Improving speakers separation with additional speech recognition task

Legchenko Anton
Novosibirsk State University
a. legchenko@g. nsu. ru

Pavlovskiy Evgeniy
Novosibirsk State University
e. pavlovskiy@g. nsu. ru

Abstract

In this work, we propose a method of multi-task learning in a cocktail-party problem with a single microphone, as an additional task, we selected the task of automatic speech recognition for each of the speakers in the audio signal. The paper presents two different approaches to the hierarchy of tasks in which speech recognition is performed after the separation of speech signals or simultaneously with it. For training and testing of models we used open source resources and public datasets.

The proposed method showed an improvement in results compared to the single-task model, giving an increase of about 1% according to various estimates of the quality of speaker separation. To confirm the results, a hierarchical model was fine-tunned using only the loss function of the speech recognition task, which also gave an improvement in the quality of speaker separation by 0.4%.

Keywords: Automatic speech recognition, cocktail-party problem, deep learning, multi-task learning
DOI: 10.28995/2075-7182-2022-20-XX-XX

Форматирование докладов на ДИАЛОГ-2022

Легченко Антон
Новосибирский
Государственный Университет
a.legchenko@g.nsu.ru

Евгений Павловский
Новосибирский
Государственный Университет
e.pavlovskiy@g.nsu.ru

Аннотация

В данной работе мы предлагаем метод многозадачного обучения в задачи коктейльной вечеринки с единственным микрофном, в качестве дополнительной задачи была выбрана задача автоматического распознавания речи для каждого из дикторов в звуковом сигнале. В работе предсталвены два различных подхода к иерархии задач, в которых распознавание речи производиться после разделения речевых сигналов или же одновременно с ним. Обучение и тестирование моделей проводилось на открытых ресурсах и наборах данных.

Предложенный метод показал улучшение результатов по сравнению с однозадачной моделью, дав прирост около 1% по различным оценкам качества разделения дикторов. Для подтверждения результатов был произведено дообучение иерархической модели используя только функцию потерь задачи распознавания речи, что дало также дало улучшение качества разделения дикторов на 0.4%.

Ключевые слова: Автоматическое распознавание речи, сверточные сети, глубокое обучение

1 Introduction

Automatic speech recognition (ASR) is one of the important problems in the modern world, there are many solutions that allow it to be solved with high accuracy for speech signals of good quality, but in the real world, speech recognition systems face contamination of the speech signal by different noise, especially difficult is the case of the presence of voices of extraneous speakers. To solve this problem, are used systems that implement the separation of the signal into speech signals by belonging to different speakers. In cases of a consistent speech, this problem is well studied and is known as the speaker diarization problem, in cases of significant overlap of the speakers's speech signals, the task becomes much more difficult and is known as the cocktail party problem. To solve the problem of a cocktail party with one microphone, various approaches were used, most modern solutions are based on deep learning, this approach allows you to achieve high results, but is associated with problems such as retraining, which are especially acute with a small amount and variety of training data, which are often unavailable. One of the methods to improve the generalizing ability of deep learning models is the multitasking learning method, when one model is trained to solve various problems from the same subject area, so that the model can use the similarities and differences between them. In many cases, the method of multitasking training allows you to achieve better results of model predictions from specific tasks, compared with models of students on one specific task. In this work, we propose a method of

multi-task learning in a cocktail-party problem with a single microphone, as an additional task, we selected the task of automatic speech recognition for each of the speakers in the audio signal, in two different formulations, when speech recognition is carried out on the basis of signals predicted by the separation model and when the transcription of the speech of individual speakers is predicted simultaneously with the prediction of the separated speech signals of speakers.
The main goal of the work was to implement the proposed model and compare the results to show the advantages of the multitasking approach in speaker separation task with mono-channel signals.

1 Literature review

Currently, many end-to-end models have been proposed to solve the cocktail party problem, the most interesting in this work are models using multitasking learning with the solution of speech separation and recognition problems. For a problem with multiple microphones in 2019, a MIMO-Speech model [1] was proposed that implements a hierarchical approach to the problem when the speech recognition problem is solved after the speaker separation problem. For a more complex task with a single microphone, in 2018, a model [2] was proposed that does not use multitasking training, but solves the problem of speech recognition of multiple speakers in an end-to-end style. Unlike MIMO-Speech, which implements a hierarchical approach to multitasking learning, an alternative is to use a single block of feature extraction for heads solving various equal tasks, an example of such an approach is Multi-task BERT model from [3] for the NLU task. Thus, work [2] shows the possibility of considering the recognition problem as one of the independent tasks and using an additional head to the block of extracting the features of the original problem to solve it. Models [1] and [2] are based on the Transformer architecture, which makes it difficult to train them in conditions of limited computing resources, and therefore attention was drawn to earlier and lighter models that solve individual problems.

