Research article
DOI: https://doi.org/10.18721/JCSTCS.15406
UDC 004.8

MULTI-CHANNEL TRANSFORMER: A TRANSFORMER-BASED MODEL FOR MULTI-SPEAKER SPEECH RECOGNITION

E.S. Fadeeva[1] , *V.A. Ershov*[2] ✉

[1,2] Yandex LLC, St. Petersburg, Russian Federation

✉ noxoomo@yandex-team.ru

Abstract. Most of the modern approaches to multi-speaker speech recognition are either not applicable in case of overlapping speech or require a lot of time to run, which can be critical, for example, in case of real-time speech recognition. In this paper, a transformer-based end-to-end model for overlapping speech recognition is presented. It is implemented by using a generalization of the standard approach to speech recognition. The introduced model achieves results comparable in quality to modern state-of-the-art models, but requires less model calls, which speeds up the inference. In addition, a procedure for generating synthetic data for model training is described. This procedure allows to compensate for the lack of real multi-speaker speech training data by creating a stream of data from the initial collection.

Keywords: speech recognition, multi-speaker speech recognition, diarization, speech separation, voice technologies

Citation: Fadeeva E.S., Ershov V.A. Multi-channel transformer: A transformer-based model for multi-speaker speech recognition. Computing, Telecommunications and Control, 2022, Vol. 15, No. 4, Pp. 73–85. DOI: 10.18721/JCSTCS.15406

Научная статья
DOI: https://doi.org/10.18721/JCSTCS.15406
УДК 004.8

МНОГОКАНАЛЬНЫЙ ТРАНСФОРМЕР: МОДЕЛЬ ДЛЯ РАСПОЗНАВАНИЯ МНОГОГОЛОСНОЙ РЕЧИ, ОСНОВАННАЯ НА АРХИТЕКТУРЕ ТРАНСФОРМЕР

Е.С. Фадеева[1], *В.А. Ершов*[2] ✉

[1,2] ООО «Яндекс», Санкт-Петербург, Российская Федерация

✉ noxoomo@yandex-team.ru

Аннотация. Многие современные подходы для решения задачи распознавания многоголосной речи либо не предназначены для работы с пересекающейся речью, либо требуют много времени для запуска, что может быть критичным, например, в случае распознавания речи в реальном времени. В статье предложена трансформерная end-to-end модель для распознавания многоголосной речи с возможными пересечениями. Предложенная архитектура является обобщением архитектуры из стандартного подхода к распознаванию речи. Такая модель позволяет достичь результатов, сопоставимых по качеству с современными решениями, но требует меньше запусков модели для получения текстового распознавания многоголосной речи, что ускоряет время работы такой системы. Описана процедура генерации синтетических данных для обучения модели. Эта процедура позволяет компенсировать отсутствие реальных данных для обучения модели для распознавания многоголосной речи путем создания потока данных из первоначального набора.

Ключевые слова: распознавание речи, распознавание многоголосной речи, диаризация, разделение речи, голосовые технологии

Для цитирования: Fadeeva E.S., Ershov V.A. Multi-channel transformer: A transformer-based model for multi-speaker speech recognition // Computing, Telecommunications and Control. 2022. Т. 15, № 4. С. 73–85. DOI: 10.18721/JCSTCS.15406

Introduction

Speech recognition is a problem of determining a text spoken on an audio signal. Voice technologies are used in many aspects of human life: voice assistants, smartphones for blind people, transcribing voice messages.

Periodically, a situation of multi-speaker speech arises in speech technologies, that is, when several people are speaking on an audio recording. Solutions to the problem of voice recognition in such situations usually give out the whole speech in one text, without identifying the specific speakers of the recognized words. But if it is necessary to determine the spoken texts for each speaker separately, the problem of multi-speaker speech recognition arises.

Systems for automatic multi-speaker speech recognition can be used for mark-up and further analysis of various audio data, for example, dialogues at conferences or personal meetings.

In the general case of the problem of multi-speaker speech recognition, the number of people participating in the conversation is not known in advance, as well as any additional information about the speakers is unknown. Another difficulty of the problem is the periodically occurring situation of overlapping speech (when people interrupt each other).

Thus, in the most general formulation of the problem of multi-speaker speech recognition, an audio signal comes to the input of the system, where the speech of several people sounds (the number of which is unknown), and the system should output several voice-recognition texts of each of the participants in the conversation.

Diarization is a problem of dividing input audio recording into several segments with a designated unique number for each the speaker whose voice sounds throughout this segment. In situations of overlapping speech, these segments should overlap.

Multi-speaker speech recognition is closely related to the problem of diarization, as is its combination with a problem of conventional speech recognition in case of disjoint (non-overlapping) speech.

In this article, a model for multi-speaker speech recognition is proposed. We will release code to reproduce our experiments here: https://github.com/cant-access-rediska0123/multi-channel-tranformer. The presented approach works correctly if no more than two people participate in the conversation on the input audio recording, however, the architecture of the presented model can be generalised to the situation of any limited number of speakers. In all other aspects, the problem statement is the most general (including the input speech may overlap).

