Single channel speech enhancement using convolutional neural network

Tomas Kounovsky and Jiri Malek
Institute of Information Technology and Electronics,
Faculty of Mechatronics, Technical University of Liberec,
Liberec, Czech Republic
Email: {tomas.kounovsky, jiri.malek}@tul.cz

***Abstract*—Neural networks can be used to identify and remove noise from noisy speech spectrum (denoisisng autoencoders, DAEs). The DAEs are typically implemented using the fully-connected feed-forward topology. Usually one of the following possibilities is used as DA target: 1) Ideal frequency ratio mask, which is applied to noisy spectrum to estimate the clean speech spectrum (masking) or 2) Clean speech spectrum directly (mapping). Recent research in the area of automatic speech recognition shows that convolutional neural networks are very promising in speech modeling. In this paper we thus suggest, construct the DAs with the convolutional topology. We investigate the suitability of both above described target types and compare the results with the fully connected DAs. Our experiments indicate that mapping-based convolutional networks estimating log-power spectra achieve significant improvement over all competing topologies and target types. Performance gains of 8 % in PESQ scores over the ideal ratio masking are observed. We also investigate the ability of DAEs to enhance speech of unseen language based on the language diversity of the training set. Our experiments suggest that training DAEs with language-diverse training sets does not yield any significant benefit for the task of speech enhancement.**

***Index Terms*—Denoising autoencoder, Single channel speech enhancement, Convolutional neural network, Multilingual training**

I. Introduction

The goal of speech enhancement is to improve the quality and intelligibility of noisy speech recordings. This problem has been widely studied for decades. Algorithms such as spectral subtraction [1] or minimum mean-square error short-time spectral amplitude (MMSE STSA) estimator [2] were proven to be successful in attenuating noise if the noise is approximately stationary and SNR levels are not too low. Several modern algoritms such as optimally modified log-spectral amplitude (OM-LSA) [3] or non-negative matrix factorization (NMF) [4] achieve better results at the cost of higher computational complexity. Most of the aforementioned algorithms require some apriori conditions to be met, such as periods without speech activity, where noise spectrum can be estimated.

Neural networks built as denoising autoencoders (DAEs), on the other hand, do not require any such apriori conditions to be met when applying the enhancement. The network learns the information about background noise during the training phase. It has been shown that DAEs estimating log-power (LP) spectra of clean speech obtain superior results to conventional methods, provided that a large and sufficiently diverse training datasets are used to train the DAE [5][6]. Modern DAEs usually use a deep neural network (DNN) with multiple layers of fully connected neurons. DAEs used in speech enhancement and automatic speech recognition (ASR) usually estimate either 1) masks that give the desired clean speech spectra after multiplying the noisy spectra, or 2) clean speech spectra directly. In [7], a comparison between mask estimating DNNs (masking) and direct feature estimating DNNs (mapping) was made. DNNs estimating FFT and gammatone-power spectra were shown to achieve inferior results to ideal binary and ideal ratio mask estimating DNNs in terms of perceptual scores.

Recent studies point towards the use of a convolutional neural network (CNN) as a convolutional denoising autoencoder (CDAE). This type of neural network architecture is primarily used in the field of image classification and feature detection, where it surpassed all other approaches [8]. The convolutional models reflect strong correlations of speech in time and are invariant to translational variance within speech caused, e.g., by different speaking styles [9]. In [10], a type of CDAE was used to effectively separate speech and music filterbank features, improving ASR scores. In [11], CDAE was used to estimate ideal ratio masks (IRMs) for gammatone coefficients. Results show that CNNs can perform on par with or slightly better than the fully-connected DNN architecture in speech enhancement when measured with objective criteria. In [12], a SNR-aware mapping-based CDAE was created. This system utilizes several CDAEs trained in different SNR levels. First, the system estimates the SNR level in LP noisy spectrum. Then it selects a specific CDAE, which corresponds to the given noise condition and uses it to enhance the speech. The authors claim that this approach achieves better results than single general CDAE trained for all SNR levels.

To our best knowledge, nobody performed an investigation of most suitable targets for CDAE, such as was done for DAE in [7]. In this paper, we consider: 1) mapping-based targets, i.e., we estimate directly the desired log-power spectra of the clean speech and 2) masking-based targets, i.e., we estimate the ideal ratio masks, which are subsequently used to produce clean speech spectrum. As a baseline speech-enhancement approach for comparison, we utilize the fully-connected DAE.

