SPEECH ENHANCEMENT FOR FORENSIC PURPOSES

Gheorghe POP[1], Dragoș BURILEANU[2]

The research community interest in forensic audio processing has increased exponentially in the latest years, mostly as an effect of media coverage of cases where the debates in the courts of law have benefitted from forensic speech recording analyses. The enhancement of speech signals quality for forensic purposes targets the recordings that are difficult to understand and use in legal contexts, whereas speech telecommunications or artistic performance have different performance criteria.

In this work, the forensic speech enhancement domain is discussed, together with an enhancement method we propose for forensic applications.

Keywords: speech enhancement, forensic audio, deep neural networks

1. Introduction

Since the beginnings of audio recordings, there was a need for a practice code in making the recorded contents sound best. This need is served since February 17, 1948, when the Audio Engineering Society was created in New York. The technical achievements of that moment, which are now more than 70 years old, were a new electronic phonograph pick-up, with less than 15 grams of pressure, and the dual-cone loudspeakers, both presented there by Harry F. Olson, from RCA Laboratories [1]. All those devices are today remarkable museum items, after several technical revolutions in speech processing.

Pioneering efforts in speech enhancement were made all along, after the World War II [2]. Since digital computers first became available in the 1970s [3], *digital signal processing* (DSP) techniques were introduced, with many important breakthroughs, such as the adaptive noise cancelling [4], the spectral subtraction method [5], decision-based noise filtering [6], and minimum mean-square noise estimators [7]. We are now in the middle of a deep learning era, built on ideas introduced in the 1980s, which has recently started a revolution in science and technology, which is expected to power up achievements of human kind at a similar scale as the widespread of electricity networks has since the 1880s [8].

[1] PhD student, Speech and Dialogue Laboratory, Faculty of Electronics, Telecommunications and Information Technology, University POLITEHNICA of Bucharest, Romania, e-mail: gheorghe.pop@etti.pub.ro

[2] Professor, Speech and Dialogue Laboratory, Faculty of Electronics, Telecommunications and Information Technology, University POLITEHNICA of Bucharest, Romania, e-mail: dragos.burileanu@upb.ro

The aim of the paper is to contribute to the forensic speech enhancement research, by shortly introducing the domain and by proposing a new method.

The contents of the paper are organized as follows. A short description of the speech recording chain is presented in section 2. The problems of speech quality are reviewed in section 3. Several major speech enhancement techniques are discussed in section 4, both classic and emergent, while in section 5 we propose a new speech enhancement method, based on a *deep neural network* (DNN) that uses input data about speech quality, and compare its performance to others. Conclusions are accommodated in section 6.

2. From speech to audio recording

Speech may be seen as a succession of acoustic events and treated as all other signals. Acoustic events to be recorded have to go through three stages from their acoustic source to the storage file. In the first stage, the speech is produced as an analog, acoustic signal, which reflects the ability of the person to produce understandable speech in a given language. In a similar manner to all other sources, the signal travels to each microphone through a unique acoustic channel. The channel properties depend on the positions of both the respective source and microphone, as well as on properties of the medium, such as the density, temperature etc. By virtue of the superposition principle, the total acoustic pressure at each microphone is realized as a mixture of all arriving acoustic pressure waves, weighted by the microphone's directivity.

The second analog stage is of electric nature. A microphone transforms the analog acoustic wave into an electrical signal, by converting the acoustic pressure wave into an analog voltage, which is then adapted to the needs of the signal processing chain.

The third stage is not only electric, but also digital at the same time, dedicated to digital signal processing. The conversion to digital format reduces the risk of signal contamination by noise of electric origin.

3. Problems of speech quality

Quality of recorded speech may be considered under two main angles: the overall quality, and the intelligibility of speech, as perceived by human listeners. The relationship between perceived quality and speech intelligibility is not entirely understood, although some degree of correlation is obvious. Overall subjective quality may be expressed by various scores, derived from opinions of listeners. Given the connection between human listener opinions and the goal of speech enhancement, opinion scores offer some subjective quality reference.

