information

Article

Language Identification-Based Evaluation of Single Channel Speech Separation of Overlapped Speeches

Zuhragvl Aysa, Mijit Ablimit *, Hankiz Yilahun and Askar Hamdulla

College of Information Science and Engineering, Xinjiang University, Urumqi 830017, China
* Correspondence: mijit@xju.edu.cn; Tel.: +86-133-0991-2366

Abstract: In multi-lingual, multi-speaker environments (e.g., international conference scenarios), speech, language, and background sounds can overlap. In real-world scenarios, source separation techniques are needed to separate target sounds. Downstream tasks, such as ASR, speaker recognition, speech recognition, VAD, etc., can be combined with speech separation tasks to gain a better understanding. Since most of the evaluation methods for monophonic separation are either single or subjective, this paper used the downstream recognition task as an overall evaluation criterion. Thus, the performance could be directly evaluated by the metrics of the downstream task. In this paper, we investigated a two-stage training scheme that combined speech separation and language identification tasks. To analyze and optimize the separation performance of single-channel overlapping speech, the separated speech was fed to a language identification engine to evaluate its accuracy. The speech separation model was a single-channel speech separation network trained with WSJ0-2mix. For the language identification system, we used an Oriental Language Dataset and a dataset synthesized by directly mixing different proportions of speech groups. The combined effect of these two models was evaluated for various overlapping speech scenarios. When the language identification network model was based on single-person single-speech frequency spectrum features, Chinese, Japanese, Korean, Indonesian, and Vietnamese had significantly improved recognition results over the mixed audio spectrum.

Keywords: speech separation; Conv-TasNet; language identification; overlap rate; spectrogram

Citation: Aysa, Z.; Ablimit, M.; Yilahun, H.; Hamdulla, A. Language Identification-Based Evaluation of Single Channel Speech Separation of Overlapped Speeches. *Information* **2022**, *13*, 492. https://doi.org/10.3390/info13100492

Academic Editors: David Martins De Matos and Zahir M. Hussain

Received: 29 July 2022
Accepted: 8 October 2022
Published: 11 October 2022

1. Introduction

Language identification, as a front-end system for natural language processing technologies such as machine translation and multi-lingual information services, is a hot topic of research, and language recognition in realistic noisy environments has received more attention in recent years. In noisy and multi-speaker environments, the human ear may not be able to separate and identify the language types accurately, and language features may easily be disturbed or masked, preventing a clear representation of language information. Therefore, it is increasingly important to study voices in real-world scenarios with multiple speakers and in multi-lingual environments. In today's globally integrated world, languages and dialects are mixed unprecedentedly. Neural network models brought the possibility of separating and understanding overlapped voices and boosted multi-lingual processing research, bringing significant progress for applying low-resource languages and dialects [1]. In scenarios with multiple speakers and background noise, obtaining monolingual information for each speaker is difficult. It has been a novel topic in speech separation and language identification in recent years. Therefore, it is necessary to carry out a speech separation process before recognizing monolingual speech and understanding the content. As the first step in speech research, speech separation techniques are vital in determining the effectiveness of the speech back-end. The term speech separation initially originated from "Cherry's Cocktail Party Problem" [2] in 1953, where a listener could effortlessly hear a person's speech surrounded by other people's speech and ambient noise

even in a cocktail party-like sound environment. The speech separation problem is often called the "cocktail party problem." The goal of speech separation is to extract one or more source signals from a mixed speech signal containing two or more sources, with each speaker corresponding to one of the source signals [3]. The two main speech separation methods currently under study are the single-channel speech separation method and the multi-channel speech separation method based on microphone arrays. This paper focuses on a single-channel two-speaker overlapped speech separation technique [4]. Although humans can focus on one of the overlapping speech sounds, research methods still face difficulties in achieving this. Speech separation methods have enabled many speech processing algorithms, such as automatic speech recognition (ASR), to achieve better performance under multi-peak conditions.

