancement experiments on the LRE17 task. The corruption of a speech signal can be seen as two types: additive and convolutional. Additive noise is typically independent of background noise, whereas convolutional noise can come from reverberation in rooms, and will be correlated with the speech signal. In this study we only consider additive noise, where we adopt the signal model for the noisy speech signal y as

$$y(t) = s(t) + n(t) \tag{1}$$

where s is the speech signal and n is the noise signal.

In the dataset, noisy speech signals are created for each SNR level of {-3, 0, 3, 6, 9, 12, 15} dB equally. Simple voice activation detection is used to account for silence regions in speech signals, when calculating the energy. The training and validation datasets have no overlap and are split into 90 and 10 percents, respectively. The speech signals are taken from the LRE17 training set consisting of 2069 hours of telephone conversations. They are all sampled with 8 kHz with a mix of precision encodings. The noise signals come from the audio signals in the LRE17 development video domain. Most of these audios except for the talk shows contain noisy speech segments. Examples of background noise are babble, television, clapping, laughing, kitchen work and wind. The dataset also includes signals with reverberation which are left as is. Speech segments in these signals have been manually marked as speech intervals. A noise signal is a concatenation of all non-speech intervals in a noisy speech signal. The concatenation is performed with 128 samples of overlap and using a Hanning window of length 256 samples. Noise intervals less than 125 milliseconds are discarded. This results in 6.6 hours of noise signals, which are expected to be closer to the noise sources in the target domain. Note that these noise signals potentially contain background speech since some recordings are annotated with segments of dominant speakers, and the aforementioned approach unintentionally includes speech segments of non-dominant speakers as noises. The noise signals are repeated to create 2069 hours of speech and noisy speech signal pairs, which are then cut into 5 seconds long segments.

Now we describe the input feature for our BLSTM speech enhancement system. First, the noisy speech signal in the time domain is transformed using short time Fourier transform (STFT) into a time-frequency domain spectrogram. We use a modified Hanning window w of length 256 samples and an overlap/step of 128 samples.

$$w[k] = \frac{1}{2} + \frac{1}{2}\cos\left(2\pi\frac{\left(k - \frac{K-1}{2}\right)}{K}\right),\ k = 0, 1, \ldots, K-1 \tag{2}$$

After STFT, we extract the 100-bin log Mel filterbank coefficients. Finally, the filterbank coefficients are normalized using the global mean and variance computed over the training samples. With these input features, the BLSTM model outputs the mask for each time-frequency bin, which is then multiplied by the original noisy speech magnitude spectrogram to get the enhanced magnitude spectrogram as an approximation of the uncorrupted speech. The time domain signal of the enhanced speech can be synthesized by using the inverse STFT, where the phase is taken from the original noisy speech spectrogram.

2.3. Model and training

We adopt BLSTM-based model architecture as speech enhancement. BLSTM recurrent neural networks offer an elegant way to incorporate context information, instead of explicitly choosing the context based on feed-forward neural networks. The baseline BLSTM has 2 layers with 384 hidden units with an additional fully connected layer to transform concatenation of the bi-directional output of 768 units to 129 frequency bins for each time step. A sigmoid activation function is applied to constrain the mask to the interval from 0 to 1. By following the previous work of [8], we consider magnitude time-frequency approximation instead of a mask approximation for the objective function. First, we consider the following distance function $D(\cdot)$:

$$D(\hat{a} \circ |Y| - a \circ |Y|) \tag{3}$$

where a is the ideal mask, $\circ$ is element-wise multiplication, $\hat{a}$ is the approximated mask obtained by BLSTM, and $|Y|$ is the magnitude time-frequency representation of the noisy speech. For the sake of simplicity, we omit the time-frequency index in the formulation. Several masks have been proposed and an overview can be found in [8]. The SE system uses the ideal amplitude mask a_{iam}

$$a_{\text{iam}} = \frac{|S|}{|Y|} \tag{4}$$

where $|S|$ is the magnitude time-frequency representation of the uncorrupted speech. Equation (3) reduces to

$$D(\hat{a} \circ |Y| - a_{\text{iam}} \circ |Y|) = D(\hat{a} \circ |Y| - |S|). \tag{5}$$

With this representation, the mean squared error (MSE) based objective function is represented as:

$$\underset{\theta \in \mathbb{R}^n}{\text{minimize}} \quad \frac{1}{M}\sum_{m=0}^{M-1}\left(\hat{a} \circ |Y| - |S|\right)^2 \tag{6}$$

with M being the number of total samples in a minibatch, n being the number of BLSTM parameters, and θ is the BLSTM parameter space. Adam is used as a stochastic minimizer. The model is implemented in the PyTorch framework.

3. Language recognition system

Figure 1 shows the pipeline of a state-of-the-art i-vector language recognition system with an additional speech enhancement step. Following, we explain each of the steps.