Meanwhile, speech intelligibility is defined as the percentage of correctly understood language units, usually words, from the inputted amount.

Randomly chosen panels of listeners can produce very different results, so that panels of well-trained listeners are necessary, for reliable and repeatable results in subjective tests. The alternative speech quality assessment approach is to estimate objective quality measures. Some measures assess the degree of improvement produced by a given speech enhancement method, while others measure the intrinsic quality of a given signal, using speech perception models.

The enhancement of recorded speech consists in assessing the speech quality, and applying special processes on the signal, which would preferably tackle the problems without distorting the target speech. Most of the time, this is not a trivial task, because of the influence of reverberation and noise in the context of the medium and source-microphone geometry, as illustrated by the following possibilities:

- With only one microphone and one speaker, in fixed positions, the acoustic channel may be taken as invariant. In such a case, if the impulse response of the channel can be estimated, the reverberation can be partly reversed. If there is no reverberation, or it was reversed beforehand, the speech would be separated from noise based on their different behavior and spectral structure.
- Slowly moving speaker or microphone asks for adaptive estimation of the acoustic channel impulse response at every short time analysis point.
- When two or more microphones are involved, given that the acoustic mixtures are made from the same sources, there is a chance to, more or less, separate speakers from the mixtures.
- With an increasing number of microphones, the acoustic sources in the mixtures should be easier to separate in the absence of reverberation. With reverberation present, each acoustic channel has different reverberation properties so that each recording channel asks for independent dereverberation before the source signals are unmixed. Hopefully, speech signals would be separated from all noise signals.
- The behavior of the acoustic channel, often dependent on frequency, allows for the disentangling of component source signals to be performed on narrow frequency bands.

Two major classes of techniques are discussed in this paper, the “classic” ones, which work without using machine learning, and “emergent” techniques, which use neural networks and all other techniques based on deep learning.

4. Classic and emergent speech enhancement techniques

Considering, for start, that only one speaker was recorded, the oldest enhancement technique that came into play was the use of an analog signal equalizer device, which allowed the transfer function of the acoustic chain to be adjusted in order to promote the useful signal components, while reducing the others. Using this very intuitive technique, the *signal-to-noise ratio* (SNR) was improved, which has suggested the spectral shaping technique, and has later inspired a large class of techniques, namely the spectral subtraction techniques.

In a recent paper [9], one of the fastest enhancement techniques, an autocorrelation-based speech enhancement, is described. The spectrum of the stationary noise is shaped to follow the spectrum of the speech, thus the regions in the signal spectrum with low level speech components have to face a lower noise level, and so the SNR is increased.

Spectral subtraction-based methods rely on subtracting the estimated power spectrum of the noise from the power spectrum of the noisy speech signal, with no prior knowledge of the power spectral density of the clean speech and noise signals. Spectral subtraction can be used to suppress background noise under various assumptions, such as the one that the noise is stationary or changes slowly during the non-speech and speech activity time intervals [10].

The procedure of spectral subtraction includes a step of estimating the power spectrum of the noise on non-speech/silence intervals, followed by the step of subtracting it from the short time power spectrum of the signal. In doing so, the signal is analyzed on a short time basis, and the speech inactivity intervals are determined based on decisions made for each analysis window.

Then, Fourier transform is applied on the windowed frames of the noisy speech signal, while speech enhanced by spectral subtraction is [11]:

$$\left|\hat{S}(k)\right|^2 = \begin{cases} \left|X(k)\right|^2 - \delta\left|\hat{D}(k)\right|^2, \text{ for } \left|X(k)\right|^2 - \delta\left|\hat{D}(k)\right|^2 > \beta\left|\hat{D}(k)\right|^2 \\ \beta\left|\hat{D}(k)\right|^2 \qquad\qquad\qquad , \text{ otherwise,} \end{cases} \tag{1}$$