Existing solutions

Currently, there are several approaches to solving the problem of multi-speaker speech recognition. A detailed overview of the existing solutions of the diarization problem and approaches to multi-speaker speech recognition is presented in [1].

One of the standard problem statements in working with multi-speaker speech is audio mark-up using speaker identifiers. The first methods for multi-speaker speech recognition solve the problem of conventional speech recognition by predicting, together with parts of the text, the unique identifiers of the speakers who uttered this text (for example, [2–5]). Such approaches do not require the use of separate models for the multi-speaker speech recognition, so this problem boils down to determining speaker identifiers for different audio parts. This could be done using particular methods of diarization – determining who spoke on which segment of audio. Such methods do not usually consider situations with overlapping speech (as for example [2]).

Often, in order to solve the problem of overlapping speech, it is proposed to solve a more complex problem: to automatically divide the signal into several audio tracks (channels), each of which refers to a separate person. The signal is then recognized independently on each track using conventional speech recognition models. This approach is used, for example, in [6]. However, this approach has a number of shortcomings. Firstly, it is necessary to additionally solve the problem of signal separation, which may be more difficult than the original problem (it is shown that such a problem is difficult for human hearing, [7]). Secondly, this approach is limited in the ability to work on improving the quality of models by collecting training data. In the presence of a working neural network architecture, further and significant improvement in the quality of most machine learning systems is achieved by collecting and marking up new training data (including using crowd-sourcing, for example, [8, 9]). Finally, in modern machine learning, there is a trend towards integral end-to-end approaches that struggle with the problem of all cascading approaches, when errors of one model affect subsequent ones. For example, later solutions to the problems of diarization, speech recognition, and speech synthesis have an end-to-end architecture [5, 10, 11].

There are also similar modern approaches that solve the problem of overlapping speech in different ways [3–5]. Those approaches use modern state-of-the-art architectures like transformers to achieve best possible results, but recognize speech of different speakers consequently. This process requires

$$\sum_{i=1\ldots n} |h_i|$$

auto-regressive model calls to transcribe n text hypotheses $h_1, h_2, \ldots, h_n$. The runtime of the multi-speaker speech recognition system can be critical in many situations, for example in case of real-time multi-speaker speech recognition, when the model needs to be inferenced fast to transcribe texts with minimum possible delay.

In this article, an end-to-end model for multi-speaker speech recognition with possible overlaps is proposed. This model achieves results that can be compared with the results of modern state-of-the-art solutions, but it reduces the runtime by recognizing multiple speakers in parallel, so that only $\max\limits_{i=1\ldots n}|h_i|$ auto-regressive model calls are required to transcribe text hypotheses h_1, h_2, ..., h_n. The structure of the presented model is based on the assumption that no more than two people participate in the dialogue, but it can be generalised to the case of any limited number of participants in the dialogue.

The successfully released transformer architecture [12] is now used in the best models that solve a wide variety of machine learning problems, including speech-related problems. The most recent models often use this architecture [5, 13, 14]. This model structure looks promising in such problems, so the presented model is completely based on the transformer architecture.

Metrics

A common metric for evaluating speech recognition models is Word Error Rate (WER), which is defined as the ratio of the minimum number of inserts, deletions and substitutions of words necessary to convert the hypothesis of the model into the text spoken on the audio recording to the number of words in this text:

$$\mathrm{WER}(h,r)=\frac{I(h,r)+D(h,r)+S(h,r)}{|r|},$$

where h is the hypothesis of the model about the speech spoken on the audio recording; r is the target text (spoken on the audio recording); I, D, S are the numbers of inserts, deletions and substitutions of words, respectively.

When evaluating the models of multi-speaker speech recognition, the exact correspondence of the speech recognition texts to the spoken texts is unknown, since the order of speakers in the prediction of the model is unknown. Therefore, the most widely used metric is the Concatenated minimum-Permutation Word Error Rate (cpWER, [15]), which generalises the WER metric to the case of several speakers. CpWER is calculated using the following formula:

$$\mathrm{cpWER}(h,r)=\frac{\min\limits_{\pi\in\Pi}\left(I\left(h_{\pi_i},r_i\right)+D\left(h_{\pi_i},r_i\right)+S\left(h_{\pi_i},r_i\right)\right)}{\sum\limits_i |r_i|},$$

where h is several texts; the hypothesis of the model about the speech spoken by each of the participants of the conversation on the audio recording; r is the texts spoken by the speakers on the audio recording; n is the number of speakers on the audio recording; the number of texts in r and in h; $\Pi(n)$ is all possible permutations from n elements; π is the permutation that determines the correspondence of the texts from the hypothesis of the model h to the texts on the audio recording r.

At the same time, if the number of people in the prediction of the model does not match their number on the audio recording, the missing people are filled with empty texts.

In this paper, in order to assess the quality of the model, the cpWER metric is used for the case of two speakers. When calculating it, all (two) possible permutations of the target texts were iterated through.

