

Speech Enhancement Based on Deep AutoEncoder for Remote Arabic Speech Recognition

Bilal Dendani^{1,2(\Box)}, Halima Bahi¹, and Toufik Sari^{1,2}

¹ Computer Science Department, University Badji Mokhtar Annaba, Annaba, Algeria {bilal.dendani,halima.bahi, toufik.sari}@univ-annaba.org
² Labged Laboratory, University Badji Mokhtar Annaba, Annaba, Algeria

Abstract. Remote applications that deal with speech need the speech signal to be compressed. First, speech coding transforms the continuous waveform into a numerical form. Then, the digitized signal is compressed with or without loss of information. This transformation affects the original waveform and degrades performances for further recognition of the speech signal. Meanwhile, the transmission is another source of speech degradation. To restore the original "clean" speech, speech enhancement (SE) is widely used, and deep learning algorithms are state-of-the-art, nowadays. In this paper, the target application is a remote Arabic speech recognition system, and the aim of using SE is to improve the accuracy of the speech recognizer. For that purpose, a Deep Auto Encoder (DAE) is used. The effect of the DAE-based SE is studied through different configurations, and the performances are evaluated through accuracy. The results showed an improvement of about 3.17 between the accuracy prior to the SE and that computed with the enhanced speech.

Keywords: Mobile speech recognition \cdot Arabic language \cdot Speech enhancement \cdot Deep learning \cdot Deep AutoEncoder (DAE)

1 Introduction

IN the last few years, there has been a large availability of communication technology, a growing number of people are accessing the internet easily, and digital devices have become increasingly powerful and cheaper; this sparkled the emergence of mobile Automatic Speech Recognition (ASR) applications. Indeed, speech recognition could be used as a core component of several applications such as healthcare [1, 2] or language learning [3, 4]. Two approaches can be used for the deployment of mobile ASR applications, a client-based approach or a remote server-based approach. The remote speech recognition architectures are Network Speech Recognition (NSR) and Distributed Speech Recognition (DSR) [5]; in both of them, the speech signal is captured at the client-side and is transmitted to the server-side for decoding and recognition. The feature extraction stage is located at the client-side for DSR architectures and the server-side for NSR ones.

© Springer Nature Switzerland AG 2020

A. El Moataz et al. (Eds.): ICISP 2020, LNCS 12119, pp. 221–229, 2020. https://doi.org/10.1007/978-3-030-51935-3_24 In remote speech recognition, the speech signal has to be transmitted, and for that purpose, the speech signal needs to be coded. The coding process aims to minimize bits used to represent a speech signal while preserving the quality of the transmitted speech [6]. An ideal speech coder represents the input speech by a few bits without quality degradation. Obviously, there is a trade-off between the codec bit rate and the quality of the transmitted voice. The transmission process is another source of signal degradation [6]. This paper is about a remote speech recognition system of which performances highly depend on the quality of the coded and transmitted (transcoded) speech signal; to restore the original signal, SE is required.

The SE technique is intended to recover the clean speech signal from its corrupted form. For that purpose, many methods exist such as the minimum mean square error [7] or the Wiener filtering [8], these traditional methods perform well when the speech is corrupted by stationary noises, but are limited in the non-stationary noise environments. Recently, Deep Learning (DL) algorithms have become state-of-the-art in SE, and they deal with stationary and non-stationary noises as well [9–20].

While a great number of researches have addressed the SE of noisy speech, researches that deal with coded SE are rare. However, in [12], the authors have implemented a Convolutional Neural Network (CNN) to enhance the coded speech. The proposed CNN architecture included three kinds of layers: convolutional, max pooling, and up-sampling layers. The authors compared their solution to the G.711 post-processing as a baseline and found that their proposition improved the speech quality in terms of PESQ (Perceptual Evaluation of Speech Quality) for G.711, G.726, and AMR-WB coders. On another side, works focusing on the enhancement of Arabic speech, and based on DL models are quasi-inexistent.

