Abstract

The sixth oriental language recognition (OLR) challenge focuses on language identification (LID) and automatic speech recognition (ASR) within multilingual scenarios. OLR 2021 includes four tasks: (1) constrained LID, (2) unconstrained LID, (3) constrained multilingual ASR, (4) unconstrained multilingual ASR. The LID tasks, C_avg and equal error rate (EER) are considered as primary and secondary metrics, respectively. For the ASR tasks, character error rate (CER) is set as the unique metric. In this challenge, there were 48 participating teams and about half of the teams submitted valid results. Compared to the official baseline systems, the top-ranking teams achieved great progress. Among them, the best submission of constrained LID reduced $C_{avg}$ from 0.0817 to 0.0025, meanwhile, the unconstrained LID achieved a very low $C_{avg}$ of 0.0039. As for ASR tasks, the CER was reduced to 13.1% in the constrained ASR and 12.6% in the unconstrained ASR, respectively. This paper describes the above four tasks, database profile, and gives the analysis of results in detail.

Index Terms: language identification, speech recognition, multilingual, oriental language, OLR 2021 challenge

1. Introduction

The oriental language recognition (OLR) challenge [1–5] is an annual competition, which aims at improving the research on multilingual phenomena and advancing the development of multilingual speech technologies.

OLR 2021 [6] includes more languages, dialects, and real-life data, and it focuses on more practical and challenging tasks within multilingual sessions. OLR 2021 not only includes traditional LID tasks but also introduces multilingual ASR tasks to improve practical value. Identifying language categories is a prior process for multilingual ASR. Therefore, we would like to encourage participants to deeply investigate the close interaction between LID and ASR, and treat them as an integrated solution for the multilingual phenomenon. Meanwhile, OLR 2021 introduces open-training tasks for both LID and multilingual ASR tracks, which has brought more potential and meaningful space for participants to play.

OLR 2021 includes four tasks:

- **Constrained LID**: A cross-domain identification task involves 13 languages with constrained training data.
- **Unconstrained LID**: An open-training identification task involves 17 languages. And the test utterances in this task are obtained from real-life environments.
- **Constrained multilingual ASR**: A fixed-training multilingual ASR task. Only the data provided by the organizer can be used to train the acoustic and language models.
- **Unconstrained multilingual ASR**: An open-training multilingual ASR task. Any data is allowed to train the acoustic and language models.

The remaining sections are organized as follows. Section 2 describes the data profile. Section 3 introduces the evaluation metrics and the baseline systems of LID and ASR tasks, successively. Section 4 analyses all the submission results and concludes the algorithms involved in this competition. Section 5 discusses some novel approaches. Section 6 is an overview of the OLR 2021 challenge.

2. Database profile

The OLR 2021 database covers 17 languages, some of which are Chinese dialects, and is provided by Speechocean and the NSFC M2ASR project [7]. It consists of a training set, which was recomposed from the last five years’ data, and two newly provided test sets.

The training set, up to 280 hours, was released to the participants for system construction, and these are all the audio data permitted for the constrained tasks. The duration distribution of each language in the training set is shown in Figure 1.

Besides the speech signals, the OLR16-OL7, OLR17-OL3, OLR20-dialect, and OLR20-test datasets also provide lexicons of all the corresponding languages, as well as the transcriptions of the training utterances. These resources allow the training of voice-based, speech-based, and even telephone-based speech recognition systems. More information about the training set can be found in the paper [6].

We prepared two standard test sets for the OLR 2021 challenge: the OLR21-cross-domain-test and the OLR21-wild-test. They both are divided into 30% progress subset and 70% evaluation subset.

- **OLR21-cross-domain-test**: A test set designed for three tasks: the constrained LID task, the constrained ASR task, and the unconstrained ASR task. It contains 13 languages and was recorded by different equipments in various environments.
- **OLR21-wild-test**: A test set designed for the unconstrained LID task. It contains 17 languages. Utterances in this subset are obtained from real-life environments.