where $X(k)$, $\hat{S}(k)$, and $\hat{D}(k)$ are the magnitude power spectrum of window *k* of corrupted speech, estimated speech, and estimated noise respectively, δ is the over subtraction factor, which depends on the a posteriori segmental SNR, and β is the spectral factor with values between 0 and 1. A compromise value of β must be found, given that a high spectral floor makes the remaining noise audible, while a small value of β reduces noise a great deal, but the remnant noise becomes annoying. The enhanced speech signal $\hat{s}(t)$ is obtained by applying the inverse Fourier transform to the estimated spectrum of the speech with the phase data taken from the direct Fourier transform.

This class of methods is known to produce the so-called “musical noise”, which consists in some tones which randomly and rapidly change, noticeable in the background of the useful signal. Parametric spectral subtraction methods and gain function filtering were first tried, as means to reduce the musical noise, then more efficient methods were found.

Another class of speech enhancement techniques is the inverse filtering class. This means that if we know the impulse response of a filter the signal has previously been passed through, we can remove its effects. In favorable conditions, the unknown filter could be blindly estimated. For example, in [12] reverberation and noise parameters are blindly estimated. While the reverberation is removed using an inverse filtering, a Bayesian filtering is applied, controlled by voice activity in the signal, to spectrally subtract the noise from the noisy speech.

For slowly variable noise spectra, the Wiener adaptive filtering technique [2], also called "optimal filtering," is available, which works as a compromise between inverse filtering and spectral subtraction.

Using the time-domain structure of the speech, which presents repetitive peaks, speech enhancement techniques based on wavelet decomposition were implemented. The removal of noise components by thresholding the wavelet coefficients relies on the assumption that the energy of speech in a noisy speech signal is mainly concentrated in a small number of wavelet dimensions [13].

The wavelet thresholding is adequate for enhancement of forensic audio recordings corrupted with different types of colored noise, with different distributions in different frequency subbands, even at high levels of noise. The level-dependent threshold, λ, can be represented by [13]:

$$\lambda = \sigma_j \sqrt{2 \log N_j} \text{, where} \tag{2}$$

$$\sigma_j = \frac{\mathrm{MAD}\left(D_j\right)}{0.6745} \tag{3}$$

is just a coefficient depending on the level *j*, MAD is the *median absolute deviation* of the detail coefficients D_j, and N_j is the length of the noisy speech signal, for the same level. In selecting the wavelet coefficients to keep, both hard (T_{hard}) and soft (T_{soft}) thresholds can be used, defined by

$$T_{hard}\left(D_j\right) = \begin{cases} D_j, \text{ for } \left|D_j\right| > \lambda \\ 0, \text{ for } \left|D_j\right| \le \lambda \end{cases} \text{, and} \tag{4}$$

$$T_{soft}\left(D_j\right) = \begin{cases} sign\left(D_j\right) * \left(\left|D_j\right| - \lambda\right), \text{ for } \left|D_j\right| > \lambda \\ 0 \qquad\qquad \text{, for } \left|D_j\right| \le \lambda \text{ .} \end{cases} \tag{5}$$

All speech enhancement techniques discussed so far work without caring about the number of channels of the input signal, so that they qualify as single-channel techniques. Multi-channel techniques must consider the spatial diversity of the sources, which allows the suppression of a given source and improve the quality of the speech under noisy conditions. *Independent component analysis* (ICA) may be used in multi-channel speech enhancement to separate the speech from noise, if noisy speech is transformed into components which are statistically independent [14]. The independent components are estimated based on maximizing the non-Gaussian distribution of one independent component. The difference between a Gaussian distribution and the distribution of the independent component is measured using a higher order statistic, which is fixed for Gaussian distributions. Preferably, the contrast parameter is either the *kurtosis* or the *excess of kurtosis* [15].

