Recent Developments in Electronics and Communication Systems
KVS Ramachandra Murthy et al. (Eds.)

doi:10.3233/ATDE221279

Improving Speech Quality Using Deep Neural Network-Based Manipulation of Cepstral Excitation

Mahesh K. Singh[a,1], D. Lavanya[b], Ch. Asha Madhuri[c], P. Ramesh[d], V. Satyanarayana[e]

[a]*Adjunct Prof., Department of ECE, Aditya Engineering College, Surampalem*
[b,c,d,e]*Department of ECE, Aditya College of Engineering, Surampalem, India*

Abstract. The source-filter paradigm, which is used in human speech development, would be used in this proposal to increase speech quality. Excitation signals in the cepstral domain are amplified using a sophisticated neural network (DNN). As a result, they compared the effects of normalization on two different types of goals. Using the cepstral excitation manipulation approach (CEM) instead of traditional signal processing-based cepstral excitation, we were able to reduce noise by 1.5 dB. Studying various combinations of data demonstrates that envelope and excitation enhancement are also present. When adopting a low signal-to-noise ratio, good speech intelligibility can be achieved even with a lot of noise attenuation. It can be shown that a traditional pure stats system can attenuate noise better than its modern counterpart because it is older and has a larger sample size. Older classic pure stats systems use less processing power, hence they're more efficient. We've created a new a priori method for measuring SNR for voice-enhancement apps. The cepstral domain spectral envelope is estimated using a DNN. We can improve the quality of low-order noise models while also providing listeners with a natural and pleasurable background noise experience using our CEM approach.

Keywords: DNN, CEM, Cepstral analysis, Speech quality, Speech recognition

1. Introduction

The field of speech enhancement is one that's expanding and becoming increasingly important. In order to communicate in the most natural way possible, it aims to improve the quality of your voice and your ability to speak clearly. Noise coupling, echoes, and bandwidth restriction are all ways that speech communications might be distorted. To address this issue, a number of algorithms have been created and improved throughout time [1, 2, 3, 4]. Recent improvements in speech technology, which are state-of-the-art, are increasing the usage of new deep learning technologies, although single-channel noise reduction is still a problem. Standard DNN-based improvement models have a hurdle when analyzing huge signals in a frame-based manner, as shown in [5, 6, 7, 8, 9]. Sections following this one will cover some of the most recent changes [3, 11]. Following the DNN-based learning of spectral weighting rules, a sketch of less integrated

[1] Corresponding Author, Adjunct Professor, Department of ECE, Aditya Engineering College, Surampalem, India; E-mail: mahesh.singh@accendere.co.in

techniques has been examined [12], which respects traditional and numerical speech in parts while utilizing current technology to its advantage [10, 11, 12, 13, 14, 15].

When moving from end-to-end solutions to classic structures while taking advantage of new technologies, the publication demonstrates numerous granular levels as shown in figure 1. Consider the ideal binary screen and the ideal ratio screen, for instance. Rosenkranz et al. have recently revitalized AR model-based spectral envelope codebook work in the cepstral domain. They integrate a previous auto-regressive-based method with a noise estimator to avoid needing a noise codebook [16, 17, 18, 19]. Figure 1 shows the numerous speech-producing organs, as well as the elements that go along with them. The vocal tract, which extends from the vocal cords to the mouth, is a non-uniform tube. It enables the transmission of acoustically created vocal tract sounds. There are a variety of noises that can be expelled from the vocal tract by using the oral cavity or nasal cavity. Depending on how the vocal cords work, basic speech sounds are classified as either voiced or unvoiced in speech processing [4, 20, 21, 22, 23, 24].

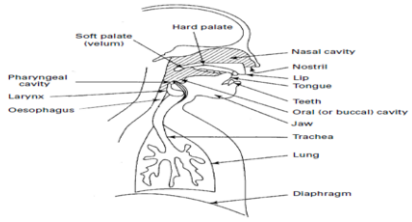

Figure 1. Schematic representation of human speech production organ [5].

To make up for the lack of noise reduction between harmonics, an SPP estimator was used instead of a spectral envelope estimator to estimate clean speech. Another study examined this issue previous to writing this one, and it's brought up here because of the impact of target normalization on harmonic preservation [14]. To ensure that the person speaking is who they claim to be, SES uses biometrics. Users can restrict access to services such as voice authentication, speaker authentication, or speaker detection, database access enhancement, and security control for sensitive data by using this method, which is also known as a true-false binary decision issue [17, 18].