Further, authors in [10] suggest that CDAE used for feature enhancement in ASR context are largely language independent. This means that networks trained on language A can be used to enhance utterances in language B. Such scenario constitutes mismatched training-test conditions, which usually result into distortions in reconstructed speech. However, the performance of ASR usually does not deteriorate due to this phenomenon, because the acoustic models can be trained very robust with respect to speech distortions. This was shown, e.g., for multichannel beamforming in [13]. In this paper, we investigate whether the claim of independence is valid also in the speech enhancement context, where the goal is to avoid distortion of estimated speech and increase its intelligibility.

Specifically, experiments in [10] showed that DAEs for ASR trained on a dataset containing only one language (monolingual networks) can be used to enhance utterances in unseen language, albeit with lower performance. It was also shown that DAEs can

This paper was partly supported by the Student Grant Scheme 2017 project of the Technical University in Liberec.

be trained using datasets containing multiple languages (multilingual networks) to achieve better performance over all training languages, aleviating the problems of monolingual networks. The multilingual network was, however, not tested using languages not included in the training set. Doing so could provide insight into the general language dependence of the CDAE, such as if increasing the language diversity of the training set also increases the performance in unseen language conditions. In this paper, we train a monolingual and multilingual network and compare their performances in scenarios with seen and unseen test languages.

II. Problem formulation

A. Speech enhancement

Speech enhancement techniques often assume the following model:

$$x[n] = s[n] + v[n], \tag{1}$$

where s is the clean speech signal, v is the noise signal and x is the distorted speech mixture. The same can be transformed by the short-time Fourier transform into the time-frequency domain, giving

$$X[k,m] = S[k,m] + V[k,m], \tag{2}$$

where X, S and V denote, respectively, mixture, speech and noise STFT spectra. The variable k denotes the frequency bin and m the index of time frame. The goal of speech enhancement is to determine an estimate of the speech spectrum $\hat{S}$. To this end, we consider two following approaches in this paper: 1) estimate $\hat{S}$ directly, or 2) estimate a mask

$$W = \frac{S[k,m]^2}{S[k,m]^2 + V[k,m]^2}, \tag{3}$$

so that

$$\hat{S}[k,m] = \sqrt{X[k,m]^2 \cdot W[k,m]} \tag{4}$$

Usually, the processing is limited to the magnitude part of the spectrum, because phase is of less importance for speech intelligibility [14]. The estimated clean speech is reconstructed using the noisy speech phase information:

$$\hat{S}_f[k,m] = \hat{S}[k,m] \cdot e^{i \cdot \angle X[k,m]} \tag{5}$$

B. Fully-connected denoising autoencoders

Our DNN system is very similar to the one used in [5] and can be seen in Fig. 1. For the mapping-based network, noisy and clean training recordings are first transformed into time-spectral domain by STFT. Log-power features are then computed as

$$X_{LP}[k,m] = ln(X[k,m]^2), \tag{6}$$

similarly for the clean spectra S. The input vector consists of 11 consecutive noisy frames, 5 preceding and 5 following the current frame. The target vector is the current clean frame. All training vectors are normalized to zero mean and unit variance over each frequency bin. For the masking-based network, this procedure is the same except that targets are formed by (3) from the magnitude spectra without any normalization.

These pairs serve as inputs and targets to the network during training. The network itself consists of 4 hidden layers with 1024 neurons within each layer. The layers are fully-connected, meaning that all neurons from the preceding layer contribute to the activation of each neuron in the following layer. Activation functions used in hidden layers are all ReLU and the output layer is linear.

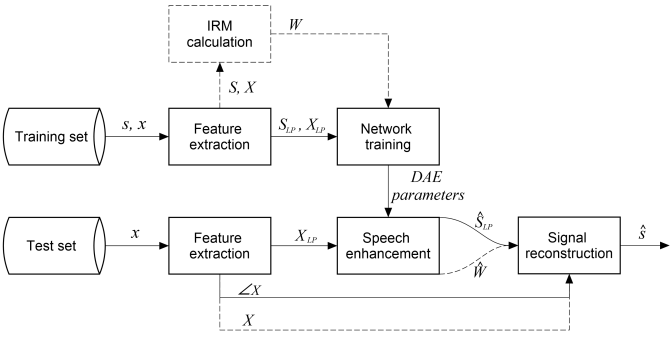

Fig. 1. Denoising system used throughout the experiments. Dashed lines denote the components used in the masking-based system only.