Language identification technology is the process of determining the type of language to which speech content belongs using automated methods [1]. In addition, in noisy and multi-speaker environments, the human ear may not be able to identify language types accurately, and language features may be easily disturbed or masked, preventing a clear representation of language information. Therefore, it is increasingly important to study language identification in real-world scenarios with multiple speakers and a multi-lingual environment. The key to language identification is feature extraction and the construction of language models. Currently, standard language features are mainly based on acoustic and phonetic features. The mainstream acoustic layer features include Mel frequency cepstral coefficients, Gamma pass frequency cepstral coefficients, and so on [5]. The above features are prone to noise, resulting in poor identification results. The phoneme layer-based language identification method divides the speech into a sequence of phonemes and then recognizes the language according to the phoneme pairing between different languages [6]. The phoneme layer-based features are less affected by noise, but phoneme segmentation is difficult to extract, resulting in a degraded identification performance. Recent research has begun to apply deep learning methods to language identification tasks and to use this method to train neural networks as phonological models. The method treats the language identification problem as a classification problem and thus trains the CNN to use the relevant linguistically labeled speech spectrograms.

However, most deep learning-based studies of single-channel speech separation rely on a single metric to evaluate the system, and the evaluation metric is relatively homogeneous. In our experiments, we indirectly analyzed the performance of the front-end network by looking at the evaluation metrics or the good or bad performance of the downstream tasks. Therefore, this paper focuses on the Single-Channel Speech Separation (SCSS) problem and applies the SCSS problem to complex multi-speaker multi-lingual scenarios. The experimental results obtained from the language identification network can be used as an intermediate reference index for the final SCSS network to optimize our separation network.

The rest of the paper is organized as follows. The current state of relevant research is presented in Section 2, the algorithmic model structure is described in Section 3, the experimental setup is described in Section 4, and the experimental results and analysis are presented in Section 5.

2. Related Work

Deep neural networks have been used for speech separation tasks with the development of deep learning techniques. One of the most common frameworks is the estimation of time-frequency masking of speech signals using neural networks. The basic theory is that after the short-time Fourier transform of speech, the energy at each time-frequency point is the sum of the powers of all target speech at that time-frequency point. There are differences in the distribution characteristics between different speakers' speech in the time frequency domain, and speech separation can be performed by estimating the correct time-frequency masking matrix. In a similar speech enhancement task where the interference is a noisy signal, researchers have used different loss functions. Qi, J. et al. [7] proposed to improve

vector-to-vector regression for feature-based speech enhancement using a new distribution loss. The approach essentially takes into account the estimation uncertainty, and it avoids the burden of deploying additional speech enhancement models to obtain a final clean speech. In the same year, Qi, J. [8] also proposed to use the properties of mean absolute error as a loss function for vector-to-vector regression based on deep neural networks for speech enhancement experiments and verified that DNN-based regression optimized using the MAE loss function can obtain lower loss values than the MSE counterpart. Siniscalchi, S.M. [9] proposed an upper bound for the vector-to-vector regression MAE based on DNNs, which is valid with or without the "over-parameterization" technique. Single-channel speech separation is the process of recovering multiple speaker speech signals from a one-dimensional mixture of speech. Moreover, the training goal of our speech separation system was to maximize the scale-invariant signal-to-noise ratio (SI-SNR). Signal-to-noise ratio improvement (SDRi) was used as the main objective metric to assess separation accuracy. The specific calculation formula is presented in Part 4 of the text. In 2014, Wang De Liang et al. presented a supervised training method for the speech separation task [10], comparing and analyzing the effects of different time-frequency features and different time-frequency masking matrices on the speech separation task, including Ideal Binary Mask (IBM) and Ideal Ratio Mask (IRM), and compared with the NMF-based method for speech. In 2016, Deep Clustering (DPCL) [11] was proposed by mapping each time-frequency point to a high-dimensional space, and taking the mean square error between the affinity matrix of the corresponding high-dimensional vector and the affinity matrix of the label composition vector as the training target. It has an SDRi value of 10.8. This is a clever solution to the label substitution problem and allows effective speech separation even when the number of speakers in the testing phase of the network is different from that in the training phase by setting a different number of clustering centers. In 2017, Yu D et al. [12] proposed a permutation invariant training (PIT) algorithm, which first calculates all combinations of network outputs and labels, and also selects the smallest of these results as the value of the loss function, solving the label replacement problem in this way. It has an SDRi value of 10.9. In 2017, Y. Luo et al. presented the Deep Attractor Network (DANet) [13], which uses IRM to weight the embedding vector corresponding to each time-frequency point in the training phase to find prime cluster points and generate time-frequency masks based on the distance between different prime points and each time-frequency point. In 2019, the literature [14] presented a computerized auditory scene analysis system based on deep learning (CASA), which combines the respective advantages of PIT and DPCL for speech separation through two stages of training, with the first stage using PIT to separate speech at the time-frame level and the second stage referring to DPCL for clustering and combining separated speech from different time-frames.