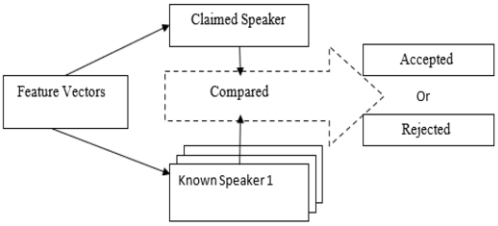

Figure 2. Schematic representation of speaker enhancement system

The claimed speaker is allowed or denied based on a comparison of the suitable matching score to a predetermined threshold. Establishing the identity of the speaker isn't vital in these systems; only improvement is. A biometric system is the most critical SES application. The speaker enhancement system is another method of identifying a single target speaker in an open-set environment [15, 16]. The authors seek to eliminate both codebooks by predicting the parameters of the AR models for speech and noise using a single network that assumes line spectral pairings. In order to address the failure to keep harmonic noise to a minimum. Estimates based on SPP are also used. By employing the

estimated entities in both scenarios to design a wiener filter, it's a step closer to integrating DNNs into a statistical speech improvement architecture [15, 19].

2. Literature Review

Several combinations of imaginary and determined signals, such as the noise explosion formed after grating closure and the determined thick pulses, were discovered to be capable of decaying in 1995, and the algorithm was named after the discovery. This unique method is a powerful tool for analyzing relevant characteristics of speech signal source components and it has been implemented in several applications [2]. In 1997, we proposed unique algorithms for natural quality variable rate spectral speech coding, with an average speed of 2,2 kbps for speech conversation and 2,8 kbps for speech, with an average speed of 2,2 kbps for speech. A Fourier spectrum is used to model each of the coder's models, and recent advancements to the MBE method to classical multiband excitations are taken into consideration [3]. A number of strategies have been documented that contribute significantly to the achievement of high performance natural-sounding speech with a variable-rate spectral modelling coder operating at an average rate of 2.8 kbps.

These techniques include: Using a wiener-type spectral amplitude gain function, which was intended for implementation in January 2009, the speech enhancement technique sought to improve the quality of denoised speech by including voicing information into the function [19]. When comparing enhanced speech to speech collected using traditional wiener-type approaches that do not take into account the gain function computation model of speech output, the technique is proven. According to one approach of estimating the gain function is to use spectral estimates obtained from the MTW [4, 20]. The procedure can be resolved as a postprocessor without the need for a side report to be generated in the process [6]. Using interpolated VQ and nonlinear activity production, this work demonstrates how narrowband speech can be converted to wideband speech. The arranged wideband speech that results outperforms narrowband speech in terms of quality and is preferred by all listeners is the most popular [5].

Improve respectful speech by increasing the signal-to-reverberation ratio (SRR) regions in the temporal domain and by increasing the high signal-to-reverberation ratio (SRR) regions in spectral domain with temporal and spectral sensory processing identification [16]. Low late reverberations by spectrum reductions and increased high SRR regions through temporal processing were important benefits of both methods that were combined into a single solution [6, 7]. According to a recent report, new speech augmentation techniques are being developed constantly in 2017 to minimize the degradation of speech signals due to various auditory ambient components. Before executing the strategy [17]. The strategy is then implemented in two stages. Due to the fact that a priori signal-to-noise ratio (SNR) transmits important information about the mixing of speech with noise, the estimation of the SNR is a hot research topic in 2017. The source and filter uncertainty of a preliminary noised signal is reduced, allowing the degraded source to be subjected to an idealized excitation [8]. New research on a novel priority SNR estimator for speech enhancement applications published in 2018 reveals the authors' conclusions. In a noise reduction scenario with three different spectral weighting rules, the performance of a novel a priori SNR estimator was investigated [18]. Speech model-based speech quality will be improved by using a human voice output source-filter as part of this contribution for 2019. The proposed method for improving

the excitation signal in the cepstral domain makes use of a deep visual network to accomplish this. It was found that one of the target representations was unsuitable for use in actual systems, whilst the other was found to be excellent [9]. A semi-formal comparative division rating (CCG) subjective hearing test demonstrates that the proposed strategy outperforms the alternative approach. Using a semi-formal comparative category ration (CCR) listening test, the combination of cepstral envelope enhancement and DNN-based CEM exceeds the standard DD approach by two orders of magnitude [10].

3. Methodology for Hyper-Dimensional Computing:

The CEM approach's core configuration is presented in figure 2 with switch s3 in the top position. After that, the cosine transform type-2 is used, followed by an algorithm for predicting pitch. The pitch frequency reference bin index is calculated by selecting the bin with the greatest amplitude from a set of fundamental frequency representing bins.