The network implemented in this paper was trained for 50 epochs (passes over the entire dataset) with the stochastic gradient descent algorithm. Learning rate was set to 0.015 initially with a 2 % decay starting from epoch 40. Batch size was 512. Error criterion was set to mean-square error. All training vectors were randomly permutated before each epoch.

During testing, noisy recordings are prepared in the same manner as the training recordings. The network computes its estimates of the targets. For the masking-based network, the estimated mask is used to compute the estimated clean speech spectra by (4). Finally, the estimated magnitude spectra are combined with the original noisy phase by (5) and transformed back into the time domain by inverse STFT with overlap-add.

III. Convolutional denoising autoencoders

The idea behind convolutional neural networks for image classification is that each small region of the input image is convoluted by a multitude of filters (called kernels), which form the first convolutional layer. Each kernel represents a certain feature (e.g. line, curve). This allows the network to create feature maps which describe the positions of the aforementioned simple features in the image. Stacking more convolutional layers on top enables the network to recognize more complex features (e.g. faces, wings of an airplane). When used for speech enhancement, the network learns to recognize important localized time-frequency speech features, such as formants. Furthermore, the CNN is invariant to translational variance due to the fact that each part of the feature map is processed by the same kernels. This enables the network to efficiently learn and process speech from e.g. male and female speakers, which usually have different voice pitch and therefore different fundamental frequencies.

Our CDAE-based system is mostly the same as the DAE system, with the network being the only difference. Both our mapping and masking-based CDAEs use the same structure. Both networks have 5 layers, 2 convolutional with a max-pooling layer in-between and 2 fully-connected layers on the top. The input feature vectors are structured as 11 feature maps of size 129x1. The first convolutional layer uses 52 feature maps and the second convolutional layer uses 78 feature maps. The kernel for both convolution layers has a size 5x1. We use max-pooling by factor 3. Both fully-connected layers have 1024 neurons. All activation functions are ReLU. The output layer has 129 neurons and uses a linear activation function. The resulting network has a similar amount of parameters as the DNN-based DAE (~4.6 million).

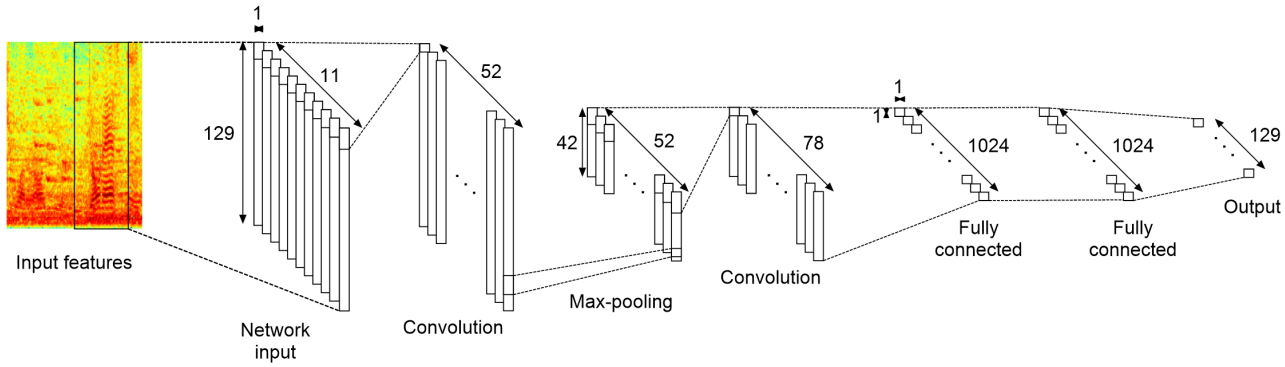

Fig. 2. Schema of the convolutional denoising autoencoder

IV. Experimental setup

In order to investigate the language dependence of the investigated DAEs, we choose to utilize 4 different languages in our experiments - English, Chinese, Polish and Russian. The TIMIT database [15] serves as the source of clean English speech, with 6000 utterances (96 % of total) kept for training purposes and the remaining 256 utterances for creating the first test set. For Chinese language we have used a portion of the THCHS-30 database [16]. The first 1250 utterances were kept for training and 85 for the second test set. The Globalphone dataset [17] was used as the source of Polish and Russian languages. The first 1200 Polish utterances were kept for training. The Russian language was not used in training and was left for test purposes as a language unseen by the autoencoders. The first 94 utterances were chosen as the basis for the third test set. The Russian and Chinese utterances are generally longer than the English utterances, but the total time duration of all test sets was the same.