In order to solve this problem, the transformer architecture is used. The model (Fig. 1) is constructed for the case of no more than two speakers on an audio recording, but it can be generalised to the case of any limited number of people. The model is constructed by analogy with the transformer model for speech recognition [13].

Model: multi-channel transformer

Standard methods are used for pre-processing of the input audio recordings and target texts. Similarly to [13], 80-dimensional mel-spectrograms are used to extract features from the audio recording.

In order to convert target texts, a 8000 tokens-size dictionary (sentencepiece, [16]) pre-trained on training data is used. Thus, one target text corresponding to one speaker is divided into several tokens from this dictionary, where the token is one or more letters in a row, which are often found in the training data in this combination.

The architecture of the multi-channel transformer consists of two parts – encoder and auto-regressive decoder. The encoder does not depend on the recognized texts in any way, so any encoder suitable for the speech recognition problem can be used in its place. The experiments used the encoder inspired by the work of [13], consisting of sequentially applied layers of two convolutions, Positional Encoding and ten Transformer Encoder blocks [12]. Additionally, experiments were carried out with the encoder from [17]. As experiments showed, for the stated problem, the most critical part of the model is the decoder and the network training method, and not the specific architecture of the encoder.

The decoder architecture has been modified for the case of multiple texts. In the presented model, two prefixes of prediction texts independently pass through the same Embedding and Positional Encoding blocks. To get rid of the extra dimension (two blocks of features from two speakers) a linear layer is used reducing the number of features by half. The final Transformer Decoder block predicts the distributions of the next tokens.

Instead of one distribution of the next token $p\left(x_t \middle| x_1,\ y_1,\ \ldots,\ x_{t-1},\ \ y_{t-1},\ e\right)$ as it would be in a conventional transformer, the decoder predicts two independent distributions of the following tokens for each speaker:

$$p\left(x_t \middle| x_1,\ y_1,\ \ldots,\ x_{t-1},\ \ y_{t-1},\ e\right),\ p\left(y_t \middle| x_1,\ y_1,\ \ldots,\ x_{t-1},\ \ y_{t-1},\ e\right),$$

where e is the output of the encoder encoding the audio recording; x, y is the recognition texts of two speakers. In order to ensure the same length of the sequences x and y, the smaller of them is padded with END tokens denoting the end of the sentence.

This paper does not consider the problem of mark-up of recognized text using speaker identifiers. Therefore, in the training data the first text corresponds to the text of the person who started speaking first, and the second text corresponds to the text of the remaining second person.

Experiments were conducted with two decoding variants. In the first variant, experiments had two decoders corresponding to each of the speakers on the audio recording and recognizing the two hypotheses independently. This variant showed results comparable to the results of the current variant. But due to the doubled number of decoders, the number of parameters in such a model increased by ~ 1.5 times, and both decoders had to learn the same information about the identity of the speakers. Therefore, this approach has been replaced with the second variant – with a single decoder that recognizes both texts together.

In total, the final model for recognizing two speakers turned out to have ~ 1.94×10^8 parameters. The single-channel transformer for conventional speech recognition, by analogy with which a multi-channel transformer was built, have approximately the same (~ 1.8×10^8) number of parameters.

Data

Data with speech recordings of one or two speakers were used to train the multi-channel transformer. In the absence of open-source datasets with short recordings of conversations between two people, data from the widespread LibriSpeech speech recognition dataset [18], which includes 960 hours of audiobook recordings, was taken to train the model. Parts of train-clean and train-other data are used in the process of training.

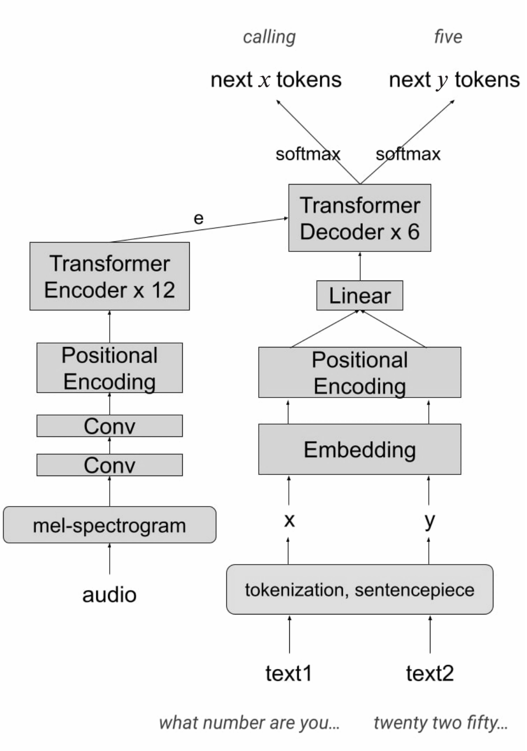


Fig. 1. Scheme of multi-channel transformer

There are no dialogues samples in the LibriSpeech data. Therefore, in addition to single-speaker original examples, synthetic data obtained by mixing pairs of audio recordings are used in training the model to recognize two people (Fig. 2). When mixing, both audio recordings are padded with silence to the same size, and then their average is taken. Pairs of audio recordings are selected so that they are spoken by different people, that is, the unique speaker identifiers given in LibriSpeech for each audio recording were different for these two samples.

