PROCEEDINGS OF THE YEREVAN STATE UNIVERSITY

Physical and Mathematical Sciences 2024, **58**(1), p. 22–29

Informatics

ADAPTIVE NOISE CANCELLATION FOR ROBUST SPEECH RECOGNITION IN NOISY ENVIRONMENTS

D. S. KARAMYAN [1,2*]

[1] *Russian-Armenian University, Armenia*
[2] *Krisp.ai, Armenia*

In this paper, we address the challenges faced when combining noise cancellation and automatic speech recognition (ASR) models. When these models are combined directly, the performance of word recognition often suffers, because the distribution of input data changes. To overcome this limitation, we propose a novel method for combining these models, which enhances the ability of the speech recognition model to perform well in noisy environments.

The key feature of the proposed method is the introduction of a mechanism to control the aggressiveness of noise reduction. This mechanism enables us to customize the noise reduction process according to the specific requirements of the ASR model, without necessitating any retraining. This advantage makes our method applicable to any ASR model, facilitating its implementation in practical scenarios.

https://doi.org/10.46991/PYSU:A.2024.58.1.022

MSC2010: 68T10.

Keywords: automatic speech recognition, noise cancellation, noise robustness, domain adaptation.

Introduction. Advancements in big data and computing power made it possible effective using of the automatic speech recognition (ASR) technology in various applications. However, ensuring noise robustness in these applications is a challenging task, as they need to function effectively in different acoustic environments. Even though deep neural networks have achieved high accuracy in large-vocabulary speech recognition [1–3], they require a significant amount of text-audio paired data, which is time-consuming and expensive to collect. Researchers have explored various approaches to improve noise robustness, including algorithms in the feature domain [4–7] or in the model itself [8–10]. Another popular method is multi-condition training [11], where the acoustic model is trained using noisy speech data.

* E-mail: dkaramyan@krisp.ai

In this work, we explore how to combine noise cancellation (NC) and speech recognition systems. It turns out that chaining these systems directly harms the model performance. This occurs because the noise cancellation forcibly removes noisy segments, making a shift in the input data distribution, which adversely affects to the speech recognition performance.

To address this issue, we propose a new method called *Weakly Noise Cancellation* that softens the effect of noise reduction without requiring retraining ASR model. The proposed method was able to improve the accuracy of speech recognition compared to the baseline ASR model obtained with multi-condition training [11].

Additionally, we found that augmenting the training process with noise cancellation, further improves word recognition accuracy. For example, when there is 0 *DB* of noise, meaning that the noise level is the same as the actual voice, we aim to decrease the word error rate (WER) by 1.2% compared to the baseline model.

Noise Injection. Adding background noise ($N(t)$) to a speech signal ($S(t)$) involves mixing noise and speech signals in the time domain based on a specified signal-to-noise ratio (SNR), which is defined as the ratio of the power of a speech signal to the power of background noise:

$$\text{SNR} = \frac{E[S(t)]}{E[N(t)]}, \qquad E[X(t)] = \frac{1}{N}\sum_{t=1}^{N} X(t)^2, \tag{1}$$

where E is the power of the signal and N is the length of the signal. Typically, it is expressed in decibels: $\text{SNR}_{DB} = 10\log_{10}(\text{SNR})$.

Assuming that both the speech and noise signals have the same amount of power (if not, we can simply scale the noise signal by a factor of $E[S(t)]/E[N(t)]$), we can represent the mixed signal $Y(t)$ as follows:

$$Y(t) = S(t) + \gamma N(t), \tag{2}$$

where $\gamma = 10^{-\frac{\text{SNR}_{DB}}{20}}$. Higher positive SNR_{DB} values indicate better speech signal quality with less audible noise, while negative SNR_{DB} values imply a higher noise level and worse speech signal quality, with the speech potentially being obscured by the noise.

Ideal Ratio Mask. The goal of noise cancellation is to reconstruct or estimate the original speech signal $S(t)$ from the mixed signal $Y(t)$ as accurately as possible. This is achieved by minimizing the error between the clean speech and the estimated target speech signal.