The 3rd CHiME challenge database of noise [18] was used to corrupt the clean speech recordings. Namely Cafeteria, Pedestrians and Bus noise types were chosen for training and Street noise type was left for testing. All training data were corrupted with each training noise type at 3 different SNR levels (0 dB, 5 dB and 10 dB). To test the performance of our networks in mismatched conditions, both testing datasets were mixed similarly as the training dataset, with 1 unseen noise type (the Street) and 3 additional unseen SNR levels (-3 dB, 2 dB and 7 dB). About 8.33 % of all training utterances were randomly left undistorted by noise in order to reduce the distortion of clean speech during enhancement.

A. Monolingual training dataset

As the name suggests, only the English language was used in creation of this training dataset. All 6000 clean utterances of the TIMIT dataset were duplicated to form 12000 utterances, which were equally split into thirds and each part was distorted by one of the training noise types. The duration of the resulting dataset was almost 26 hours.

B. Multilingual training dataset

The multilingual dataset was carefully created to match the length of the monolingual dataset. It consists of three equally-sized language parts; the English, the Polish and the Chinese. The first 4000 TIMIT utterances were selected to form the first part. A number of utterances from the Chinese and Polish datasets were picked, 1250 and 1200 respectively. All utterances were mixed with noise such that each noise type was present in each language part equally.

All signals were downsampled to 8 kHz sampling frequency before mixing. Short-time Fourier transform of length 256 samples (32 ms) and overlap of 128 samples (16 ms) was used to create the log-power spectra and ideal ratio masks. These were used as inputs and targets of the networks throughout the experiments. All four networks were trained as described in the previous section.

Two main criteria were used for objective performance evaluation, namely, perceptual evaluation of speech quality (PESQ) [19] and log-spectral distance (LSD), as defined in [20]. PESQ has a high correlation with subjective listening tests and can provide relative comparison of intelligibility. LSD measures logarithmic distance in spectra and provides information about signal similarity. It can be assumed that estimations with low LSD have two positive properties: 1) exhibit high SNR and 2) do not contain significant distortions of speech.

V. Experiments and results

A. Monolingual training: comparison of topologies and target types

In this experiment we compare the performance of masking and mapping-based convolutional and fully-connected DAEs (FDAEs). All networks were trained on the monolingual dataset. Tests were performed on the English test set and the results were averaged over all noise types. We skip the detailed results with respect to specific noise types due to clarity of presentation. The performance of the DAE does not differ significantly for the different background noises. This holds even for the Street noise, which was not included in the training set.

As can be seen in Table I, all networks noticeably improved the quality and intelligibility of noisy speech. The PESQ improvements from 13 to 25 % were observed. The LSD improvements ranged from 24 to 39 %. The FDAEs achieved slightly lower performance than CDAEs in terms of both PESQ and LSD, which is in agreement with [11].

Comparing the target types in the terms of PESQ, we find the mapping superior to masking, especially for lower SNR levels (below 10 dB). Considering the CDAE topology, mapping achieves higher PESQ by more than 8 % compared to masking. This result partly contradicts the findings in [7]. There the authors claim that masking achieves better results for magnitude spectrum inputs and both approaches are comparable for gammatone frequency power spectra. However, the paper confirms better performance of mapping in scenarios with low SNR. We thus conclude that the choice between mapping and masking targets depends strongly on the specific input.

TABLE I
AVERAGE PESQ AND LSD RESULTS FOR MASKING/MAPPING-BASED DNN AND CNN DAE

			DNN				CNN			
	Noisy		Masking		Mapping		Masking		Mapping	
SNR	PESQ	LSD	PESQ	LSD	PESQ	LSD	PESQ	LSD	PESQ	LSD
-3	1,88	16,90	2,15	14,13	**2,22**	**11,09**	2,19	13,61	**2,36**	**10,42**
0	2,05	16,01	2,34	12,83	**2,43**	**10,13**	2,39	12,26	**2,57**	**9,58**
2	2,17	15,29	2,47	11,86	**2,57**	**9,54**	2,53	11,32	**2,71**	**9,07**
5	2,36	13,94	2,71	10,05	**2,75**	**8,83**	2,79	9,53	**2,89**	**8,43**
7	2,49	13,00	2,84	9,29	**2,87**	**8,39**	2,92	8,85	**3,00**	**8,03**
10	2,68	11,54	**3,03**	8,26	3,02	**7,85**	3,11	7,94	**3,15**	**7,50**

Further, the log-power spectrum seems to be an optimal input for mapping-based networks, regardless of the autoencoder topology.