The paper tackles the problem of SE of transcoded speech in a remote context, implemented throughout a network-based ASR architecture. In this paper, we propose the use of a DAE (Deep AutoEncoder) algorithm for SE; the effect of the DAE is studied in different contexts of the input window. The performances of the proposed architecture are assessed in terms of the accuracy of the speech recognizer. The speech recognizer is considered as a black box and is based on Hidden Markov Models (HMMs) [21, 22].

The next section outlines the DL-based SE models. Section 3 presents the target application and the context of the study from speech coding towards SE and recognition. Results and discussion are presented in Sect. 4. Finally, a conclusion is drawn.

2 Related Works

Recently, different DL models, known as data-driven models, have been explored, and they provided good results for SE. Most of the time, the DL-based SE methods model the non-linearity between noisy speech signals and clean speech signals, when the clean speech is accessible, thus they can recover clean speech signals from corrupted versions.

The fully connected-based models exploit the mapping function to enhance speech. These models include the deep denoising auto-encoders (DDAE) [9, 13], and the fully connected deep neural network (DNN) models [14, 15]. Lu et al. [9] have applied a

DDAE algorithm for noise reduction and SE, in the training stage, both input and output were clean speech signals. Thus, the DDAE is expected to only encode statistical information of the clean speech. In their further experiments reported in [13], the DDAE was trained using noisy speech as input and clean speech as output. Therefore, the DDAE is expected to explicitly learn the difference between clean and noisy signals. Furthermore, Xu et al. [14, 15] proposed a regression-based SE DNN approach to estimate the mapping function between noisy and clean utterances. The DNN model was trained on a large data of about 100 h with multiple noise conditions. Other works focused on recurrent neural networks (RNN) using noisy/clean speech pairs as input and output respectively during the training stage [16].

While the fully connected models use a lot of parameters, CNN-based SE architectures use a fewer number of parameters. Indeed, the CNN-based models deal with the local temporal and spectral structure of the speech. Fu et al. [17] investigated the CNN model to restore clean speech from a noisy version using SNR-aware algorithms. Later, they applied a fully convolution network (FCN) to model simultaneously the high and the low-frequency components of a raw waveform [19]. The results of the FCN algorithm outperformed those of the CNN and the DNN when the waveforms are used as inputs. The FCN has been applied also in an embedded system to enhance speech for hearing aids [18]. The authors in [11] used both convolutional and recurrent neural network architectures for SE. The proposed model improves the PESQ and the WER (Word Error Rate) by 0.6 and 1% respectively, the model also generalizes better on unseen noise. Other DL-based models have been tested for SE such as the generative adversarial networks [20].

3 Speech Enhancement Based on Deep AutoEncoder for Remote Arabic Speech Recognition

Our investigation is about the impact of SE on network speech recognizer performances. In the NSR model, the speech signal is captured at the client-side, then, it is transmitted to the server where the recognition is performed [23, 24]. Figure 1 depicts the block diagram of the proposed remote Arabic speech recognition system.

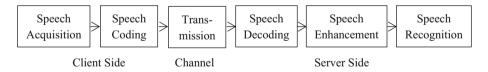


Fig. 1. Block diagram for remote Arabic speech recognition

The received signal is coded at the client-side. For the purpose of this study, the G.711 codec is applied. Then, the coded speech is transmitted to the server-side in reallife conditions. At the server-side, the received signal is enhanced by DAE previously trained. Finally, the HMM-based Arabic speech recognizer is fed with the enhanced speech signal. The computed accuracy serves to assess the performances of the DAEbased SE system.

3.1 G.711 Codec

G.711 is a waveform, lossless compression speech codec, developed by ITU-T [25] and released in 1972. G.711 codec is mainly used for speech transmission over the telephony system, provides a bit rate at 64 kbps and exists in two versions: μ -law and A-law. Both versions code speech sample with 8 bits, provide a reduction of 50% of file size and use a bandwidth compared to the original signal which uses 16 bits as a sample size. In [24], experiments, made in the context of the Arabic language, showed that the narrowband high bit rate codec's G.711 provides good performances.