When generating another synthetic example in the training process, first audio recording is uniformly selected from all the data (train-clean and train-other), then the second audio recording is uniformly selected among all, excluding audio recordings with the same speaker, then these two audio recordings are mixed with indentation. The indentation is selected randomly with a lower bound of 1 second in order to avoid the possibility of incorrectly determining the order of speakers. This streaming method of generating examples greatly increases the size of the training data.

At the training stage data is pre-processed using several augmentations: changing the audio speed with a coefficient chosen randomly from the interval [0.75, 1.25] and augmenting the spectrogram [19].

Similar data is used in [6] to train the model to separate audio for each of the speakers and subsequent speech recognition on each of the resulting channels. However, in this work, such data are used to train the model on signal separation, and not for end-to-end multi-speaker speech recognition. Also in this article, the case of a large number of speakers and long audio recordings is considered, so such data are not suitable for the multi-channel transformer training.

In order to test the models, several datasets were collected from the LibriSpeech test cases (test-clean and test-other): a dataset with one speaker on each audio recording (original data) and mixed pairs of these audio recordings (each audio recording participated in no more than one synthetic example). The overlap size of audio recordings is measured as a percentage of the length of the segment with overlapping audio from the length of the first audio recording. Mixing was carried out for all possible overlaps from 0 % to 100 % in 10 % increments, forming eleven synthetic test datasets.

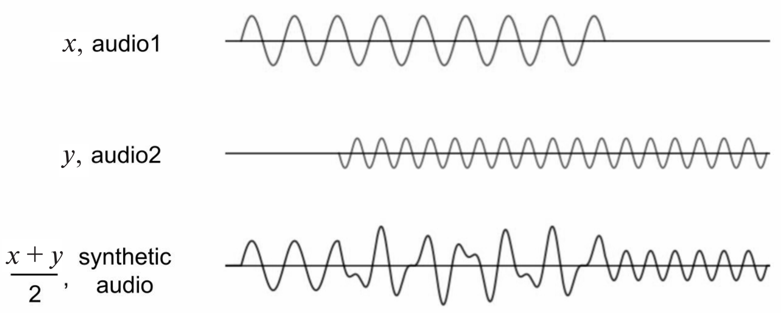


Fig. 2. Synthetic data for training multi-channel transformer with two speakers

Baseline solutions

In existing articles with different approaches to solving the problem of multi-speaker speech recognition there is no source code or pre-trained models. Therefore, similarly to [4], in order to compare the results of the multi-channel transformer, a baseline solution for multi-speaker speech recognition was built.

Firstly, the "single-speaker transformer" model was built, recognizing speech on each channel with their known separation. It is using a single-channel transformer model for speech recognition. The architecture and hyperparameters of this model were similar to the multi-channel transformer. The single-speaker transformer does not solve the stated problem of multi-speaker speech recognition because it uses generally unknown information about channel separation, but this model can be considered as a lower estimate for the potential quality of the multi-channel transformer, since their architectures are almost the same.

Secondly, the "transformer with channel separation" model was built. In it, similarly to [6], a model for separating the audio signal is used first (specifically, the Sepformer model, [20], pre-trained in [21]), and then the same single-channel transformer for speech recognition is used for each channel.

The single-channel speech recognition transformer and multi-channel transformer were trained on 8 A100 GPUs with a batch size of 16 audio recordings on each GPU. In the training process of the multi-channel transformer, 30 % of the examples were taken uniformly from all audio with one speaker, and 70 % – uniformly from synthetic data with two speakers. In the training of the single-channel transformer, only data with one speaker was used. The multi-channel transformer was trained for eight days, and the single-channel transformer was trained for three days. In both models, the AdamW optimizer ([22]) was used with the parameters weight decay of 0.01, learning rate of 10^{-4} and the subsequent two decreases of the learning rate by ten times when the model error function reached a plateau. Both models used 12 layers of Transformer Encoder, 6 layers of Transformer Decoder, 16 attention-heads, $d_{model} = 512$.

Results

Solutions were tested on test data from mixed pairs of LibriSpeech data. Table 1 shows the results of the cpWER metric for three models: single-speaker transformer, transformer with channel separation, and the presented model. The results are presented for several different overlaps of two audio recordings when they are mixed (0 %, 20, 50, 70, 90 %). The experiments were carried out with transformer models generating recognition using beam search with different search widths (1 and 5).

The results of the single-speaker transformer do not depend on the size of the overlap of the mixed audio recordings, since this model is fed to the input channels of each speaker separately, which are known even before the mixing process and do not depend on its results. The results of the single-speaker transformer can be considered as a lower estimate for the potential quality of the multi-channel transformer, since their architectures are almost the same.