In the frequency domain, Eq. 2 can be expressed as:

$$Y(t,f) = S(t,f) + \gamma N(t,f) = |Y| \cdot e^{i\theta_Y}. \tag{3}$$

Here, $Y(t,f)$, $S(t,f)$ and $N(t,f)$ represent the spectra of the mixed, speech, and noise signals at a given frame t and frequency f, respectively. $|Y|$ denotes the magnitude of the mixed signal, while θ_Y represents its phase. For simplicity, the time-frequency (T-F) indexes (t,f) will be omitted from now on.

Next, let's define the ground truth ratio mask (M) as the ratio of the magnitudes of the speech signal ($|S|$) and the mixed signal ($|Y|$):

$$M = \frac{|S|}{|Y|} = \sqrt{\frac{|S|^2}{|S|^2 + \gamma^2|N|^2 + \gamma SN^* + \gamma S^*N}}, \tag{4}$$

where $\{\}^*$ is the conjugate transpose operator. Speech and noise are generally assumed statistically independent so that expectations of the last two terms in denominator are zero, although this assumption does not hold in real-world environments. The relaxed version of the ratio mask is known as the Ideal Ratio Mask (IRM) [12, 13]:

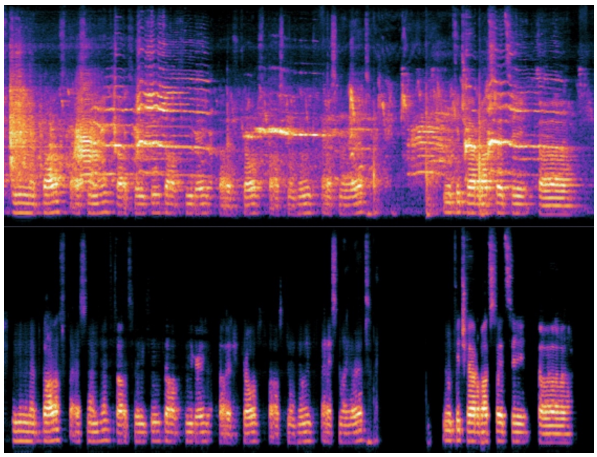
$$\text{IRM} = \sqrt{\frac{|S|^2}{|S|^2 + \gamma^2|N|^2}}. \tag{5}$$

Finally, by estimating the IRM one can estimate the speech component as follows:

$$\hat{S}(t,f) = \hat{M} \cdot |Y| \cdot e^{i\theta_Y}. \tag{6}$$

Here, $\hat{M}$ represents the estimated ratio mask, which is typically modelled using Deep Neural Networks. It is important to note that only the magnitude part of the mixed signal is transformed, while the phase remains unchanged.

Weakly Noise Cancellation. Applying noise cancellation before speech recognition may negatively impact the accuracy of word recognition, even if the audio remains audible to humans. This is because noise cancellation eliminates noisy segments forcefully, causing a shift in the input data distribution that ultimately affects speech recognition (see Figure). In this section, we present a new method for controlling the level of noise reduction.



The top image represents the audio spectrogram before noise cancellation and the bottom one displays the audio spectrogram after noise cancellation.

For estimated ratio mask $\hat{M}$, we construct a new adjusted ratio mask by applying the following transformation:

$$\hat{M}(\alpha) = \alpha + (1-\alpha)\hat{M}, \tag{7}$$

where α takes values from $[0,1]$. This provides us with a new estimation of the speech signal:

$$\begin{aligned}\hat{S}(t,f)(\alpha) = \hat{M}(\alpha)\cdot|Y|\cdot e^{i\theta_Y} = \\ \hat{M}(\alpha)Y(t,f) = \\ \alpha Y(t,f) + (1-\alpha)\hat{M}Y(t,f) = \\ \alpha Y(t,f) + (1-\alpha)\hat{S}(t,f) = \\ \alpha S(t,f) + \alpha\gamma N(t,f) + \hat{S}(t,f) - \alpha\hat{S}(t,f) = \\ \hat{S}(t,f) + \alpha\gamma N(t,f) + \alpha(S(t,f) - \hat{S}(t,f)).\end{aligned} \tag{8}$$