3.1. DNN- Supported Cepstral Excitation Manipulation:

Figure 3 illustrates how a neutral network, rather than the classical signal processing methods that have been used up to this point, can absorb the unique deep learning chances we seek to examine the potential of the CEM idea. The traditional base line CEM is detailed in the upper route, whereas the traditional base line CEM is detailed in the lower route.

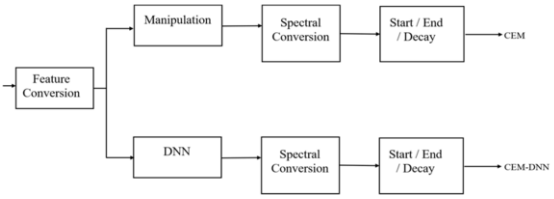

Figure 3. Block diagram of the CEM-DNN approach by using DNN

3.2. New Two-Stage Speech Enhancement

We will quickly present our code in this section, followed by the new two-stage speech enhancement technique, which will be detailed in the picture and used as a reference throughout the section.

4. DNN-Based Cepstral Excitation Manipulation

Experiment: The databases used in the creation of our system, as well as the quality metrics included in the final computation of the baselines and the proposed strategy, are described in the following section.

Data Bases: To make it easier to link our results to earlier work, we use the same database configuration for training, with improvement sets serving as training sets and the TIMIT database's test set serving as development sets. Finally, we present results from the NTT great wideband database, which acts as a test set and enables us to provide results from many databases.

Oracle and Motivation Experiments: First and foremost, we participate in two Oracle experiments that serve as inspiration for our study. In the context of MS's noise power estimate, both Oracle experiments show the performance of a priority SNR estimator with varied uses of restricted Oracle knowledge.

Table 1. Based on a -5dB SNR situation, the MSE loss for various network topologies is evaluated. with C_1^s (m) set of development objectives

N_4	$N_{N=64}$	$N_{N=128}$	$N_{N=256}$	$N_{N=512}$	$N_{N=1024}$
1	0.839	0.811	0.796	0.787	0.782
2	0.830	0.794	0.771	0.758	0.753
3	0.823	0.785	0.761	0.746	0.742
4	0.816	0.779	0.754	0.742	0.742
5	0.815	0.776	0.752	0.741	0.742
6	0.813	0.774	0.750	0.739	0.744

Table 1. display the MSE loss under the -5Db SNR condition for various covered lager structures NH and their node count Nn. This development set has to be constructed on a small quantity of data because the training method under all SNR conditions is time-consuming to complete. SNR conditions and C ls(m) development set targets are used to evaluate MSEL loss for various network-to-policy scenarios as in Table 2.

Table 2. MSEL loss for various network-to-policy scenarios

N_H	$N_{N=64}$	$N_{N=128}$	$N_{N=256}$	$N_{N=512}$	$N_{N=1024}$
1	0.744	0.679	0.643	0.654	0.648
2	0.732	0.661	0.632	0.622	0.615
3	0.726	0.654	0.631	0.608	0.604
4	0.721	0.648	0.632	0.603	0.600
5	0.719	0.645	0.633	0.601	0.601
6	0.718	0.643	0.636	0.600	0.603

The MSE loss appears to be the same for N N512,1024 in table-2 because the duration of the loss falls as a result of the targets' mean and variance being normalized. The microphone signal is then processed with CEM-DNN for a network trained with goal normalization, with only one reduction at the start and end.

5. Conclusion

This work investigated the usage of a DNN to CEM approach for estimating a priori SNR in a voice enhancement project. Two target representations were compared: One is poor for functional systems while the other is excellent. It would also be possible to demonstrate the significance of target normalization, and evaluate the usefulness of

reducing the predicted residual signal at the beginning and the end. In this study, we introduce a novel two-stage speech augmentation method based on a source-filter strategy for analyzing a denounced signal in advance. An adjustment is made to the cepstral domain excitation signal to create a better speech estimate, which is then used to adjust the a priori SNR calculation.

References

[1] Eishamy S, Madhu N, Tirry W, Fingscheidt T. A priori SNR computation for speech enhancement based on cepstral envelope estimation. In2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC) 2018 Sep 17 (pp. 351-355). IEEE.

[2] d'Alessandro C, Yegnanarayana B, Darsinos V. Decomposition of speech signals into deterministic and stochastic components. In1995 International Conference on Acoustics, Speech, and Signal Processing 1995 May 9 (Vol. 1, pp. 760-763). IEEE.

[3] Erzin E, Kumar A, Gersho A. Natural quality variable-rate spectral speech coding below 3.0 kbps. In1997 IEEE International Conference on Acoustics, Speech, and Signal Processing 1997 Apr 21 (Vol. 2, pp. 1579-1582). IEEE.