Table 1

CpWER on synthetically mixed data from test-clean and test-other when generating transformer models recognitions using beam search with different search widths (1 and 5) for different overlaps of audio recordings in three compared models: single-speaker transformer (transformer), transformer with channel separation (sepformer baseline) and the presented model

Synthetic test-clean, cpWER, %						
	Overlap	0	20	50	70	90
Beam search, 1	transformer	4.6 ± 0.2				
	sepformer baseline	21.2 ± 0.4	20.0 ± 0.3	20.2 ± 0.4	20.1 ± 0.2	**19.8 ± 0.3**
	multi-channel transformer*	**5.9 ± 0.4**	**6.0 ± 0.3**	**9.8 ± 0.2**	**12.0 ± 0.3**	21.3 ± 0.7
Beam search, 2	transformer	4.4 ± 0.2				
	sepformer baseline	20.6 ± 0.5	19.6 ± 0.5	20.0 ± 0.4	18.6 ± 0.4	**18.9 ± 0.7**
	multi-channel transformer*	**5.7 ± 0.3**	**5.8 ± 0.4**	**9.5 ± 0.4**	**11.5 ± 0.5**	20.2 ± 0.7
Beam search, 1	transformer	10.6 ± 0.5				
	sepformer baseline	29.7 ± 0.5	29.5 ± 0.5	29.1 ± 0.6	29.1 ± 0.5	29.5 ± 0.6
	multi-channel transformer*	**11.6 ± 0.5**	**12.3 ± 0.5**	**16.5 ± 0.5**	**19.3 ± 0.6**	**25.5 ± 0.5**
Beam search, 2	transformer	10.0 ± 0.4				
	sepformer baseline	29.1 ± 0.6	29.1 ± 0.5	28.3 ± 0.5	29.1 ± 0.5	29.3 ± 0.5
	multi-channel transformer*	**11.0 ± 0.4**	**11.9 ± 0.5**	**16.0 ± 0.4**	**18.7 ± 0.5**	**24.8 ± 0.6**

Additionally, Fig. 3 shows a diagram of the dependence of the cpWER metric on different overlaps of audio recordings. To generate the results of this diagram, the model was applied using beam search with a search width of 5.

For synthetic data from test-clean and test-other with overlaps of 0—20 %, the multi-channel transformer shows quality of cpWER ~ 15 % better than the transformer with channel separation, and ~ 1—2 % worse than the transformer with known channel separation. Moreover, it can be seen from Fig. 3 that the multi-channel transformer beats the transformer with channel separation in case of small overlaps (up to 80 % on synthetic test-clean and all overlaps for test-other).

On synthetic audio recordings with a large overlap, the multi-channel transformer tends to poorly, but consonantly, recognize overlapping parts of utterances. An example of such a recognition result is presented in Fig. 4. For an overlap of 90 % in the example presented, the hypothesis of the multi-channel transformer gets cpWER ~ 40 % worse than its predictions for an overlap of 10 %. This may be due to the complexity of speech recognition in the case of large overlaps, which turns into a channel separation problem.

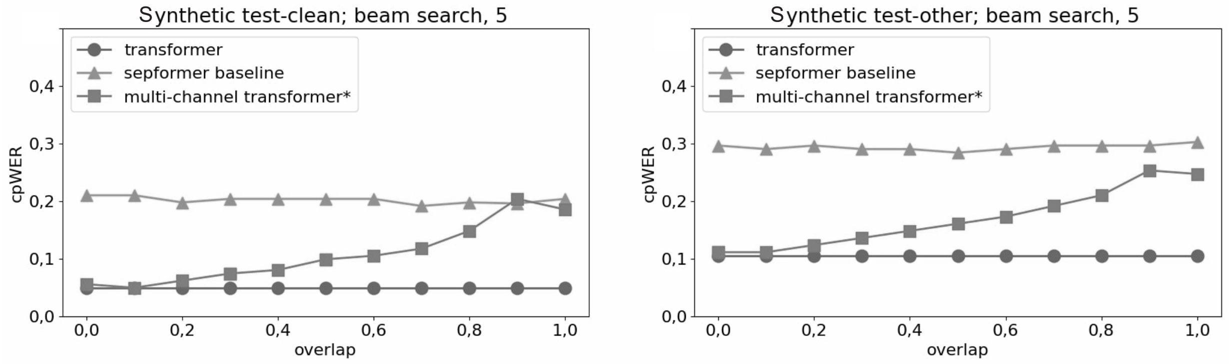
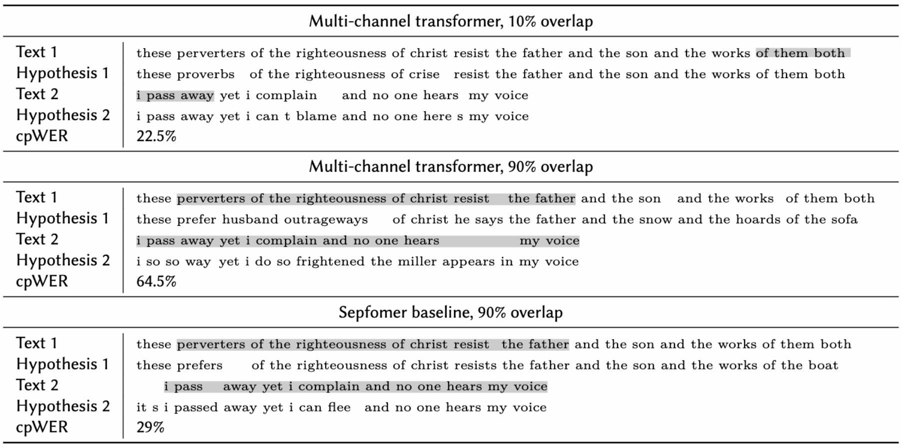
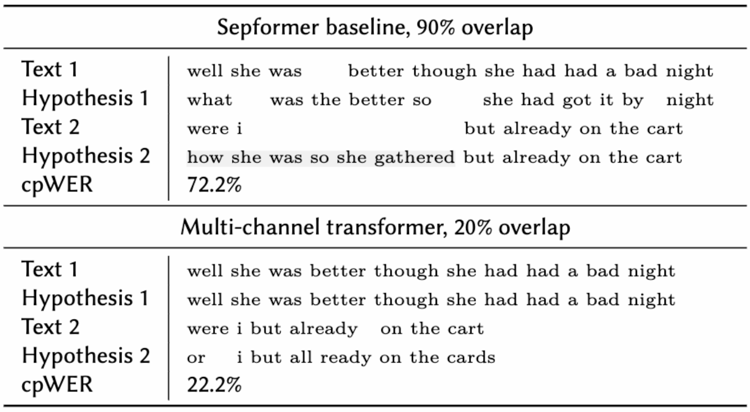