As we can see when $\alpha = 0$, we achieve the maximum possible noise reduction, and when $\alpha = 1$, we do not achieve any noise reduction at all. The term $\alpha\gamma N(t,f)$ in the last row of Eq. (8) is responsible for controlling the level of noise reduction. Moreover, if we express α as $10^{-\frac{d}{20}}$, with d ranging from zero to infinity, the parameter d takes on a physical meaning, indicating the desired noise level to keep in decibels.

Experiments and Results.

ASR Model. In all of our experiments, for an ASR model we use Conformer-CTC architecture [2] which effectively combines convolutional and transformer blocks to model both local and global dependencies of an audio sequence. We use a medium-size pre-trained Conformer checkpoint (https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_en_conformer_ctc_medium) that was made available by Nvidia. The model generates a probability distribution across subword units with a vocabulary size of 128. During inference, we apply a beam search with a width of 8 and do not use an external language model.

NC Model. For noise cancellation model we use a stack of 4 GRU layers [14], incorporating layer normalization between each layer. The final dense layer with sigmoid activation is responsible for generating ratio mask. This mask is then applied to the input spectrogram and transformed into a waveform using the Inverse Fourier Transform. The model was trained on the LibriSpeech dataset [15], while the noise data used was taken from the MUSAN dataset [16].

Test Dataset and Evaluation Metric. To evaluate the performance of the speech recognition, we have gathered an audio dataset consisting of 10.5 h of recordings. These recordings feature conversational content and involve multiple speakers, with 2 to 7 speakers per recording. To evaluate the system's performance in noisy environments, we also created augmented versions of the dataset by introducing varying levels of background noise ($0DB, 5DB$, and $10DB$). The results of our experiments are presented in terms of Word Error Rate (WER), which represents the percentage of words that were inaccurately predicted. All the relevant hyperparameters have been carefully adjusted using this dataset, and the optimal values are presented in Table.

Results. Augmenting training dataset with noise has been recognized as a simple yet effective method for enhancing noise-robustness in speech recognition models [11]. We consider this approach to be a solid baseline for our work. We fine-tune the ASR model on an in-house dataset with around 75000 h of English speech.

During the training process, we introduce various levels of noise to the speech signals, spanning from 0 to 50 *DB* signal-to-noise ratios. The first row in Table (labeled *Multi-Condition training*) showcases the baseline WER results of the model trained with noise augmentation.

WER results of differnet methods

	Orig	0*DB*	5*DB*	10*DB*
Multi-Condition training (Baseline [11])	8.78	17.1	12.07	10.26
Full Noise Cancellation	8.85	19.97	13.41	10.82
Power of Ratio Mask ($\hat{M}^{0.8}$)	8.8	19.08	12.96	10.65
Weakly Noise Cancellation (α=0.6)	8.75	16.53	11.83	10.17
Weakly Noise Cancellation (α=0.3) + NC-Augment	9.07	15.9	11.77	10.24
Teacher-Student (Amazon [8])	9.09	14.78	11.35	10.1

Next, we implemented a cascaded approach, where we first applied noise cancellation before feeding the signals into the speech recognition system. The second row in Table illustrates that direct chaining of noise cancellation and speech recognition models did not yield improvements. In fact, the performance was significantly worse compared to the baseline.

To overcome this issue, we applied proposed *Weakly Noise Cancellation* method to reduce the aggressiveness of the noise cancellation model before providing the audio inputs to the ASR model. This approach produces better results compared to the *Full Noise Cancellation*, and it also improves the model's performance compared to the multi-condition trained baseline.

Additionally, we experimented with an exponential ratio mask transformation, which involves raising the ratio mask $\hat{M}$ to a power β. By choosing a value of β less than 1, we can reduce the aggressiveness of the noise cancellation. However, our empirical findings showed that the results were much worse compared to the proposed approach.