To solve the separation problem in 2017, the ConvTasNet [4] model based on the CNN architecture was proposed, due to which its training time is significantly shorter, while the model shows good quality scores.

Modern approaches to the task of automatic speech recognition are also based on the Transformer architecture, for example Wav2Vec2.0 [5], within the framework of this work, their use is difficult and a model based on the CNN Wav2Latter [6] architecture was also chosen as a speech recognition model.

In the work we also used the PIT [7] method, which has become standard for solving the separation problem and significantly improves the convergence of model used in [4].

For the training separation models there are many synthetic data sets based on existing speech corpora, similar in principle to generation. The largest of the public datasets is Librimix [8], based on the English-language speech corpus of recordings from LibriSpeech [9] that contains high quality speech examples from audio books.

2 Method

The paper compares three different composite models with the same architecture of the part solving the problem of separating speech signals and representing ConvTasNet with a reduced size of up to 21 million parameters to reduce training time. The differences of the models are in the part of the model solving the problem of speech recognition.

The hierarchical model is a serial connection of ConvTasNet and Wav2Letter models adapted to the sampling rate of 8khz.

The model with a common feature space is a ConvTasNet with an additional decoder block implementing end-to-end speech recognition.

The basic solution is the usual component model with ConvTasNet and Wav2Letter as in hierarchical model.

All models were implemented in the pytorch framework.

4 Experiments and results

The experiments are done on a synthesized data set based on the LibriSpeech public corpus containing audio recordings of English speech with a low noise level, part of train-clean-100 was used for training and validation and test-dev-clean was used for the test. The synthesis of examples for the data set was similar to LibriMix , to reduce the learning time, speech examples were limited in duration to four seconds, in this regard, the markup of the speech recognition task for each individual speaker's speech example was carried out synthetically using a modern Wav2Vec 2.0 model with high accuracy on LibriSpeech, WER 1.7%, which significantly exceeds the possible quality indicators of the models considered in this paper. in total, 27,000 examples were synthesized for training and 4,000 examples each for testing and validation.
When training models, we used the gradient clipping technique, the Knowledge Distillation method [10] was used for the speech recognition unit, which allows to improve the convergence of the model, the Wav2Vec 2.0 model was used as a teacher. PIT training method was used for the separation unit. One model was trained for one day on one Nvidia Tesla V-100 with batch size 32 for 100 epochs. In the experiments, model indicators were considered for different values of the multiplier k with an additional loss function of the transcription task, Table 1 shows PESQ indicators for different values of the parameter k.

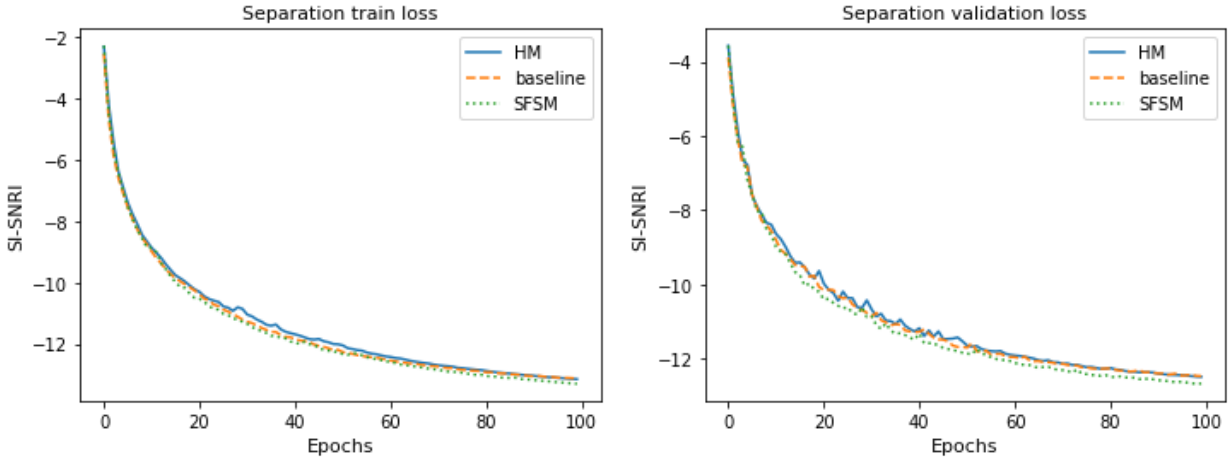
Table 1: PESQ score of separation block

Model	k = 10	k = 1	k = 0.1
HM	2.12	2.31	2.39
SFSM	2.02	2.29	2.38

The results obtained show that in both approaches, the best separation result corresponds to a lower coefficient k for the loss function of the ASR task.