In terms of time domain speech separation, Y. Luo et al. presented a time domain speech separation network (Tas-Net) [15], which directly uses speech time domain information as input to the network, encodes and decodes the speech using a convolutional layer with thresholding, and then separates the speech by estimating a mask matrix of encoded features using a separator consisting of an LSTM network. It has an SDRi value of 10.8. In 2019, Y. Luo improved on the Tas-Net network model by replacing the LSTM with a Temporal Convolutional Network (TCN) to propose a fully-convolutional time-domain speech separation network (Conv-TasNet) [16], which replaces the separator LSTM of the Tas-Net network with a TCN, which can obtain a large perceptual field with a small number of parameters, thus better capturing long-distance contextual information. It can reach an SDRi value of 15.3. In 2020, Y. Luo proposed the Dual-Path Recurrent Neural Network (DPRNN) [17], which performs speech separation by partitioning long sequences into smaller blocks and applying intra- and inter-block RNNs. In 2021, the literature [18] presented the SepFormer network, which performs speech separation by replacing the original RNN network with a Transformer network. In the same year, a self-attentive network with an hourglass shape (Sandglasset) [19] was proposed, which enhances speech separation network performance with a smaller model size and computational cost by

focusing on multi-granularity features and capturing richer contextual information. Moreover, it has been relatively effective, with an SDRi value of 21.0. In 2022, the SepIt [20] speech separation network was proposed, which improves the performance of speech separation networks by iteratively improving the estimation of different speakers. This is the latest method that came out with a relatively high SDRi value of 22.4.

The process of language identification is a process of classification and judgment, and the key is the acquisition of valuable features and the construction of language models. For a general language identification system, the most important thing is the recognition rate, the accuracy of the recognition results must be guaranteed, and this is the starting point of all evaluation metrics. The following will introduce the more commonly used performance evaluation metrics, which are Accuracy, Precision, Recall, and F1 value. Moreover, the specific formulas for calculating these indicators are described in Part 4 of our text. The traditional methods of language identification are mainly based on the Gaussian mixed model and identity vector (i-vector) based components. The Gaussian hybrid-universal background model (GMM-UBM) approach was proposed in the literature [21], requiring enormous data to estimate the covariance matrix. In the literature [22], a Gaussian mixed model-support vector machine (GMM-SVM) mean super vector classification algorithm was proposed, which has improved the recognition performance compared to the GMM-UBM method. The literature [23,24] used i-vector features extracted from audio for language identification, which effectively enhanced the recognition results and became one of the primary language identification methods.