But can we further improve? By incorporating NC-like augmentation (see the 5th row in Table) in the training process and applying the *Weakly Noise Cancellation* method, as described earlier, we were able to achieve even better results. In the 0 *DB* scenario, this approach resulted in a significant 1.2% absolute reduction in WER compared to the baseline.

Lastly, we also reproduced the results of the study by Movsner et al. [8], where the authors adopt the Teacher-Student learning technique using a parallel clean and noisy corpus for improving ASR performance under multimedia noise. In the proposed approach, clean and noisy audios were fed to the Teacher and the Student models, respectively, to enforce similarity between the output distributions. On top of that, they apply a logits selection method, which only preserves the $k = 20$ highest values to prevent wrong emphasis of knowledge from the Teacher and to reduce bandwidth needed for transferring data. As a Teacher model, we use the one obtained via a multi-condition training baseline. This model processes clean audio to generate logits,

guiding the training of the Student model on noisy inputs. We fine-tuned the Student model on an in-house dataset with around 75000 *h* of noisy speech data. As shown in the Table, the Teacher-Student method outperforms all previous methods, but it requires retraining the ASR model, which may not always be possible.

Adapting models to perform better in noisy environments (the last two rows in Table) can sometimes lead to overfitting to noise characteristics, which affects the model's ability to accurately recognize speech in clean audio conditions. Although these strategies significantly improve the model's performance in challenging noisy scenarios, they can introduce a slight degradation in performance on clean audio due to the model's increased sensitivity to noise characteristics rather than focusing solely on the speech signal.

Conclusion. In conclusion, this paper presents a novel solution, Weakly Noise Cancellation, to address the challenge of integrating noise cancellation with speech recognition models. By introducing a parameter for the controlled noise reduction, we were able to enhance the model's performance in noisy environments compared to the baseline model trained with the noise augmentation. Additionally, we demonstrate that even better results can be achieved by incorporating NC-based augmentation during the training phase.

In future studies, we aim to further develop the concept of Weakly Noise Cancellation by making the α hyperparameter trainable. Additionally, we intend to explore the possibility of having a separate reduction parameter for each time-frequency index, making the method more flexible and adaptable.

Received 28.12.2023
Reviewed 19.02.2024
Accepted 28.02.2024

REFERENCES

1. Radford A., Kim J., et al. Robust Speech Recognition Via Large-scale Weak Supervision. *International Conference on Machine Learning* (2023), 28492–28518. https://doi.org/10.48550/arXiv.2212.04356
2. Gulati A., Qin J., et al. Conformer: Convolution-augmented Transformer for Speech Recognition. *Electrical Engineering and Systems Science* (2020), 5036–5040. https://doi.org/10.48550/arXiv.2005.08100
3. Li J., Lavrukhin V., et al. Jasper: An End-to-End Convolutional Neural Acoustic Model. *Electrical Engineering and Systems Science* (2019). https://doi.org/10.48550/arXiv.1904.03288
4. Boll S. Suppression of Acoustic Noise in Speech Using Spectral Subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **27** (1979), 113–120. https://doi.org/10.1109/TASSP.1979.1163209
5. Acero A. *Acoustical and Environmental Robustness in Automatic Speech Recognition.* Springer Science and Business Media (1992). https://doi.org/10.1007/978-1-4615-3122-7