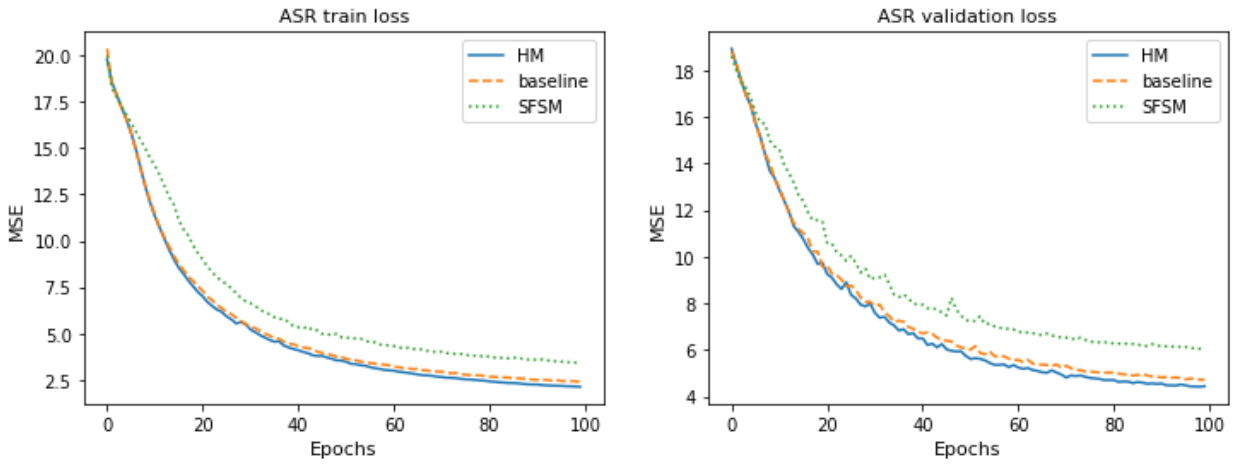
4.1 Base experimnet

Figure 1 shows the convergence graphs of the separation unit, Figure 2 shows the convergence graphs of the speech recognition unit, both figures show models with the optimal parameter.



(a) Training loss (b) Validation loss

Figure 1: The separation block training curves



(a) Training loss (b) Validation loss

Figure 2: The transcription block training curve

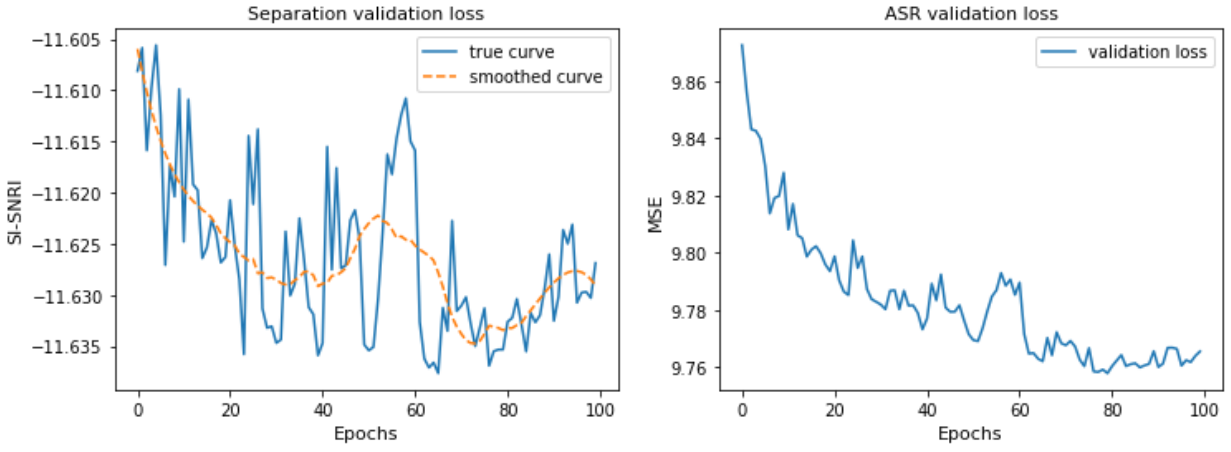
Table 2 shows the various metrics of the trained models on the test dataset. The results show that both multitasking methods yielded results superior to the base solution in terms of separation quality, and the hierarchical model outperformed the component system in terms of CER for the ASR problem as well.