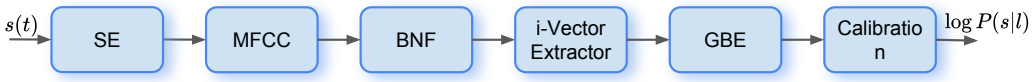

Figure 1: *Proposed i-Vector language recognition system with single-channel enhancement.*

3.1. Feature extraction

We computed 20 dimensional Mel-frequency cesptral coefficients (MFCCs) from the noisy/enhanced speech signal. From MFCCs, we obtained phonetic discriminant bottleneck features (BNF). The bottleneck network was trained on 1800 hours of Fisher English using Kaldi NNet2 [19]. The network consisted of 7 hidden layers, the 6th layer was an 80 dimensional linear bottleneck layer; the rest were TDNN layers with p-norm activations with input/output dimension equal to 3500/350. The output layer was a softmax that classifies 5577 senone acoustic units. Short-term mean and variance normalization was applied with 3 second sliding window and silence frames were removed.

3.2. i-Vectors

The i-vector paradigm [20] transforms the sequence of BNFs into a fixed-dimensional embedding. Each speech segment is modeled by a Gaussian mixture model (GMM) whose super-vector mean $\mathbf{M}$ is assumed to be

$$\mathbf{M}_s = \mathbf{m} + \mathbf{T}\mathbf{w}_s \tag{7}$$

where $\mathbf{m}$ is the GMM-UBM mean super-vector, $\mathbf{T}$ is a low-rank matrix and $\mathbf{w}$ is a standard normal distributed vector. $\mathbf{M}$ defines the total variability space, i.e. the directions in which we can move the UBM to adapt it to a specific segment. The GMM-UBM represents the speaker-independent distribution of feature vectors. The *maximum a posteriori* (MAP) point estimate of $\mathbf{w}$ is the i-vector embedding.

3.3. Gaussian back-end (GBE) with domain adaptation

We used a linear Gaussian classifier to compute the language log-likelihood scores from the i-vectors. This back-end models each class with a Gaussian where the within-class covariance matrix is shared across languages. We equalized the weight of each language in the covariance estimation.

For domain adaptation, we computed the *a priori* back-end means and covariances on out-domain data and applied *Maximum a posteriori* (MAP) adaptation using in-domain data. The adaptation equations for the Gaussian classifier are

$$\boldsymbol{\mu}_l = \alpha_l \boldsymbol{\mu}_{\mathrm{ML}_l} + (1-\alpha_l)\boldsymbol{\mu}_{0_l} \qquad l = 1, \ldots, L \tag{8}$$

$$\begin{aligned}\mathbf{S}_{\mathbf{W}} = \frac{1}{L}\sum_{l=1}^{L} \big[&\beta_l \mathbf{S}_{\mathrm{ML}_l} + (1-\beta_l)\mathbf{S}_0 \\ &+\beta_l(1-\alpha_l)\left(\boldsymbol{\mu}_{\mathrm{ML}_l} - \boldsymbol{\mu}_{0_l}\right)\left(\boldsymbol{\mu}_{\mathrm{ML}_l} - \boldsymbol{\mu}_{0_l}\right)^{\mathrm{T}}\big]\end{aligned} \tag{9}$$

where

$$\alpha_l = \frac{N_l}{N_l + r_{\boldsymbol{\mu}}} \qquad \beta_l = \frac{N_l}{N_l + r_{\mathbf{W}}}\ ; \tag{10}$$

L is the number of languages, N_l is the number of samples of language l; $\boldsymbol{\mu}_{0_l}$ and $\mathbf{S}_0$ are the prior means and covariance; $\boldsymbol{\mu}_{\mathrm{ML}_l}$ and $\mathbf{S}_{\mathrm{ML}_l}$ are the maximum likelihood means and covariances for language l computed on the in-domain data; and $r_{\boldsymbol{\mu}}$ and $r_{\mathbf{W}}$ are the relevance factors.

3.4. Calibration

Finally, we applied a linear calibration function to convert the Gaussian back-end scores into well-calibrated log-likelihoods. The calibration function had a language dependent bias and a common scaling parameter, and was trained using multi-class logistic regression.

4. Experimental setup

4.1. NIST LRE17 dataset

We evaluated our approach on the NIST language recognition evaluation 2017 (LRE17) task [2]. The LRE17 task consists of closed set language identification between 14 languages from 5 language clusters (Arabic, English, Slavic, Iberian and Chinese).