As deep learning is widely used in various tasks, researchers have also started to apply deep neural networks to language identification research. The literature from 2014 [25] proposed the first large-scale application of DNN models to short-time speech segment language tasks for language identification at the speech frame-level feature level. In addition, Convolutional Neural networks (CNN), Recurrent Neural networks (RNN), and Long Short-Term Memory (LSTM) have also been applied to language identification, resulting in a breakthrough improvement in language identification performance [26–29]. Subsequently, Geng W, Raffel C, Mounika K V, et al. [30–32] proposed an end-to-end language identification framework based on the attention mechanism, which improves the effectiveness of language identification by obtaining more valuable information on speech features for language discrimination through the attention mechanism. In 2018, Snyder D [33] proposed an x-vector language identification method, which outperformed the i-vector. In 2019, Cai W et al. [34] proposed an attention-based CNN-BLSTM model that performs language recognition at the discourse level and can obtain discourse-level decisions directly from the output of the neural network. In 2020, Aitor Arronte Alvarez et al. [35] proposed an end-to-end network Res-BLSTM language identification method combining Residual Block and BLSTM network. However, most deep learning-based studies of single-channel speech separation rely on a single metric to evaluate the system, and the evaluation metric is relatively homogeneous. In our experiments, we indirectly analyzed the performance of the front-end network by looking at the evaluation metrics or the good or bad performance of the downstream tasks.

In this paper, based on the above research on single-channel speech separation and language identification, the speech separation task was applied to a complex multi-speaker multi-lingual scenario to obtain the corresponding monolingual information of each speaker to facilitate subsequent speech recognition or other operations. The experimental results obtained from the language identification network can also be used as an intermediate reference index for the final single-channel speech separation network to optimize our separation network.

3. Model Architecture

3.1. Single Channel Speech Separation

This paper used Conv-TasNet, a fully convolutional audio time-domain separation network proposed in [16] for the single-channel speech separation task. The input and

output of the model are time-domain waveforms, which avoids the problem of poor separation due to the difficulty of estimating the phase spectrum in frequency domain speech separation. The network uses 1D convolution and 1D transpose convolution as the coding and decoding layers. Between the coding and decoding layers, the Temporal Convolutional Network [36] is used for long-term time series modeling of features, which has the advantage of solving the possible gradient disappearance problem of RNN in modeling long-term time-series data. The advantage of TCN is that it can solve the gradient disappearance problem that may occur when modeling long-term time series data in RNN and obtain a sizeable perceptual field with few parameters.

Conv-TasNet consists of three processing stages, Encoder, Separator, and Decoder, as shown in Figure 1. First, the encoder module converts short segments of the mixed signal into short-time features in the intermediate feature space. Then, a separator combines the speaker-specific feature vectors to estimate the speaker-specific waveform mask. Finally, the short-time features from the encoder and the waveform mask from the separator are dot-multiplied, and the decoder module reconstructs the speaker-specific waveform by converting the masked encoder features. Moreover, the network applies utterance-level permutation invariant training (uPIT) during training to solve the label alignment problem.

① Encoder: The encoder part uses a time-domain hybrid waveform as input, and the extraction of time-domain short-time features is performed by one-dimensional convolution:

$$h = RELU(conv1d(x)). \tag{1}$$

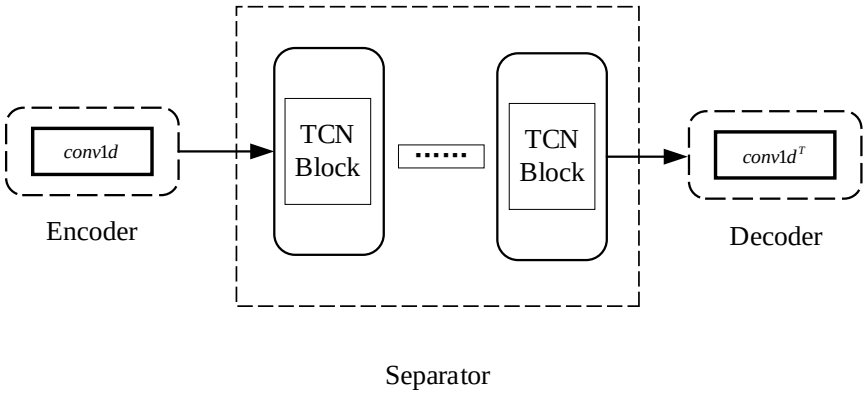

Figure 1. Network architecture of Conv-TasNet.