3. Evaluation metrics and baseline

3.1. OLR-LID

As in NIST LRE15 [8], the OLR 2021 challenge chooses $C_{avg}$ as the principle evaluation metric. First, define the pair-wise...
loss that composes the missing and false alarm probabilities for a particular target/non-target language pair:

\[ C(L_t, L_n) = P_{\text{Target}} \cdot P_{\text{Miss}}(L_t) + (1 - P_{\text{Target}}) \cdot P_{\text{FA}}(L_t, L_n) \]  \hfill (1)

where \( L_t \) and \( L_n \) are the target and non-target languages, respectively; \( P_{\text{Miss}} \) and \( P_{\text{FA}} \) are the missing and false alarm probabilities, respectively. \( P_{\text{Target}} \) is the prior probability for the target language, which is set to 0.5 in the evaluation. Then the principle metric \( C_{\text{avg}} \) is defined as the average of the above pair-wise performance:

\[ C_{\text{avg}} = \frac{1}{N} \sum_{L_t} \left\{ P_{\text{Target}} \cdot P_{\text{Miss}}(L_t) + \sum_{L_n} P_{\text{Non-Target}} \cdot P_{\text{FA}}(L_t, L_n) \right\} \]  \hfill (2)

where \( N \) is the number of languages, and \( P_{\text{Non-Target}} = (1 - P_{\text{Target}}) / (N - 1) \). For the open-set testing condition, all of the interfering languages are treated as a single unknown language in the computation of \( C_{\text{avg}} \). We have provided the evaluation scripts for system development.

The baseline LID systems in this challenge was an extended TDNN x-vector model [9] constructed with ASV-Subtools [10]. The feature extraction and back-end were all conducted with Kaldi [11]. The recipes of the LID baseline can be downloaded from the website\(^1\).

### 3.2. OLR-ASR

Considering that the dataset consists of languages with different grammatical rules, we choose CER as the primary metric for ASR tasks. The CER is calculated as the sum of deletion, insertion, and substitution errors in the ASR output compared to the reference transcript, divided by the total number of characters in the reference transcription:

\[ \text{CER} = \frac{\#\text{Deletions} + \#\text{Insertions} + \#\text{Substitutions}}{\#\text{Reference Characters}} \]  \hfill (3)

We use the CER of the entire evaluation set as the basis for the ranking and provide CER for each language as a reference. Punctuation and special tags are not included in the calculation of CER.

\(^1\) https://github.com/Snowdar/asv-subtools#3-olr-challenge-2021-baseline-recipe-language-identification

For the ASR tasks, we employed a transformer [12] based end-to-end model as a baseline, which combined encoder-decoder structure with an attention mechanism. And we utilized the standard transformer configuration in ESPnet [13] which consists of 12-layer encoder and 6-layer decoder, each layer has 2048 hidden units. The recipes of the ASR baseline can be downloaded from the website\(^2\).

### 4. Results analysis

#### 4.1. OLR-LID tasks

Figure 2 illustrates the Detection Error Tradeoff (DET) curves of the baseline system and the top 11 systems, which outperformed the baseline system in the constrained LID task. In this figure, the colored solid line, gray solid line, and colored dotted line represent the top 3 systems, other systems, and baseline system respectively. The test set for this task is cross domain data containing 13 languages. Because of the introduction of multilingual ASR tasks, some of the speech data is given test labels this year, which makes more research methods applicable to the language recognition track. From the graph, we can see that teams have done amazingly well in cross domain tasks. The top 3 systems deliver EERs below 1%. Most of the systems deliver EERs below 10%.

Figure 3 illustrates the DET curves of the top 3 systems in Unconstrained LID tasks. The test set for this task contains 17 languages.

\(^2\) https://speech.xmu.edu.cn/2021/0727/c19782a436940/page.htm
languages, which was obtained from real-life environments. It is therefore more challenging than the constrained LID task. In this task, any data (except the evaluation data) one can access is allowed to build the system. Due to the complexity of this task, only a small number of teams submitted results, and most of the results less than satisfactory. However, the top 3 systems outperform other system with a large margin.

In both constrained LID and unconstrained LID, the best systems are from X-Voice. Their strategy is to pretrain the LID model as an ASR task to integrate the phonetic information that helps to enhance language identification, and then the pre-trained model is further finetuned until its optimum; utilize the strategy of model’s fusion and ensemble for improved performances. Adopting an ASR encoder is the most effective method in this competition.