How does the ICA work, in its instantaneous form? Let the source speech and noise signals, emitted from *N* sources, be

$$\mathbf{s}(t)=\{s_1(t),s_2(t),...,s_N(t)\}. \tag{6}$$

For forensic applications, the noisy speech signals can be recorded instantaneously, by using *M* microphones in a street, and be expressed as

$$\mathbf{x}(t)=\{x_1(t),x_2(t),...,x_M(t)\}. \tag{7}$$

Instantaneous ICA is defined in [15] as a linear transformation of noisy speech signals into components which are statistically independent, by

$$\mathbf{x}=\mathbf{A}\mathbf{s}, \tag{8}$$

where **A** is an unknown mixing matrix.

The goal of ICA is to estimate the original sources from the mixed signals. The estimates of source signals, $\hat{\mathbf{s}}$, can be represented by

$$\hat{\mathbf{s}}=\mathbf{W}\mathbf{x}, \tag{9}$$

where **W** is the unmixing matrix, which equals the inverse of the mixing matrix **A**, when the matrix is square.

When the disturbing signal is also speech, or speech-like, a different problem arises, namely the co-talker interference, which can be tackled by both single- and multiple-channel enhancement techniques. In [16], two single-channel speech enhancement techniques were used to suppress co-talker interference from forensic audio recordings – the *dynamic time warping* (DTW) and the *wavelet packet thresholding* (WPT). Spectral subtraction was used to remove colored noise from mixed speech signals and convolutive ICA was used to separate one speaker from another in [17] to improve the performance of speaker identification.

Several techniques use speech models to produce hard or soft decisions on each analysis window, whether the signal is mostly speech or mostly noise. The decisions were even taken to each time-frequency cell, that is to each identifiable time-frequency unit on the signal spectrogram.

In order to separate a signal of interest from a cocktail party-like mixture, its spectrogram can be multiplied with an *ideal binary mask* (IBM), which can be estimated by using either *Gaussian mixture models* (GMM), *support vector machines* (SVM), or trained *deep neural networks* (DNN). It is clear that the IBM is not a practical applicable speech enhancement method, since it requires the knowledge of both the clean and noise signal, in isolation. For practical speech enhancement, the goal is to obtain a decent estimate of the IBM.

The idea of the IBM arises from the model of the human auditory perception, proposed by Albert Stanley Bregmans, called *auditory scene analysis* (ASA) [18]. Bregman identifies two stages of auditory analysis. First stage, often called *the segmentation stage*, decomposes the input signal into *time-frequency* (T-F) units. The second stage groups the different T-F units that are likely to come from the same source. The T-F units within the same group are then collected into a perceptual stream. It is believed that this is how humans perform sound source segregation. This model has inspired research within the field of *computational auditory scene analysis* (CASA), where the goal is to extract such streams using computer programs [18]. In a CASA system, the input signal is first transformed into a *cochleagram*, that is, a gammatone filter bank T-F representation, then divided into frames. Since the goal is to segregate a speech signal from a noise signal, the IBM has been proposed to decide whether a T-F unit in the cochleagram is dominated by noise or by speech. The IBM is given by

$$\mathrm{IBM}(n,\omega)=\begin{cases}1, \text{ for } \dfrac{\left\|x(n,\omega)\right\|^2}{\left\|v(n,\omega)\right\|^2}>\theta\\ 0, \text{ otherwise,}\end{cases} \tag{10}$$

where $\left\|x(n,\omega)\right\|^2$ is the energy of speech T-F unit, $\left\|v(n,\omega)\right\|^2$ is the energy of noise in T-F unit (n,ω), and θ is a threshold value. Considerable improvements in speech intelligibility have been reported by use of an IBM with the decision rule in equation (10) [18].

While the GMMs and SVMs need a model of speech signal, DNN algorithms learn their own representation of the data, and remain capable of finding the best features and decisions without being told what to look for. Most of the DNN concepts are inspired from nature, including our beliefs on how humans perform complex tasks such as understanding speech, and recognizing persons or objects [19]. This is entirely different from the usual machine learning paradigm where features were primarily hand-crafted, and the actual machine learning was limited to how to perform purely discriminative classification tasks.