Table 2: Architecture comparison on test dataset

Model	PESQ	SI-SDR	CER
baseline	2.37	11.11	0.42
HM	2.39	11.33	0.43
SFSM	2.38	11.16	0.49

4.2 Fine-tuning experimnet

To confirm the relationship of improved results with the use of an additional loss function, an experiment was conducted on tuning the base solution using only the loss function of the speech recognition problem. The learning process is shown in Graph 3, the metrics obtained before and after the fine-tuning are shown in Table 3.



(a) Separation task loss on validation set (b) ASR task loss on validation set

Figure 3: Tuning separation model only with ASR task loss

Table 3: ASR task fine-tuning results on test

Model	PESQ	SI-SDR	CER
Model before tuning	2.33	10.73	0.44
Model after tuning	2.34	10.81	0.43

The loss function curve for the separation unit was unstable during the training process, but the final result showed improvement in the metrics, which confirms the feasibility of the hierarchical multitasking approach.

5 Discussion

Both of the proposed multitasking methods showed an increase in results when the relationship between the loss functions was correctly chosen, with the model with a single feature space showing markedly higher quality metrics on the validation sample containing the voices present in the training sample, but inferior to the hierarchical model on the test sample, which may indicate an increased propensity for this model to over-learn by more accurate learning of the speech features of individual speakers. Unfortunately, the increase was not so large, only about 1%. Perhaps great results could be achieved using models with a large number of parameters and a more advanced architecture, as in MIMO-Speech. Since both approaches gave improved results, it was also worth considering the third statement of the problem, combining both approaches to the architecture, to consider the first transcription output as a preliminary prediction, as for example in the Inception networks [x].

6 Challenges and unrealized plans

During the research, the main problem was the lack of computing resources, which led to a significant simplification of the architecture of models and datasets. This did not allow the use of models that solve individual problems well and significantly reduced the possible number of experiments.

7 Conclusion

In this study, it was shown the possibility of improving the quality of speaker separation using the additional task of transcription of the received signals, as well as the possibility of tuning already trained models. Based on the results obtained, it is necessary to test this method for modern architectures, as well as with the addition of more tasks, such as classifying speakers, then the method may possibly give a more significant increase in metrics.

Another important task for subsequent work remains the study of changes in the hidden space of models introduced by the multitasking learning method, establishing a connection between various speech processing tasks.

References

[1] Xuankai Chang , Wangyou Zhang, et al. MIMO-SPEECH: End-to-End Multi-Channel Multi-Speaker Speech Recognition. // arXiv:1910.06522. Access mode: http://arxiv.org/abs/1910.06522.

[2] Xuankai Chang , Wangyou Zhang, et al. End-to-End Multi-speaker Speech Recognition with Transformer. // arXiv:2002.03921. Access mode: http://arxiv.org/abs/2002.03921.

[3] Xiaodong Liu, Pengcheng He, et al. Multi-Task Deep Neural Networks for Natural Language Understanding. // arXiv:1901.11504. Access mode: http://arxiv.org/abs/1901.11504.

[4] Yi Luo, Nima Mesgarani. Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation. // arXiv: 1809.07454. Access mode: http://arxiv.org/abs/1809.07454.

[5] Alexei Baevski, Henry Zhou, et al. Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation. // arXiv: 2006.11477. Access mode: http://arxiv.org/abs/2006.11477.

[6] Ronan Collobert, Christian Puhrsch, et al. Wav2Letter: an End-to-End ConvNet-based Speech Recognition System. // arXiv: 1609.03193. Access mode: http://arxiv.org/abs/1609.03193.

[7] Dong Yu, Morten Kolbæk, et al. Permutation Invariant Training of Deep Models for Speaker-Independent Multi-talker Speech Separation. // arXiv: 1607.00325. Access mode: http://arxiv.org/abs/1607.00325.
[8] Joris Cosentino, Manuel Pariente, et al. LibriMix: An Open-Source Dataset for Generalizable Speech Separation. // arXiv: 2005.11262. Access mode: http://arxiv.org/abs/2005.11262.
[9] Vassil Panayotov; Guoguo Chen, et al. Librispeech: An ASR corpus based on public domain audio books. // Publisher: IEEE. Access mode: https://ieeexplore.ieee.org/document/7178964.