Table 1 and Table 2 display the final rankings for the Constrained LID task and the Unconstrained task.

### 4.2. OLR-ASR tasks

For the constrained multilingual ASR task, the optimal baseline is the multilingual model based on transformer built with ESPnet. Table 3 shows the CER for all submitted systems and the ASR baseline. Almost all systems outperform the baseline system with a large margin. Moreover, while most of the systems deliver CER between 15% and 35%, the champion system, CCDL, delivers CER below 15%, even to 13.1%. They first used LID model to identify language and then performed speech recognition. They built E2E multilingual model based on Conformer [14] and used CNN-TDNNF architecture to compose hybrid monolingual model. As for LID system, they used a language identification classifier based on ResNet34-SE [15] and the edit-distance between the E2E multilingual ASR system and hybrid monolingual ASR system to determine the language tag of recordings, which can effectively alleviate the problem of language model due to the limit of training data. Further they fused their model based on the confidence of LID and ASR systems. It is worth noting that even if there are more or less differences, there are other teams that have performed speech recognition after using the LID model to identify language.

The rank of unconstrained multilingual ASR task is shown on Table 4. For this task, among the 6 participating teams, 5 teams had better results than the baseline. In this task, the top-performing system, CCDL, utilized WenetSpeech [16] dataset, Common Voice, CSJ dataset, MagicData, and Google TTS datasets except for the training set in the constrained condition. They used WeNet toolkit [17] to build the same system as constrained multilingual ASR task. In this way, they similarly achieved the best CER of 12.6%.

Descriptions of the top 3 systems for each task can be downloaded from the challenge web site.

### 5. Discussion

In this challenge, most teams completed the LID tasks by combining the ASR task with the language recognition task through transfer learning. The performance of the language recognition system is enhanced by exploiting the prior knowledge in the ASR model. Similarly, many teams completed the ASR tasks by performing LID first.

Figure 4 shows the flow of LID and ASR. We will summarize the various methods used by the OLR 2021 team in their submission by point.

#### a) OLR-LID tasks

- Augmentation: Augmentation (e.g., velocity, volume perturbations) is widely used. Some systems use background noise extracted from the training data, mp3/mp4a

---

In this paper, we present the database profile, evaluation metrics, task definitions, and baselines of the OLR 2021 challenge. We also summarize the ranking results of the participating teams and analyze the technologies that the participants used. In addition to the traditional LID task, constrained multilingual ASR and unconstrained multilingual ASR, and the unconstrained LID task were introduced in this challenge for the first time. Further, the script of the baseline systems was made public to allow more people to participate in this competition. The participants used a wide variety of approaches to improve the performance on the four tasks. In LID tasks, obtaining additional information through ASR model to reduce the impact of noise, unknown channels, and unknown test sets is the most common method. Meanwhile in ASR tasks, almost teams used LID model to identify the languages before processing their speech recognition. Naturally we come to a conclusion that LID and multilingual ASR complement each other. In the future, we hope to see more collaborative studies for both tasks.

6. Conclusions

In this paper, we present the database profile, evaluation metrics, task definitions, and baselines of the OLR 2021 challenge. We also summarize the ranking results of the participating teams and analyze the technologies that the participants used. In addition to the traditional LID task, constrained multilingual ASR and unconstrained multilingual ASR, and the unconstrained LID task were introduced in this challenge for the first time. Further, the script of the baseline systems was made public to allow more people to participate in this competition. The participants used a wide variety of approaches to improve the performance on the four tasks. In LID tasks, obtaining additional information through ASR model to reduce the impact of noise, unknown channels, and unknown test sets is the most common method. Meanwhile in ASR tasks, almost teams used LID model to identify the languages before processing their speech recognition. Naturally we come to a conclusion that LID and multilingual ASR complement each other. In the future, we hope to see more collaborative studies for both tasks.

7. Acknowledgements

We would like to thank Ming Li at Duke Kunshan University, Xiaolei Zhang at Northwestern Polytechnical University, Limin Lai at Xiann University for their help in organizing this OLR 2021 challenge. This work was supported by the National Natural Science Foundation of China No.61876160, No.61633013, and No.62001405.
8. References