It was also observed that performance in unseen SNR levels did not noticeably deteriorate, showing that all networks are able to generalize well in this aspect. This is mainly caused by the fact that all networks learn the task frame-by-frame. Segmental SNR in one frame varies significantly, which forces the network to account for many SNR levels during training.

B. Performance of mono- and multilingual models

In the context of feature enhancement for ASR, the paper [10] suggests that CDAEs are largely language independent. This experiment investigates whether the claim of independence is valid also in the speech enhancement context, where the goal is to avoid distortion of estimated speech and increase its intelligibility. We also test whether increasing the training set language diversity (i.e., inclusion of several languages in single training set) increases the ability of the network to enhance unseen languages.

First, both mono- and multilingual mapping-based CDAEs are tested in seen language conditions, meaning the English test set. Then, both networks are tested on the Chinese test set, which constitutes seen language conditions for the multilingual network, but unseen language conditions for the monolingual network. These two experiments represent the speech enhancement version of the experiments done in [10] with the difference of one additional language being present in our multilingual training set. Finally, both networks are tested in unseen language conditions using the Russian test set.

Results in Table II show that in seen language conditions (the English language), the monolingual network consistently achieves slightly better performance (PESQ gains of 3 %) compared to the multilingual-trained networks. This can be expected, because the multilingual training set is more diverse with respect to both language and acoustics. This forces the same-sized network to accommodate a more complex task.

Results in Table III show the performance when using the Chinese language, i.e. in seen language conditions for the multilingual network and unseen language conditions for the monolingual network. While the multilingual network did consistently surpass the monolingual network, the PESQ score increase was only 2.4 % on average. This does not completely reflect the results in [10], where switching from monolingual network in unseen language conditions to multilingual network in seen language conditions improved the word error rate during ASR by a minimum of 53 %. The reason for this difference is not entirely clear. It is possible that the monolingual network produces artifacts, which are not reflected by the considered evaluation criteria (energy- and perception-based), but are detrimental for the automatic recognizer and result in higher word error rate. This opens a topic for further research beyond the scope of the current paper.

TABLE II
AVERAGE PESQ AND LSD RESULTS OF MONOLINGUAL AND MULTILINGUAL CDAE IN SEEN LANGUAGE CONDITIONS (ENGLISH TEST SET)

	Noisy		MonoCDAE		MultiCDAE	
SNR	PESQ	LSD	PESQ	LSD	PESQ	LSD
-3	1,88	16,94	**2,36**	**10,42**	2,29	10,91
0	2,06	16,00	**2,57**	**9,58**	2,49	10,02
5	2,36	13,95	**2,89**	**8,43**	2,82	8,77
10	2,68	11,55	**3,15**	**7,50**	3,11	7,78

TABLE III
AVERAGE PESQ AND LSD RESULTS OF MONOLINGUAL CDAE IN UNSEEN LANGUAGE CONDITIONS AND MULTILINGUAL CDAE IN SEEN LANGUAGE CONDITIONS (CHINESE TEST SET)

	Noisy		MonoCDAE		MultiCDAE	
SNR	PESQ	LSD	PESQ	LSD	PESQ	LSD
-3	1,86	17,43	2,22	12,64	**2,24**	**12,41**
0	2,03	16,44	2,41	11,59	**2,45**	**11,38**
5	2,32	14,37	2,70	10,27	**2,77**	**10,13**
10	2,65	11,90	2,96	9,28	**3,05**	**9,25**

TABLE IV
AVERAGE PESQ AND LSD RESULTS OF MONOLINGUAL AND MULTILINGUAL CDAE IN UNSEEN LANGUAGE CONDITIONS (RUSSIAN TEST SET)

	Noisy		MonoCDAE		MultiCDAE	
SNR	PESQ	LSD	PESQ	LSD	PESQ	LSD
-3	2,02	15,15	**2,42**	**10,15**	2,39	10,26
0	2,18	14,38	**2,59**	**9,52**	2,59	9,62
5	2,45	12,35	2,88	8,73	**2,90**	**8,72**
10	2,75	9,94	3,13	8,07	**3,17**	**7,95**

Results in Table IV show that both types of networks performed similarly in unseen language conditions. Monolingual network achieves slightly better results in lower SNR levels, while multilingual network achieves slightly better results in higher SNR levels. Overall the difference in performance is negligible. These results suggest that multilingual training does not yield the benefits we expected, that is a better language generalization with a more diverse training set.