3.2 Deep AutoEncoder for Arabic Speech Enhancement

DAE is a feedforward neural network with many hidden layers. Given a pattern X, the encoding layers reduce X into a smaller dimension, then, the decoder layers reconstruct the pattern Y (a version close to X). When applied in SE, DAEs aim to recover the clean version of the speech from its corrupted one. For that purpose, the model is trained on noisy/clean pairs of speech. In the training stage, the DAE is fed with a corrupted version of the speech signal; in our case with the transcoded speech, as input and the output is the original signal. During the test stage, the DAE is fed with the transcoded version of the speech and it is intended to provide the clean one (Fig. 2).

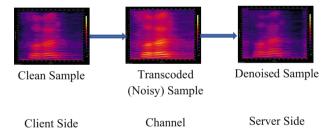


Fig. 2. Spectrograms for a sample corresponding to the word "Zero", in several locations

We endeavor to restore the original (clean) speech from transcoded speech. We trained the DAE based on clean speech samples, using an Arabic speech corpus for isolated words. The incoming speech is framed into windows of 512 samples. From each frame are extracted coefficients corresponding to the spectrum power.

3.3 Speech Recognizer

For the recognition purpose, HMMs stand for the acoustic models. The training dataset contains approximately 70% of the utterances in which the list of isolated words is pronounced by 50 native Arabic speakers. Testing data represents 30% of the total

number of utterances. The HMM-based speech recognizer is used as a black box without tuning during our experiments related to SE.

4 Experimental Results and Discussion

4.1 Arabic Speech Data Set

The corpus used in experimentations is an Arabic speech corpus for isolated words [26]. The corpus has been developed by the Department of Management Information Systems of King Faisal University. It contains 9992 recorded utterances of 50 speakers pronouncing 20 words. The list of the pronounced words is:

بنعم التسديد الرصيد التحويل التنشيط بتسعه بثمانيه سبعه سته خمسه أربعه بثلاثه اثنان واحد صفر } إنعم التسديد الرصيد التحويل التنشيط بسعه بمانيه مسبعه بسته بخمسه أربعه بثلاثه التيان واحد مصفر }

Translated as: {Zero, One, Two, Three, Four, Five, Six, Seven, Eight, Nine, Activation, Transfer, Balance, Payment, Yes, No, Funding, Data, Account, End}.

The recordings are sampled at 44100 Hz (stereo channel) with 16 bits precision. The corpus has been adapted by down-sampling it to 8 kHz sampling rate as if the recordings were done through mobile devices, and a mono channel is considered.

The DAE is trained with approximately 1500 utterances from 50 speakers. The test dataset contains 445 utterances from 15 speakers.

4.2 Effects of Coding, Transmission, and Enhancement

Prior to the coding and transmission tasks, the accuracy is computed at the client-side. Then, to assess the effect of the transmission process, the accuracy is computed at the server-side after the reception of the uncompressed signal. Finally, to assess coding and transmission effects, the accuracy is computed at the server-side after the reception of the compressed signal. The HMM-based recognition takes place without prior enhancement of the incoming signal. Table 1 reports the accuracy of the ASR system in the three situations considering the utterances of the test dataset.

Mode	Offline	Online Uncompressed	Online G.711
Accuracy	64,44	64,33	62,55

Table 1. Accuracy Offline, Online with and without compression

Table 1 shows that transmission slightly decreased the accuracy while the coding/transmission clearly did with a loss of about 1.89.

To overcome the degradation of the accuracy due to the use of coding and transmission, we suggest the use of the DAE algorithm to enhance the corrupted signal. Figure 3 shows the impact of the SE stage on the accuracy computed at the server-side.

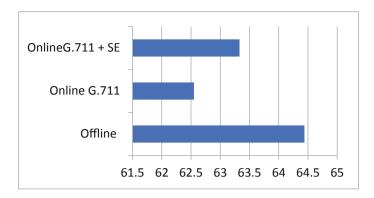


Fig. 3. Accuracy computed in offline mode, after transmission, and after SE

Figure 3 shows an improvement of the accuracy when the speech signal submitted to the decoder was already enhanced.