The convolution kernel, with step sizes set to 160 and 80, respectively.

② Separator: The separator uses TCN in which each layer of the network is connected by a one-dimensional null convolutional network and a RELU activation function. Null convolution allows arbitrarily large sensory fields to capture multi-scale contextual information without additional parameters. Furthermore, the expansion coefficients of each layer are 1, 2, 4, 8, 16, 32, 64 from the first layer onwards, and residual blocks connect the layers. Because each layer is an inflated convolution, the final output node of the network can contain the time-frequency feature information of all previous input nodes, and each layer has a larger perceptual field compared to a convolutional neural network.

$$m_x = TCN(h, e). \tag{2}$$

③ Decoder: A one-dimensional transposed convolution is used in the decoder, with the same convolution kernel and step size as in the encoder. The input is the result of a dot product of the short-time features at the output of the decoder and the target speaker waveform mask at the outcome of the separator.

$$\hat{x} = conv1d(mx \cdot h). \tag{3}$$

where $conv1d^T$ is the one-dimensional transposed convolution, convolution kernel and step size are the same as the one-dimensional convolution in the encoder.

3.2. Language Identification

In language identification research, extracting valuable speech features is the most critical step, followed by constructing a suitable classification model. Deep learning has better advantages in feature extraction and model building, as it can automatically extract features and build optimal models and save them during the training process to provide a basis for subsequent classification judgments. In the language identification task, the CNN-CBAM-BLSTM neural network structure is used to classify and recognize multiple languages. Spectral images are typically processed through numerous convolutional and pooling layers, as well as one or two fully connected layers. The CNN model is fed with a speech spectral map, and then feature extraction is performed by convolutional operations. As the model deepens, the local features extracted from the shallow layers are continuously processed and integrated to obtain the deeper, higher-dimensional parts. However, for tasks related to time series, the performance of CNN can be relatively inferior. To solve such problems, RNN networks are introduced, which can handle time-series tasks well. LSTM networks are representative of RNNs. BLSTM consists of forwarding and backward LSTMs and is a bi-directional network structure. In this case, the forward LSTM learns information before the current moment, and the backward LSTM learns information after the present moment so that this network can learn the temporal contextual information contained in the speech sequences, thus compensating for the shortcomings of the CNN network. Therefore, the CNN-BLSTM network based on the strengths of both networks can be combined and used to extract local features and temporally relevant features. Then the CBAM attention module can be used to focus on global information to obtain the features that contribute more to the linguistic information. Our language identification system recognizes the five languages simultaneously based on the randomly mixed input speech spectra or the single language speech spectra separated by the separation network, and scores them to calculate the final accuracy, recall, and other evaluation index information. The structure of this network is shown in Figure 2.

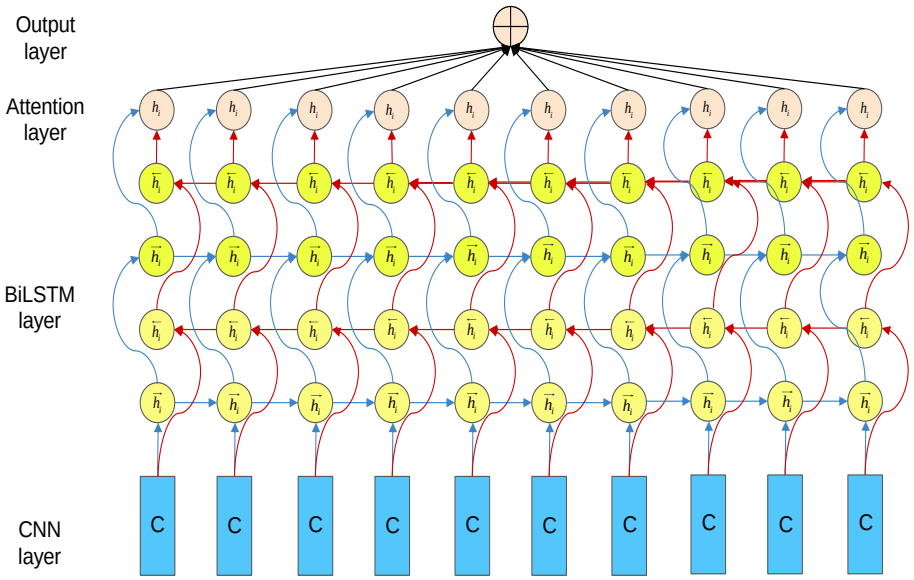

Figure 2. Overall network structure for language identification.