With a proper training, *fully connected deep neural networks* (FCDNNs) are able to recognize or synthesize speech, or to solve various tasks, including speech enhancement. One of the deep network structures that is both simple and stackable is the *denoising autoencoder* (DAE) [20]. Starting from a given set of data points, such as the set of samples in an analysis window, an autoencoder learns a different set of features, together with the reversed transformation, in order to reconstruct the initial data with a minimal error. A DAE is trained a little bit differently, by reconstructing the clean signal from a version of the signal corrupted with noise. In [20] it has been shown that minimizing the expected reconstruction error is the same as maximizing a lower bound on the *mutual information*, I(X,Y). Denoising autoencoders can thus be justified by the objective that output Y captures as much information as possible about input X, although Y is a function of corrupted input.

With *convolutional neural networks* (CNNs), the temporal structure of speech can also be exploited. For example, the WaveNet DNN is built from stacked, dilated, convolution layers [21]. It works directly on the input waveform and predicts the current output signal sample based on the knowledge of a few preceding and following input samples (output too, for autoregressive models).

A Bayesian application of the WaveNet was also implemented, which directly predicts the clean speech audio samples by estimating the prior distribution and the likelihood function of clean speech using WaveNet-like architectures [22]. The Bayesian models broaden the generalizability of deep learning, while the accuracy of modeling is improved by deep learning.

Using the known equivalence of CNNs to regression networks, if the former is trained with the same receptive field size and a lot more speech and noise examples of the same distributions, a 4-layer DNN regression-based speech enhancement was proposed in [23], which we detail in the next section. By using temporal context data, the spectral bin independence assumption is relaxed.

5. Proposed method and performance comparisons

The method we propose for use in forensics extends the capabilities of the method described in [23], by also capturing measures of the quality of input speech. This is done by using the three supplemental input features described in [24], namely the *inverse linear cepstral peak* (ILCP), the *log-windowed autocorrelation lag energy* (logWALE), and the *modified spectral autocorrelation peak-to-valley ratio* (MSAPVR). The proposed network, shown in Fig. 1, uses data packs of 350 features to encode each 32 ms analysis window of input signal, with 50% overlap. A data pack is made of 257 log-spectral features and 93 cepstral features – 31 *Mel-frequency cepstral coefficients* (MFCCs), including the energy, as well as their variation speeds and accelerations. A set of 8 consecutive data packs, and the three new features, computed only for the current window, are fed into the DNN at once.

The network includes three 2803-node hidden layers, with sigmoid activation, and a final layer with linear activation. The output of the DNN is made of two log-spectra (one per channel) and the same 93 cepstral features. The output log-spectra are then inverse Fourier transformed into a time domain series of enhanced speech windows, which are assembled by 50% *overlap and add* (OLA).

The training data for the neural network was extracted from 4620 TIMIT corpus speech files [25], after corruption with noises from Hu corpus [26].