Fig. 3. Diagram of the cpWER dependence on the overlap of mixed audio recordings on synthetic data from test-clean and test-other for the three models compared



Multi-channel transformer, 10% overlap	
Text 1	these perverters of the righteousness of christ resist the father and the son and the works of them both
Hypothesis 1	these proverbs of the righteousness of crise resist the father and the son and the works of them both
Text 2	i pass away yet i complain and no one hears my voice
Hypothesis 2	i pass away yet i can t blame and no one here s my voice
cpWER	22.5%
Multi-channel transformer, 90% overlap	
Text 1	these perverters of the righteousness of christ resist the father and the son and the works of them both
Hypothesis 1	these prefer husband outrageways of christ he says the father and the snow and the hoards of the sofa
Text 2	i pass away yet i complain and no one hears my voice
Hypothesis 2	i so so way yet i do so frightened the miller appears in my voice
cpWER	64.5%
Sepfomer baseline, 90% overlap	
Text 1	these perverters of the righteousness of christ resist the father and the son and the works of them both
Hypothesis 1	these prefers of the righteousness of christ resists the father and the son and the works of the boat
Text 2	i pass away yet i complain and no one hears my voice
Hypothesis 2	it s i passed away yet i can flee and no one hears my voice
cpWER	29%

Fig. 4. Recognition of the multi-channel transformer and transformer with channel separation on one example from synthetic test-clean for different overlap sizes. The words correctly recognized by the model are aligned with their originals. The part of the texts that overlaps is highlighted in grey



Sepformer baseline, 90% overlap	
Text 1	well she was better though she had had a bad night
Hypothesis 1	what was the better so she had got it by night
Text 2	were i but already on the cart
Hypothesis 2	how she was so she gathered but already on the cart
cpWER	72.2%
Multi-channel transformer, 20% overlap	
Text 1	well she was better though she had had a bad night
Hypothesis 1	well she was better though she had had a bad night
Text 2	were i but already on the cart
Hypothesis 2	or i but all ready on the cards
cpWER	22.2%

Fig. 5. Recognition of the transformer with channel separation and multi-channel transformer on one example from synthetic test-clean. The words correctly recognized by the model are aligned with their originals. The part of speech recognition obtained after the artefacts of separating the audio recording into channels is highlighted in grey

On the other hand, the single-speaker transformer tends to recognize multi-speaker speech worse due to various artefacts in the audio recording channels after separation. For example, in the sample from Fig. 5, after splitting into channels, part of the first audio recording was duplicated in a quiet tone on the second channel, which is why both texts from subsequent speech recognition contain recognition of this piece.

Table 2

CpWER in the case of one speaker for test-clean and test-other for the single-speaker (transformer) and multi-channel transformer when using beam search with different search widths (1 and 5)

	Test-clean, cpWER, %		Test-other, cpWER, %	
	beam search, 1	beam search, 5	beam search, 1	beam search, 5
Transformer	4.5 ± 0.3	4.3 ± 0.3	10.0 ± 0.3	9.3 ± 0.4
Multi-channel transformer*	5.3 ± 0.3	4.7 ± 0.3	10.8 ± 0.4	9.8 ± 0.4

Table 2 shows the results of experiments with the multi-channel transformer and single-channel transformer on audio recordings with one speaker. The results show that in the case of one speaker instead of two, the presented model differs in quality from the usual transformer model that can recognize the speech of only one speaker in only ~ 1 % cpWER.