6. Cui X., Iseli M., et al. Evaluation of Noise Robust Features on the Aurora Databases. *Proc. 7th International Conference on Spoken Language Processing (ICSLP 2002)* (2002), 481–484.
https://doi.org/10.21437/ICSLP.2002-24
7. Hermansky H., Morgan N. RASTA Processing of Speech. *IEEE Transactions on Speech and Audio Processing* **2** (1994), 578–589.
https://doi.org/10.1109/89.326616
8. Mošner L., Wu M., et al. Improving Noise Robustness of Automatic Speech Recognition Via Parallel Data and Teacher-Student Learning. *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2019), 6475–6479.
https://doi.org/10.48550/arXiv.1901.02348
9. Gales M., Young S. Robust Continuous Speech Recognition Using Parallel Model Combination. *IEEE Transactions on Speech and Audio Processing* **4** (1996), 352–359.
https://doi.org/10.1109/89.536929
10. Gong Y. Speech Recognition in Noisy Environments: A Survey. *Speech Communication* **16** (1995), 261–291.
https://doi.org/10.1016/0167-6393(94)00059-J
11. Lippmann R., Martin E., Paul D. Multi-style Training for Robust Isolated-word Speech Recognition. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* **12** (1987), 705–708.
https://doi.org/10.1109/ICASSP.1987.1169544
12. Wang Z., Wang X., et al. Oracle Performance Investigation of the Ideal Masks. *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)* (2016), 1–5.
https://doi.org/10.1109/IWAENC.2016.7602888
13. Xia S., Li H., Zhang X. *Using Optimal Ratio Mask as Training Target for Supervised Speech Separation. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. Malaysia, Kuala Lumpur, IEEE (2017).
https://doi.org/10.48550/arXiv.1709.00917
14. Cho K., Merriënboer B., et al. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation* **10** (2014), 103–111.
https://doi.org/10.3115/v1/W14-4012
15. Panayotov V., Chen G., et al. Librispeech: An ASR Corpus Based on Public Domain Audio Books. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2015), 5206–5210.
https://doi.org/10.1109/ICASSP.2015.7178964
16. Snyder D., Chen G., Povey D. Musan: A Music, Speech, and Noise Corpus (2015).
https://doi.org/10.48550/arXiv.1510.08484

Դ. Ս. ՔԱՐԱՄՅԱՆ

ԱԴԱՊՏԻՎ ԱՂՄՈՒԿԻ ՀԵՌԱՑՈՒՄ ԱՂՄԿՈՑ ՄԻՋԱՎԱՅՐՈՒՄ ԽՈՍՔԻ ՀՈՒՍԱԼԻ ՃԱՆԱՉՄԱՆ ՀԱՄԱՐ

Այս հոդվածում անդրադարձ է կատարվել այն խնդիրներին, որոնք առաջանում են աղմուկի հեռացման և խոսքի ճանաչման մոդելները համապատեղելու

ժամանակ: Երբ այս մոդելները ուղղակիորեն համակցվում են, բառերի ճանաչման ճշգրտությունը հաճախ փուժում է, քանի որ մուտքային փվյալների բաշխումը փոխվում է: Այս սահմանափակումը հաղթահարելու համար առաջարկվել է համակցման նոր մեթոդ, որը լավացնում է խոսքի ճանաչման մոդելի ճշգրտությունը աղմուկի պայմաններում:

Առաջարկվող մեթոդի հիմնական առանձնահատկությունը աղմուկի հեռացման մոդելի ագրեսիվությունը վերահսկելու մեխանիզմի ներդրումն է: Այս մեխանիզմը հնարավորություն է փալիս հարմարեցնել աղմուկի հեռացման գործընթացը՝ համաձայն ASR մոդելի պահանջների՝ առանց որևէ վերաուսուցման անհրաժեշտության:

Д. С. КАРАМЯН

АДАПТИВНОЕ ШУМОПОДАВЛЕНИЕ ДЛЯ НАДЕЖНОГО РАСПОЗНАВАНИЯ РЕЧИ В УСЛОВИЯХ ШУМА

В данной статье рассматриваются проблемы, которые появляются при объединении моделей шумоподавления и автоматического распознавания речи (АРР). Когда эти модели объединяются напрямую, производительность распознавания слов часто страдает из-за изменения распределения входных данных. Чтобы преодолеть это ограничение, в данной статье рассматривается новый метод объединения этих моделей, который повышает способность модели распознавания речи хорошо работать в шумной среде.

Ключевой особенностью предлагаемого метода является введение механизма управления агрессивностью шумоподавления. Этот механизм позволяет настроить процесс снижения шума в соответствии с конкретными требованиями модели АРР без необходимости какого-либо переобучения. Это преимущество делает данный метод применимым к любой модели АРР, облегчая его реализацию в практических сценариях.