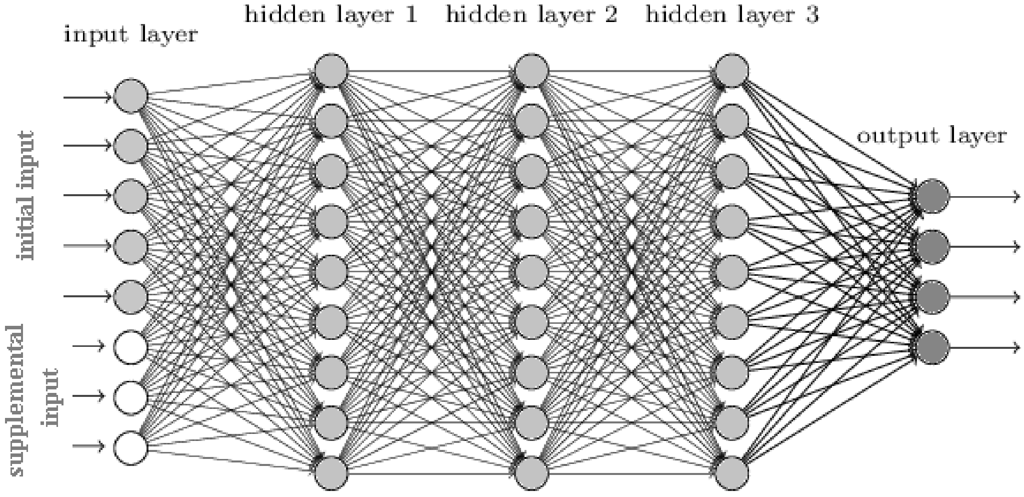

Fig. 1. The schematic of the proposed neural network structure

80% of the speech files were corrupted with one of the 100 points noise types, at one randomly selected SNR from 15 dB, 10 dB, 5 dB, 0 dB, and –5 dB, while on the other 20%, random isotropic noises and reverberations were applied.

For initialization, each hidden layer received unsupervised pre-training for 20 epochs as *restricted Boltzmann machines* (RBMs), with a learning rate of 0.0005, except for the output layer, initialized randomly.

Over the 50 epochs of training, the learning rate was set to decrease from 0.1 with 10% per epoch, with a momentum of 0.8. For regularization, the dropout technique was used, by 10% at the input layer and 20% at each hidden layer. The training ended on an early stopping basis, after 100 h of training data, using the *minimum mean squared error* (MMSE) cost function.

The performance of proposed method was compared to the methods described in [12], [22] and [23]. Three comparisons were performed in terms of commonly used objective measures, namely the *perceptual evaluation of speech quality* (PESQ) [27], the *short-time objective intelligibility* (STOI) [28], and the *Itakura-Saito* (IS) measure [29], on the NSDTSEA *test* corpus [30].

The preservation of speech throughout the enhancement process is paramount in forensics, so that we also estimated the *word error rate* (WER) on the SSC-eval corpus, made of 3035 spontaneous and noisy speech files covering about 3.5 hours, and on the NSDTSEA testset, covering about 1.2 hours in 824 speech files. The WER on SSC-eval was measured with the *automatic speech* *recognition* (ASR) for Romanian described in [31], while on NSDTSEA with the baseline Kaldi ASR, trained on the *train-clean-100* LibriSpeech corpus [32].

The results of performance comparisons are shown in Table 1, from where it comes out that proposed method outperforms the methods it was compared to.

Table 1

Comparison of proposed method to other methods

Method	WER [%] on SSC-eval	WER [%] on NSDTSEA	PESQ	STOI	IS
Blind reverberation and noise estimation, with Bayesian filtering [12]	21.73	9.42	2.47	0.88	53.1
Dilated CNN (WavNet) [22]	47.32	10.23	1.15	0.06	57.0
Regression DNN-GV [23]	20.68	8.47	2.26	0.87	19.4
Unprocessed	20.02	8.80	–	–	–
Proposed	**20.34**	**8.21**	**2.49**	**0.90**	**16.6**

Given that the dilated CNN [22] works on bare waveform, and had no training whatsoever for enhancement of reverberated speech, its results indicate a low generalization capability to reverberated speech.

The method in [23] and the proposed network, both achieve better WER scores than for unprocessed speech in the NSDTSEA dataset, mainly because of the normalization effect of the more hand-crafted features used. However, the ASR system used on the SSC-eval is already very robust against noises presented.

The inclusion of quality measures as inputs helps the neural network to generalize better to unseen speakers, noise types, and reverberated rooms.

6. Conclusions and future work

In this paper, a palette of forensic speech enhancement techniques was described, as an introduction to the field, and a DNN-based speech enhancement method was proposed. The field of forensics deals with audio that is often obtained in difficult conditions and is likely to be relied upon in a court of law. Thus it must preserve the authentic features of speech and speakers.