4.3 Effect of the Context Length

Although the proposed DNN is only intended to enhance the speech, the first tuning concerns the inclusion of the left/right context of the target window. As already said, from each window of 512 samples were extracted the corresponding spectrum power coefficients, and as the obtained vector is symmetric, only 257 coefficients are kept. Thus, the input layer, as well as the output layer, comprises 257 neurons. Alternately, as the context of neighbor frames is important in speech, we make experiments where both sides of a target window are considered. Figure 4 depicts DAE architecture where two frames from the left and two from the right of the target window are considered.

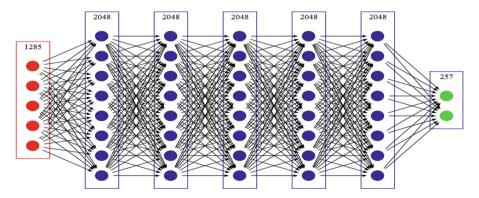


Fig. 4. DAE where the input corresponds to five consecutive frames and the output stands for the middle one

Table 2 reports accuracy obtained after the enhancement of the received signal considering models of DAE trained with several context length of the target window. Considering the DAE depicted in Fig. 4, and the utterances of the test dataset, the results are reported in Table 2.

#Input frames	01	03	05	07	11	13
Accuracy	63.333	64.276	65.724	53.153	50.922	48.981

Table 2. Accuracy of enhanced speech considering several context lengths

Table 2 shows that an increasing number of windows surrounding the target speech frame impacts accuracy. Indeed, adding one window at each side improves the accuracy, and by enlarging the context to two windows at each side, the accuracy grows to 65.72%. Although, when the context becomes larger and the complexity of the DAE increases (three windows per side or more), the accuracy decreased drastically.

In summary, the DAE-based SE process including one or two windows at each side of the target frame (or even only the target frame) improved the accuracy of the ASR application and provides a robust remote ASR system even when the HMMs were trained on clean speech.

5 Conclusion

Remote speech recognition is a challenging task and is for much importance for many applications such as security, healthcare, education, etc. The presented work aims to improve the communication client-server on mobile networks by reducing the impact of degraded ASR performance which is introduced by speech G.711 codec. A deep auto-encoder algorithm is used to enhance speech degraded by the G.711 codec and the transmission. Results showed that the accuracy of speech recognizer has been increased when enhancement of the speech is performed. Future works will address several configurations of the DAE architecture depth, and the number of neurons in each layer. One of the remaining challenges is to explore other algorithms such as the CNN and to consider larger Arabic speech corpora.

References

- 1. Hossain, M.S.: Patient sate recognition system for healthcare using speech and facial expressions. J. Med. Syst. 40, 12 (2016)
- Necibi, K., Bahi, H., Sari, T.: Automatic speech recognition technology for speech disorders analysis. In: Speech, Image and Language Processing for Human-Computer Interaction: Multi-modal Advancements. IGI-global (2012)
- 3. O'Brien, M.G., et al.: Directions for the future of technology in pronunciation research and teaching. J. Second Lang. Pronunciation **4**(2), 182–207 (2018)
- Lee, J., Lee, C.H., Kim, D.-W., Kang, B.-Y.: Smartphone-assisted pronunciation learning technique for ambient intelligence. IEEE Access 5, 312–325 (2016)