We focused on the fixed condition where the organizers constrained the datasets allowed for system development. NIST provided a training set (TRN17) consisting of narrow-band telephony speech built from previous NIST evaluations (around 2000h). Switchboard and Fisher English telephony corpora were also allowed for training. Additionally, NIST provided a development set (DEV17) containing around 60 hours of speech from a domain similar to the evaluation set. Both, development and evaluation sets contain audios from two sources: narrow-band telephony and broadcast radio (MLS14); and wide-band video (VAST). MLS14 audio files consisted of segments of 3, 10 and 30 seconds while VAST audio files contained the full duration of the original source video file.

Language recognition systems were requested to provide a vector of calibrated log-likelihoods, one for each target language. Performance was measured using a detection cost function which is a weighted average of miss and false alarm rates.

$$C(\gamma) = \frac{1}{L}\sum_{i=0}^{L}\left[P_{\mathrm{Miss}}(i,\gamma) + \frac{\gamma}{L-1}\sum_{j\neq i} P_{\mathrm{FA}}(i,j,\gamma)\right] \tag{11}$$

where $\gamma = (1-P_{\mathcal{T}})/P_{\mathcal{T}}$, $P_{\mathcal{T}}$ is the target language prior, and L the number of languages. $P_{\mathrm{Miss}}(i,\gamma)$ is the miss rate for language i and $P_{\mathrm{FA}}(i,j,\gamma)$ is the probability of detecting language i in an audio containing language j. Miss and false alarms are computed by applying detection thresholds $\log(\gamma)$ to the language log-likelihood ratios (derived from the calibrated log-likelihoods). The primary metric averages (11) for two operating points, $P_{\mathcal{T}} = 0.5$ and $P_{\mathcal{T}} = 0.1$. Also, the counts of each corpus (MLS14 and VAST) are equalized when computing the cost function so both have the same weight in the metric.

4.2. Experiments

The baseline is the language recognition system described in Section 3. We considered systems with Gaussian back-end non-adapted to the LRE17 development set; adapted to the full development set (condition independent); and adapted to the specific domain (condition dependent), i.e., different adapted

Table 1: *Result for the speech quality (PESQ), speech intelligibility (STOI, eSTOI), and audio source separation (SDR) for the simulated validation set. The values should be compared relative to the reference values. Higher is better for all speech enhancement measures.*

System	PESQ	STOI	eSTOI	SDR
All SNRs:				
Reference	2.456	0.733	0.565	4.395
OM-LSA	2.379	0.708	0.546	6.502
BLSTM	**2.815**	**0.793**	**0.634**	**12.333**
15 dB SNR:				
Reference	3.042	0.875	0.761	13.507
OM-LSA	2.895	0.844	0.730	13.249
BLSTM	**3.305**	**0.895**	**0.801**	**18.670**
-3 dB SNR:				
Reference	1.895	0.568	0.362	−4.626
OM-LSA	1.809	0.541	0.346	−1.722
BLSTM	**2.291**	**0.665**	**0.440**	**5.517**

model for MLS14 and VAST. We also considered condition independent and dependent score calibration. We processed the development and evaluation data with the OM-LSA and BLSTM SE methods. Thus, speech enhancement was included in the back-end adaptation and calibration steps.

5. Results

5.1. Speech quality measures

Table 1 shows the SE performance with four performance measures (PESQ, STOI, eSTOI, and SDR), as introduced in Section 2.1. The performance of OM-LSA was slightly degraded on the PESQ, STOI, eSTOI scores, but improved on the SDR score. This is because OM-LSA tends to remove noise components overly, which would affect the speech quality and intelligibility, especially for the high SNR setting. On the other hand, the BLSTM SE system outperformed OM-LSA for all measures consistently in both high and low SNR settings.

5.2. Language recognition

Table 2 presents language recognition in terms of the detection cost as defined in Section 4.1. The OM-LSA method improved the VAST (noisy video) performance from the baseline in most of the cases, but significantly degraded the MLS14 (telephone) in all cases. Meanwhile, the proposed BLSTM improved the performance in all the adaptation conditions, outperforming OM-LSA. For the MLS14 case, the BLSTM performance was degraded in some cases, but not significantly. For the VAST case, the improvement was very significant in all conditions. The best language recognizer, including condition dependent back-end and calibration, achieved 11.3% relative improvement when using our BLSTM SE. In average of the MLS14 and VAST cases, the relative improvement of BLSTM SE was around 6.3%, which is still significant.

Another thing worth mentioning is that, with applying SE, the gap between condition-dependent and condition-independent back-end systems was reduced. This property is useful in a real application, since we can avoid using a complicated condition-dependent system, which requires multiple domain-dependent models with a precise domain detector.