It was shown that the spectral subtraction usually produces musical noise but may be very handy for easy to remove noises, while the knowledge of the impulse response of the filter allows the removal of its convolution effects by inverse filtering. With the use of more crafted input features, which often concentrate relevant information, the speech recognition performance is improved, and the enhanced speech is easier to listen to. The training with speech quality measures and random corruption with noise was shown to achieve state-of-the-art generalization capability to different speakers and unseen noise environments.

The results compared in Table 1 show that the proposed method achieves the least loss in WER, the highest PESQ, STOI (higher are better), and the lowest IS (lower means better), although on SSC-eval is has produced a small disturbance for unprocessed speech. The training of the network with various noise and reverberation conditions and speech quality information led to increased speech recognition performance, close to that for the unprocessed corpora. However, for forensic applications, lower IS values, together with lower WER on noisy speech corpora, show better speech preservation.

In our future work, we intend to test other features, such as Gammatone filterbank power spectra and multi-resolution cochleagram feature, as improved information carrier features for DNNs input [18].

The objective quality results are expected to improve by working more on the granularity of the time-frequency representations and on a better selection of conditioning information, whose study was also left for a future work.

Acknowledgment

This work was partly supported by a grant of the Romanian Ministry of Research and Innovation, CCCDI – UEFISCDI, project number PN-III-P1-1.2-PCCDI-2017-0818 / 73PCCDI, within PNCDI III.

REFERENCES

[1] *The AES Historical Committee*, “How AES began”, AES website, 2011, visited 2019.

[2] *N. Wiener*, Extrapolation, interpolation and smoothing of stationary time series, John Wiley & Sons, Inc., New York, 1949.

[3] *S.W. Smith*, The scientist and engineer's guide to digital signal processing, California Technical Publishing, 1997.

[4] *B. Widrow, J.G.R. Glover Jr., J.M. McCool* et al., “Adaptive Noise Cancelling: Principles and Applications”, Proceedings of IEEE, **vol. 63**, no. 12, 1975, pp. 1692-1716.

[5] *S.F. Boll*, “Suppression of Acoustic Noise in Speech Using Spectral Subtraction”, ICASSP, **vol. 27**, no. 2, 1979, pp. 113-120.

[6] *R. McAulay, M. Malpass*, “Speech Enhancement Using a Soft-decision Noise Suppression Filter”, ICASSP, **vol. 28**, no. 2, 1980, pp. 137-145.

[7] *Y. Ephraim, D. Malah*, “Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator”, ICASSP, **vol. 32**, no. 6, 1984, pp. 1109-1121.

[8] *A. Ng*, “Neural Networks and Deep Learning”, Coursera, deeplearning.ai, visited 2019.

[9] *Lalchhandami, M. Pal*, “An Auto-correlation Based Speech Enhancement Algorithm”, International Journal of Engineering Research and Development, **vol. 7**, no. 5, June 2013, pp. 23-30.

[10] *M. Berouti , R. Schwartz, and J. Makhoul*, “Enhancement of Speech Corrupted by Acoustic Noise”, ICASSP, **vol. 4**, 1979, pp. 208-211.

[11] *N. Upadhyay and A. Karmakar*, “Speech Enhancement Using Spectral Subtraction-type Algorithms: A Comparison and Simulation Study”, Procedia Computing Science, **vol. 54**, 2015, pp. 574-584.

[12] *C.S.J. Doire*, Single-channel enhancement of speech corrupted by reverberation and noise, PhD Thesis, Imperial College London, 2016.

[13] *Y. Ghanbari and M.R. Karami-M.*, "A New Approach for Speech Enhancement Based on The Adaptive Thresholding of The Wavelet Packets", Speech Communications, **vol. 48**, no. 8, 2006, pp. 927-940.

[14] *X. Zou, P. Jancovic, J. Liu, and M. Kokuer*, "Speech Signal Enhancement Based on MAP Algorithm in The ICA Space", ICASSP, **vol. 56**, no. 5, 2008, pp. 1812-1820.