- Schmitt, A., Zaykovskiy, D., Minker, W.: Speech recognition for mobile devices. Int. J. Speech Technol. 11(2), 63–72 (2008)
- 6. Chu, W.C.: Speech Coding Algorithms: Foundation and Evolution of Standardized Coders. Wiley, Hoboken (2004)
- Ephraim, Y., Malah, D.: Speech enhancement using a minimum-mean square error shorttime spectral amplitude estimator. IEEE Trans. Acoust. Speech Sig. Process. 32, 1109–1121 (1984)
- Sreenivas, T.V., Pradeep, K.: Codebook constrained Wiener filtering for speech enhancement. IEEE Trans. Speech Audio Process. 4(5), 383–389 (1996)
- Lu, X., Matsuda, S., Hori, C., Kashioka, H.: Speech restoration based on deep learning autoencoder with layer-wised pretraining. In: Proceedings of INTERSPEECH, Portland, USA, pp. 1504–1507 (2012)
- Kumar, A., Florencio, D.: Speech enhancement in multiple-noise conditions using deep neural networks. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, San Francisco, USA (2016)
- 11. Zhao, H., Zarar, S., Tashev, I., Lee, C.H.: Convolutional-recurrent neural networks for speech enhancement. In: Proceedings of ICASSP, pp. 2401–2405 (2018)
- 12. Zhao, Z., Liu, H., Fingscheidt, T.: Convolutional neural networks to enhance coded speech. IEEE/ACM Trans. Audio Speech Lang. Process. **27**(4), 663–678 (2018)
- Lu, X., Tsao, Y., Matsuda, S., Hori, C.: Speech enhancement based on deep denoising autoencoder. In: Annual Conference of the International Speech Communication Association, INTERSPEECH, pp. 436–440 (2013)
- Xu, Y.: An experimental study on speech enhancement based on deep neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. 23(1), 7–19 (2015)
- Xu, Y., Du, J., Dai, L.R., Lee, C.H.: A regression approach to speech enhancement based on deep neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. 23(1), 7–19 (2015)
- Weninger, F., et al.: Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. In: Vincent, E., Yeredor, A., Koldovský, Z., Tichavský, P. (eds.) LVA/ICA 2015. LNCS, vol. 9237, pp. 91–99. Springer, Heidelberg (2015). https:// doi.org/10.1007/978-3-319-22482-4_11
- 17. Fu, S.W., Tsao, Y., Lu, X.: SNR-aware convolutional neural network modeling for speech enhancement. In: Proceedings of Interspeech (2016)
- Park, S.R., Lee, J.W.: A fully convolutional neural network for speech enhancement. In: Annual Conference of the International Speech Communication Association INTER-SPEECH, vol. 2017, no. 2, pp. 1993–1997 (2017)
- Fu, S.W., Tsao, Y., Lu, X., Kawai, H.: Raw waveform-based speech enhancement by fully convolutional networks. In: Asia-Pacific Signal & Information Processing Association Annual Summit Conference APSIPA ASC 2017, vol. 2018, no. December, pp. 6–12 (2018)
- Pascual, S., Bonafonte, A., Serra, J.: SEGAN: speech enhancement generative adversarial network. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, vol. 2017, pp. 3642–3646 (2017)
- Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77–2, 257–286 (1989)
- Bahi, H., Sellami, M.: Combination of vector quantization and hidden Markov models for Arabic speech recognition. In: Proceedings ACS/IEEE International Conference on Computer Systems and Applications, Beirut, Lebanon (2001)
- Tan, Z.H., Varga, I.: Network, distributed and embedded speech recognition: an overview. In: Tan, Z.-H., Lindberg, B. (eds.) Automatic Speech Recognition on Mobile Devices and over Communication Networks. Advances in Pattern Recognition. Springer, London (2008). https://doi.org/10.1007/978-1-84800-143-5_1

- 24. Dendani, B., Bahi, H., Sari, T.: A ubiquitous application for Arabic speech recognition. In: Proceedings of the International Conference on Artificial Intelligence and Information Technology, Ouargla, Algeria, pp. 281–284 (2019)
- Digital cellular telecommunications system (Phase 2+); Half rate speech; ANSI-C code for the GSM half rate speech codec (GSM 06.06 version 7.0.1 Release 1998) GLOBAL SYSTEM FOR MOBILE COMMUNICATIONS (2000)
- Alalshekmubarak, A., Smith, L.S.: On improving the classification capability of reservoir computing for Arabic speech recognition. In: Wermter, S., et al. (eds.) ICANN 2014. LNCS, vol. 8681, pp. 225–232. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11179-7_29