[4] Soon IY, Yeo CK. Bandwidth extension of narrowband speech using cepstral analysis. InProceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004. 2004 Oct 20 (pp. 242-245). IEEE.

[5] Singh MK, Singh AK, Singh N. Multimedia analysis for disguised voice and classification efficiency. Multimedia Tools and Applications. 2019 Oct;78(20):29395-411.

[6] Krishnamoorthy P, Prasanna SM. Reverberant speech enhancement by temporal and spectral processing. IEEE transactions on audio, speech, and language processing. 2009 Jan 13;17(2):253-66.

[7] Singh MK, Singh AK, Singh N. Disguised voice with fast and slow speech and its acoustic analysis. Int. J. Pure Appl. Math. 2018;11(14):241-6.

[8] Jyothi, K.D., Sekhar, M.S.R. and Kumar, S., 2021, October. Applications of Statistical Machine Learning Algorithms in Agriculture Management Processes. In *2021 6th International Conference on Signal Processing, Computing and Control (ISPCC)* (pp. 237-241). IEEE.

[9] Singh, M. K., Singh, A. K., & Singh, N. (2019). Multimedia utilization of non-computerized disguised voice and acoustic similarity measurement. *Multimedia Tools and Applications*, 1-16.

[10] Singh M, Nandan D, Kumar S. Statistical Analysis of Lower and Raised Pitch Voice Signal and Its Efficiency Calculation. Traitement du Signal. 2019 Oct 1;36(5):455-61.

[11] Singh MK, Singh AK, Singh N. Acoustic comparison of electronics disguised voice using different semitones. Int J Eng Technol (UAE). 2018;7(2):98.

[12] Singh MK, Singh N, Singh AK. Speaker's voice characteristics and similarity measurement using Euclidean distances. In2019 International Conference on Signal Processing and Communication (ICSC) 2019 Mar 7 (pp. 317-322). IEEE.

[13] Satya PM, Jagadish S, Satyanarayana V, Singh MK. Stripe Noise Removal from Remote Sensing Images. In2021 6th International Conference on Signal Processing, Computing and Control (ISPCC) 2021 Oct 7 (pp. 233-236). IEEE.

[14] Nandini A, Kumar RA, Singh MK. Circuits Based on the Memristor for Fundamental Operations. In2021 6th International Conference on Signal Processing, Computing and Control (ISPCC) 2021 Oct 7 (pp. 251-255). IEEE.

[15] Anushka RL, Jagadish S, Satyanarayana V, Singh MK. Lens less Cameras for Face Detection and Verification. In2021 6th International Conference on Signal Processing, Computing and Control (ISPCC) 2021 Oct 7 (pp. 242-246). IEEE.

[16] Priya, B.J., Kunda, P. and Kumar, S., 2021. Design and Implementation of Smart Real-Time Billing, GSM, and GPS-Based Theft Monitoring and Accident Notification Systems. In *Proceedings of International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications* (pp. 647-661). Springer, Singapore.

[17] Kanchana V, Nath S, Singh MK. A study of internet of things oriented smart medical systems. Materials Today: Proceedings. 2022 Jan 1; 51:961-4.

[18] Balaji VN, Srinivas PB, Singh MK. Neuromorphic advancements architecture design and its implementations technique. Materials Today: Proceedings. 2022 Jan 1; 51:850-3.

[19] Ramya, K., Bolisetti, V., Nandan, D. and Kumar, S., 2021. Compressive Sensing and Contourlet Transform Applications in Speech Signal. In *ICCCE 2020* (pp. 833-842). Springer, Singapore.

[20] Punyavathi G, Neeladri M, Singh MK. Vehicle tracking and detection techniques using IoT. Materials Today: Proceedings. 2022 Jan 1; 51:909-13.
[21] Karri, K.P., Anil Kumar, R. and Kumar, S., 2021. Multi-point Data Transmission and Control-Data Separation in Ultra-Dense Cellular Networks. In *ICCCE 2020* (pp. 853-859). Springer, Singapore.
[22] Veerendra G, Swaroop R, Dattu DS, Jyothi CA, Singh MK. Detecting plant Diseases, quantifying and classifying digital image processing techniques. Materials Today: Proceedings. 2022 Jan 1; 51:837-41.
[23] Padma U, Jagadish S, Singh MK. Recognition of plant's leaf infection by image processing approach. Materials Today: Proceedings. 2022 Jan 1; 51:914-7.
[24] Elshamy S, Fingscheidt T. Improvement of Speech Residuals for Speech Enhancement. In2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) 2019 Oct 20 (pp. 219-223). IEEE.