[15] *A. Hyvarinen and E. Oja*, "Independent Component Analysis: Algorithms and Applications", Neural Networks, **vol. 13**, no. 4, 2000, pp. 411-430.

[16] *L. Singh and S. Sridharan*, "Speech Enhancement for Forensic Applications Using Dynamic Time Warping and Wavelet Packet Analysis", the IEEE Region 10 Annual Conference on Speech and Image Technologies for Computing and Telecomm., **vol. 2**, 1997, pp. 475-478.

[17] *F. Denk, J.P.C.L. da Costa, and M.A. Silveira*, "Enhanced Forensic Multiple Speaker Recognition in The Presence of Coloured Noise", the 8th IEEE International Conference on Signal Processing for Communication Systems, 2014, pp. 1-7.

[18] *D. Wang and G.J. Brown*, Computational Auditory Scene Analysis, Wiley and Sons, 2006.

[19] *Y. Bengio*, "Learning Deep Architectures for AI", Foundamental Trends in Machine Learning, **vol. 2**, no. 1, 2009, pp. 1-127.

[20] *P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol*, "Extracting And Composing Robust Features With Denoising Autoencoders", in Proceedings of the 25-th International Conference on Machine Learning, Helsinki, Finland, July 5-9, 2008, pp. 1096-1103.

[21] *D. Rethage, J. Pons, and X. Serra*, "A WaveNet for Speech Enhancement", Proceedings of the 43rd ICASSP, Calgary, AB, Canada, April 15-20, 2018, pp. 5069-5073.

[22] *K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florencio, and M. Hasegawa-Johnson*, "Speech Enhancement Using Bayesian WavNet", in Proceedings of INTERSPEECH, Stokholm, Sweden, August 20-24, 2017, pp. 2013-2017.

[23] *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, 2015, pp. 7-19.

[24] *G. Pop, D. Drăghicescu, D. Burileanu*, "A Quality-Aware Forensic Speaker Recognition System", Romanian Journal of Information Science and Technology, **vol. 17**, no. 2, 2014, pp. 134-149.

[25] *J.S. Garofolo*, Getting started with the DARPA TIMIT CD-ROM: an acoustic phonetic continuous speech database, NIST Tech Report, 1988.

[26] *G. Hu*, 100 nonspeech environmental sounds, 2004, http://www.cse.ohio-state.edu/pnl/corpus/HuCorpus.html, visited 2019.

[27] *A. Rix, J. Beerends, M. Hollier, and A. Hekstra*, "Perceptual Evaluation of Speech Quality (PESQ) - A New Method for Speech Quality Assessment of Telephone Networks and Codecs.", in Proc of ICASSP, **vol. 2**, 2001, pp. 749-752.

[28] *C. Taal, R. Hendriks, R. Heusdens, and J. Jensen*, "A Short-Time Objective Intelligibility Measure for Time-Frequency Weighted Noisy Speech," in Proceedings of ICASSP, 2010, pp. 4214-4217.

[29] *P.C. Loizou*, Speech enhancement: Theory and practice, second edition, CRC Press, Boca Raton, London, New York, 2013.

[30] *C. Valentini-Botinhao*, Noisy speech database for training speech enhancement algorithms and TTS models [NSDTSEA dataset], University of Edinburgh, School of Informatics, Centre for Speech Technology Research (CSTR), https://doi.org/10.7488/ds/1356, 2016, visited 2019.

[31] *A.L. Georgescu, H. Cucu*, "Automatic Annotation of Speech Corpora Using Complementary GMM and DNN Acoustic Models", in Proceedings of the 40-th International Conference on Telecommunications and Signal Processing (TSP), Barcelona, Spain, July 5-7, 2018.

[32] *V. Panayotov, G. Chen, D. Povey, and S. Khudanpur*, "LibriSpeech: an ASR corpus based on public domain audio books", ICASSP 2015, available online: https://www.openslr.org/12/, visited 2019.