3.2.1. Convolutional Neural Network (CNN)

In this paper, the CNN-CBAM-BLSTM network was used to implement language identification, where the structure of the CNN neural network is shown in Figure 3. The pooling layer performs pooling operations on the output feature map to reduce the size

of the feature map to reduce the number of parameters in the network. Finally, the classification layer is used to classify the language.

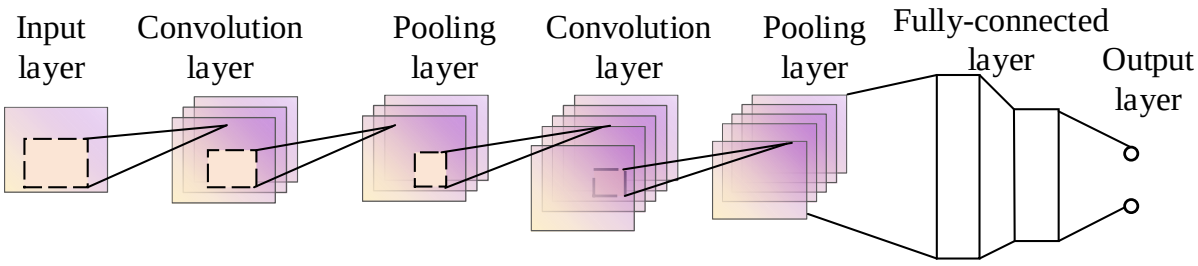

Figure 3. General structure of CNN.

3.2.2. Bidirectional Long- and Short-Term Memory Network (BLSTM)

The BLSTM consists of a forward and a backward LSTM, a bi-directional network structure. The forward LSTM learns messages before the current moment, and the backward LSTM learns messages after the current moment, so the network can learn the temporal background information contained in the speech sequence, thus making up for the shortcomings of CNN networks. The BLSTM network [37] consists of four parts: the input layer, the forward LSTM, the backward LSTM, and the output layer, and its network structure is shown in Figure 4.

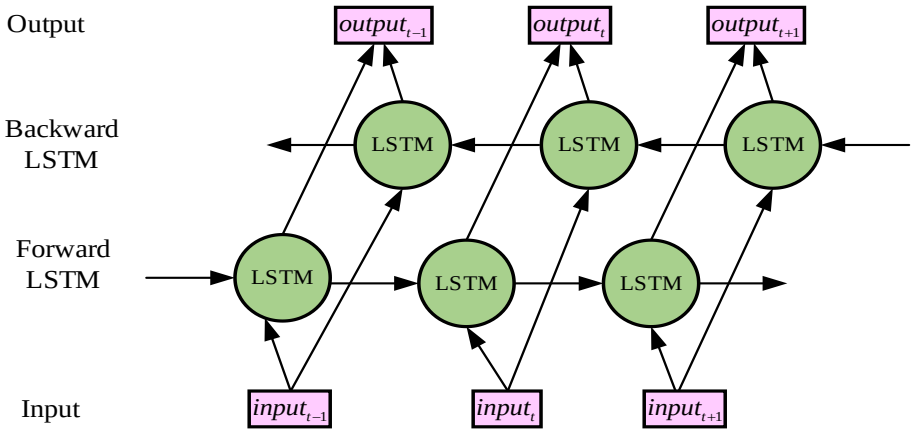

Figure 4. BLSTM network structure.