The proposed multi-channel transformer model achieves results similar to modern multi-speaker recognition systems (~4–5 % on test-clean single-speaker samples and ~ 4–6 % cpWER on two-speaker synthetic test-clean samples), but has faster inference due to parallel speakers recognition architecture. The number of model runs needed to recognize texts $h_1, h_2, ..., h_n$ is $\max_{i=1...n} |h_i|$.

The number of model runs required by modern consequent speaker recognition techniques is $\sum_{i=1...n} |h_i|$.

Thus, the proposed solution speeds up model inference by two times in the best case of two-speaker speech (when the length of different speaker hypotheses are approximately the same), and has the same speed in the worst case (when only one speaker is presented).

Conclusion

Based on the results of the study, a model was built that solves the problem of possibly overlapping multi-speaker speech recognition of no more than two speakers. The model was trained and tested on synthetic data from LibriSpeech and showed better quality than the transformer with channel separation for small overlaps of audio recordings (up to 80 % cpWER on test-clean and 90 % cpWER on test-other), as well as a small difference from the model that recognizes a single speaker, in the case of such data. The model also showed results similar to modern state-of-the-art multi-speaker speech recognition solutions, but can be inferenced faster, which could be beneficial in many tasks, for example, real-time multi-speaker speech recognition.

The presented model has a potential for many applications, for example, transcription of meeting records, or mark-up and subsequent analysis (for example, [23]) of multi-speaker speech. To solve such problems, it will only be necessary to overcome the problem of the model's inability to recognize speech on sufficiently long audio recordings (this problem, as for any similar auto-regressive models, arises due to the limitations in the memory of computing resources) and, possibly, to investigate generalisations of the model to a larger number of speakers. More details about these problems and their possible solutions can be found in the "Future work" section.

Another result of the study is a data construction scheme. The presented streaming method of generating two-speakers training examples by mixing pairs of audio recordings with different overlaps greatly expands the size of the dataset, and also forces the models to learn to recognize audio recordings with possible speech overlaps.

In addition, it is expected that the multi-channel transformer will work no worse on real audio recordings with two speakers than on synthetic data, since the problem of dividing audio into channels, which in fact the multi-channel transformer is facing in the case of large overlaps in synthetic audio recordings, is more difficult than conventional recognition of dialogue speech. However, due to the lack of open-source datasets with short dialogues, the relevant studies were conducted only on a closed-source dataset and therefore are not presented in this article. The same studies on proprietary data showed that the multi-channel transformer demonstrates high quality not only on LibriSpeech audio recordings recorded with a minimum amount of background noise by speakers in the studio, but also on noisier data.

Future work

One of the shortcomings of the proposed solution is its inability to recognize speech on long audio recordings due to the limitations in the memory of modern computing resources. A possible way to solve this problem is to combine the recognition of small audio recording windows, which may also allow solving the problem of recognizing audio recordings in real time. When applying this approach to the proposed model, it will be necessary to match the local order of speakers in the audio recording window to their global order, which is a problem for further study.

Additionally, a useful functionality of the model would be to mark up the utterances of each speaker in time inside the audio recording, that is, to determine the time intervals for each word/symbol from the recognition hypothesis. Such information, for example, would help when gluing window recognitions of a long audio recording to identify the same words from different windows.

Another field for research is the case of a large number of participants in the dialogue (more than two). In some situations, the upper limit on the number of speakers on an audio recording may be unknown, or it may be very large. It is yet to be found whether the multi-channel transformer generalised to a larger number of speakers can produce the results of the same how high quality.

Another problem is to select the optimal architecture of the proposed model. Experiments were carried out with only two encoder variants (the original and the encoder from [17]) and two decoding variants (with two and one decoder). However, it is not known how much the results could be improved by using any other architecture. Unfortunately, such experiments with the architecture of the model or increasing the upper bound by the number of speakers require a lot of time and computing power.

REFERENCES

1. **Park T.J., Kanda N., Dimitriadis D., Han K.J., Watanabe S., Narayanan S.** *A review of speaker diarization: Recent advances with deep learning*. 2021. DOI: https://doi.org/10.48550/arXiv.2101.09624

2. **Shafey L.E., Soltau H., Shafran I.** *Joint speech recognition and speaker diarization via sequence transduction*. 2019. DOI: https://doi.org/10.48550/arXiv.1907.05337

3. **Kanda N., Horiguchi S., Fujita Y., Xue Y., Nagamatsu K., Watanabe S.** *Simultaneous speech recognition and speaker diarization for monaural dialogue recordings with target-speaker acoustic models*. 2019. DOI: https://doi.org/10.48550/arXiv.1909.08103

4. **Kanda N., Gaur Y., Wang X., Meng Z., Chen Z., Zhou T., Yoshioka T.** *Joint speaker counting, speech recognition, and speaker identification for overlapped speech of any number of speakers*. 2020. DOI: https://doi.org/10.48550/arXiv.2006.10930