Table 2: *Results for the addition of a preprocessing speech enhancement step, for different language recognition systems. We consider systems with three types of back-end non-adapted to the development data, condition independent adapted (CI) and condition dependent adapted (CD); and two calibrations, condition independent and dependent. The values are from equation (11), where lower is better and the MLS14 and VAST display the result for the telephone and video audio respectively. The Baseline systems are without the preprocessing speech enhancement step.*

System	Baseline	OM-LSA	BLSTM
Cost average:			
GBE Non-adapt + Cal-CI	0.306	0.289	**0.269**
GBE Non-adapt + Cal-CD	0.292	0.277	**0.265**
GBE Adapt-CI + Cal-CI	0.234	0.238	**0.207**
GBE Adapt-CI + Cal-CD	0.221	0.227	**0.199**
GBE Adapt-CD + Cal-CI	0.219	0.235	**0.209**
GBE Adapt-CD + Cal-CD	0.206	0.218	**0.193**
MLS14:			
GBE Non-adapt + Cal-CI	0.198	0.218	**0.193**
GBE Non-adapt + Cal-CD	0.193	0.213	**0.192**
GBE Adapt-CI + Cal-CI	**0.165**	0.185	**0.165**
GBE Adapt-CI + Cal-CD	**0.162**	0.183	0.164
GBE Adapt-CD + Cal-CI	**0.168**	0.188	0.169
GBE Adapt-CD + Cal-CD	**0.164**	0.185	0.166
VAST:			
GBE Non-adapt + Cal-CI	0.414	0.360	**0.346**
GBE Non-adapt + Cal-CD	0.391	0.340	**0.337**
GBE Adapt-CI + Cal-CI	0.304	0.291	**0.249**
GBE Adapt-CI + Cal-CD	0.280	0.270	**0.235**
GBE Adapt-CD + Cal-CI	0.270	0.282	**0.249**
GBE Adapt-CD + Cal-CD	0.248	0.252	**0.220**

6. Conclusions

We proposed a BLSTM speech enhancement technique to improve language recognition in a noisy signal condition. The BLSTM is trained to estimate a time-frequency mask indicating the quality of each frequency bin. Using this mask, we obtain an enhanced version of the signal spectrogram, and recover the time domain waveform. We evaluated the quality of the enhanced signals in the recent NIST 2017 language recognition evaluation, where there is a condition with noisy audio from Internet videos. We compared results using the proposed method and baseline OM-LSA; also adapting the language recognition system to the target domain and non-adapting. In the noisy condition, we obtained performance gains around 16% for the case without adaptation and around 11% for the case where we performed condition dependent adaptation of the recognizer. Performance in clean conditions was not degraded. Also, speech enhancement contributed to reduce the gap between condition dependent and independent recognizers, which could greatly simplify the systems.

As future work, we plan to use more realistic noise databases like CHiME-4 [21], and Musan [22]. Additionally, reverberation could be simulated as well to reduce the noise mismatch further. Also, we want to perform speech enhancement in wide-band speech, instead of downsampling to 8 kHz, which should improve the language recognition performance on videos.

7. References

[1] M. H. Bahari, N. Dehak, H. Van hamme, L. Burget, A. M. Ali, and J. R. Glass, "Non-Negative Factor Analysis of Gaussian Mixture Model Weight Adaptation for Language and Dialect Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 7, pp. 1117–1129, jul 2014.

[2] "NIST 2017 Language Recognition Evaluation Plan," NIST, Tech. Rep., 2017.

[3] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 91–99.

[4] D. Michelsanti and Z.-H. Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," *Proc. Interspeech*, 2017.

[5] M. Kolbæk, Z.-H. Tan, and J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 153–167, 2017.

[6] D. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech intelligibility in background noise with ideal binary time-frequency masking," *The Journal of the Acoustical Society of America*, vol. 125, no. 4, pp. 2336–2347, 2009.

[7] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 1849–1858, 2014.

[8] H. Erdogan, J. R. Hershey, S. Watanabe, and J. L. Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 708–712.

[9] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder." in *Interspeech*, 2013, pp. 436–440.

[10] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, Jan 2015.

[11] F. Richardson, P. A. Torres-carrasquillo, J. Borgstrom, D. Sturim, Y. Gwon, J. Villalba, N. Chen, J. Trmal, and N. Dehak, "The MIT Lincoln Laboratory / JHU / EPITA-LSE LRE17 System," in *submitted to Odyssey 2018*, Les Sables d'Olonne, France, jun 2018.

[12] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403 – 2418, 2001.

[13] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, Sept 2003.

[14] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (ICASSP)*, vol. 2, 2001, pp. 749–752.

[15] ITU-T. (2005) Pesq, p.862.2. [Online]. Available: https://www.itu.int/rec/T-REC-P.862-200511-I!Amd2/en

[16] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, Sept 2011.

[17] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, Nov 2016.

[18] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[19] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, dec 2011, pp. 1–4.

[20] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis For Speaker Verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788 – 798, may 2011.

[21] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, vol. 46, pp. 535–557, 2017.

[22] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.