Results of this experiment suggest that training CDAEs with multilingual training sets for the purpose of enhancing unseen languages does not yield any significant benefits. Multilingual training does provide a very small performance gain if the network is expected to be enhancing a set of a few specific languages. Monolingual datasets are, however, far easier to create than balanced multilingual datasets.

VI. Conclusion

In this paper we have experimented with convolutional neural networks in the role of a denoising autoencoder for speech enhancement. We have compared the performance of both fully-connected and convolutional networks as well as the performance of networks estimating ideal ratio masks (masking-based) and networks estimating speech spectra directly (mapping-based). For the former, convolutional networks provided superior performance when measured with objective criteria (PESQ and LSD), which is in agreement with previous research. For the latter, mapping-based networks estimating log-power spectra of both architectures consistently outperformed their masking-based counterparts, especially in lower SNR levels. Autoencoder performance seems to be strongly dependent on the choice of input and target feature type. These results add to the previous research and suggest that log-power spectra are ideal features for mapping-based networks.

We also compared monolingual (trained using one language) and multilingual (trained using multiple languages) convolutional denoising autoencoders in 3 situations: 1) both in seen language conditions 2) monolingual in unseen and multilingual in seen language conditions, and 3) both in unseen language conditions. In situation 1, performance of the monolingual network surpassed the performance of the multilingual network by about 3 % in PESQ score. In situation 2, multilingual network surpassed the monolingual network by also about 3 % in PESQ score. This does not reflect the results of previous research, where including additional languages in the training set significantly increased performance in automatic speech recognition. In situation 3, both networks performed similarly. These results suggest that training convolutional denoising autoencoders with multiple languages does not yield significant benefits over the simplicity of creating a monolingual training set. In the future, we would like to explore the behaviour of convolutional denoising autoencoders in more detail, particularly the prominent language invariance shown in this paper.

References

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

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

[4] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 4029–4032.

[5] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks." *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, 2014.

[6] 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 (TASLP)*, vol. 23, no. 1, pp. 7–19, 2015.

[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] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[9] T. N. Sainath, B. Kingsbury, G. Saon, H. Soltau, A.-r. Mohamed, G. Dahl, and B. Ramabhadran, "Deep convolutional neural networks for large-scale speech tasks," *Neural Networks*, vol. 64, pp. 39–48, 2015.

[10] M. Zhao, D. Wang, Z. Zhang, and X. Zhang, "Music removal by convolutional denoising autoencoder in speech recognition," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2015 Asia-Pacific*. IEEE, 2015, pp. 338–341.

[11] L. Hui, M. Cai, C. Guo, L. He, W.-Q. Zhang, and J. Liu, "Convolutional maxout neural networks for speech separation," in *Signal Processing and Information Technology (ISSPIT), 2015 IEEE International Symposium on*. IEEE, 2015, pp. 24–27.

[12] S.-W. Fu, Y. Tsao, and X. Lu, "Snr-aware convolutional neural network modeling for speech enhancement," *Proceedings of the Interspeech, San Francisco, CA, USA*, pp. 8–12, 2016.

[13] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 196–200.

[14] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 4, pp. 679–681, 1982.

[15] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[16] D. Wang and X. Zhang, "Thchs-30: A free chinese speech corpus," *arXiv preprint arXiv:1512.01882*, 2015.

[17] T. Schultz, "Globalphone: a multilingual speech and text database developed at karlsruhe university." in *INTERSPEECH*, 2002.

[18] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'chime'speech separation and recognition challenge: Dataset, task and baselines," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 2015, pp. 504–511.

[19] 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 *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 2. IEEE, 2001, pp. 749–752.

[20] J. Du and Q. Huo, "A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions." in *INTERSPEECH*, 2008, pp. 569–572.