5. **Kanda N., Ye G., Gaur Y., Wang X., Meng Z., Chen Z., Yoshioka T.** *End-to-end speaker-attributed ASR with transformer*. 2021. DOI: https://doi.org/10.48550/arXiv.2104.02128

6. **Raj D., Denisov P., Chen Z., Erdogan H., Huang Z., He M., Watanabe S., Du J., Yoshioka T., Luo Y., Kanda N., Li J., Wisdom S., Hershey J.R.** *Integration of speech separation, diarization, and recognition for multi-speaker meetings: system description, comparsion, and analysis*. 2020. DOI: https://doi.org/10.48550/arXiv.2011.02014

7. **Bronkhorst A.** The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica united with Acustica*. 86 (2000): 117—128. DOI: 10.3758/s13414-015-0882-9

8. **Pavlichenko N., Stelmakh I., Ustalov D.** *Crowdspeech and voxdiy: Benchmark datasets for crowdsourced audio transcription*. 2021. DOI: https://doi.org/10.48550/arXiv.2107.01091

9. **Lyudvichenko V., Vatolin D.** *Predicting video saliency using crowdsourced mouse-tracking data*. 2019. DOI: http://dx.doi.org/10.30987/graphicon-2019-2-127-130

10. **Fujita Y., Kanda N., Horiguchi S., Xue Y., Nagamatsu K., Watanabe S.** *End-to-end neural speaker diarization with self-attention*. 2019. DOI: https://doi.org/10.1109/ASRU46091.2019.9003959

11. **Kim J., Kong J., Son J.** *Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech*. 2021. DOI: https://doi.org/10.48550/arXiv.2106.06103

12. **Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I.** *Attention is all you need*. 2017. DOI: https://doi.org/10.48550/arXiv.1706.03762

13. **Mohamed A., Okhonko D., Zettlemoyer L.** *Transformers with convolutional context for ASR*. 2020. DOI: https://doi.org/10.48550/arXiv.1904.11660

14. **Li N., Liu S., Liu Y., Zhao S., Liu M., Zhou M.** *Neural speech synthesis with transformer network*. 2018. DOI: https://doi.org/10.48550/arXiv.1809.08895

15. **Watanabe S., Mandel M., Barker J., Vincen E., Arora A., Chang X., Khudanpur S., Manohar V., Povey D., Raj D., Snyder D., Subramanian A.S., Trmal J., Yair B.B., Boeddeker C., Ni Z., Fujita Y., Horiguchi S., Kanda N., Yoshioka T., Ryant N.** *Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings*. 2020. DOI: 10.21437/CHiME.2020-1

16. **Kudo T., Richardson J.** *Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing*. 2018. DOI: https://doi.org/10.48550/arXiv.1808.06226

17. **Gulati A., Qin J., Chiu C.-C., Parmar N., Zhang Y., Jiahui Yu W.H., Wang S., Zhang Z., Wu Y., Pang R.** *Conformer: Convolution-augmented transformer for speech recognition*. 2020. DOI: 10.21437/Interspeech.2020-3015

18. **Panayotov V., Chen G., Povey D., Khudanpur S.** *Librispeech: An ASR corpus based on public domain audio books*. 2015. DOI: https://doi.org/10.1109/ICASSP.2015.7178964

19. **Park D.S., Chan W., Zhang Y., Chiu C.-C., Zoph B., Cubuk E.D., Le Q.V.** *Specaugment: A simple data augmentation method for automatic speech recognition*. 2019. DOI: 10.21437/Interspeech.2019-2680

20. **Subakan C., Ravanelli M., Cornell S., Bronzi M., Zhong J.** *Attention is all you need in speech separation*. 2021. DOI: https://doi.org/10.48550/arXiv.2010.13154

21. **Ravanelli M., Parcollet T., Plantinga P., Rouhe A., Cornell S., Lugosch L., Subakan C., Dawalatabad N., Heba A., Zhong J., Chou J.-C., Yeh S.-L., Fu S.-W., Liao C.-F., Rastorgueva E., Grondin F., Aris W., Na H., Gao Y., Mori R.D., Bengio Y.** *Speech-brain: A general-purpose speech toolkit*. 2021. DOI: http://dx.doi.org/10.21437/Interspeech.2022-10644

22. **Loshchilov I., Hutter F.** *Decoupled weight decay regularization*. 2018. DOI: https://doi.org/10.48550/arXiv.1711.05101

23. **Narayanan S., Georgiou P.** *Behavioral signal processing: Deriving human behaviorial informatics from speech and language*. 2013. DOI: https://doi.org/10.1109/JPROC.2012.2236291

INFORMATION ABOUT AUTHORS / СВЕДЕНИЯ ОБ АВТОРАХ

Фадеева Екатерина Сергеевна
Ekaterina S. Fadeeva
E-mail: rediska@yandex-team.ru

Ершов Василий Алексеевич
Vasily A. Ershov
E-mail: noxoomo@yandex-team.ru

Поступила: 29.11.2022; Одобрена: 26.12.2022; Принята: 12.01.2023.
Submitted: 29.11.2022; Approved: 26.12.2022; Accepted: 12.01.2023.