In the BLSTM network structure in Figure 4, $input_{t-1}$, $input_t$, $input_{t+1}$ denote the inputs at moments $t-1$, t, $t+1$, respectively, and $output_{t-1}$, $output_t$, $output_{t+1}$ denote the outputs corresponding to moments $t-1$, t, $t+1$, respectively. The forward LSTM refers to the calculation of the output corresponding to the forward moments along the forward order of the moments. Backward LSTM means calculating the output corresponding to the reverse moment along the reverse order of moments, and finally the output of both together as the final output at the corresponding moment.

3.2.3. Convolutional Block Attention Module (CBAM)

The features extracted by CNN and BLSTM from the language spectral map do not all contribute equally to the representation of language information, but rather some features contribute minimally, and some contribute more. Therefore, following CNN and BLSTM, an attention layer is used to selectively focus on language-related features and produce a differentiated feature representation for language identification. The advantage is that using an attention mechanism to reflect the importance of a series of high-level features to the final language differentiation, rather than simply completing the stack of features over time.

The attention layer is located after the bidirectional LSTM. The output of the bidirectional LSTM is first passed through a softmax function to calculate the normalized weight α_t, calculated as shown in (4). The normalized weight α_t is then weighted and summed over h_t to obtain the language representation c, as shown in (5).

$$\alpha_t = \frac{\exp(W \cdot h_t)}{\sum_{t=1}^{T} \exp(W \cdot h_t)} \tag{4}$$

$$c = \sum_{t=1}^{T} \alpha_t h_t. \tag{5}$$

where W denotes the weight value, h_t is the state at the current moment and the language representation c will be derived by computing Equations (4) and (5) and passed to the all-connected layer to obtain a more profound representation of the language, and a softmax classifier that maps the language representation to N different spaces for classification, where N denotes the number of classes of the language.

The general network architecture of the convolutional chunk attention module is shown in Figure 5.

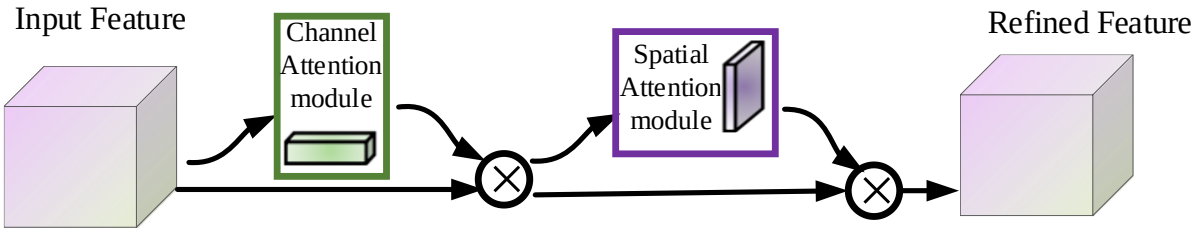

Figure 5. Convolutional Block Attention Module.

Given a feature map, the CBAM module can serially generate attentional feature map information in both the channel and spatial dimensions, and then the information from the two feature maps is multiplied with the previous original input feature map for adaptive feature correction to produce the final feature map.

As shown in the figure above, there is an input, a channel attention module, a spatial attention module, and an output. The input features $F \in R^{C*H*W}$, followed by the channel attention module $M_c \in R^{C*1*1}$, multiply the result of the convolution by the original image, and the output of the channel attention module is used as input for the two-dimensional convolution of the spatial attention module $M_s \in R^{1*H*W}$, and then the output is multiplied by the original image.

$$F' = M_c(F) \otimes F. \tag{6}$$

$$F'' = M_s(F') \otimes F'. \tag{7}$$

Equation (6) focuses on the features on the channel, by keeping the channel dimension constant and compressing the spatial dimension, focusing on the meaningful information in the input image. Moreover, Equation (7) focuses on features on space by keeping the spatial dimension constant, compressing the channel dimension, and focusing on the location information of the target.

4. Experimental Settings

4.1. Dataset

Speech Separation: This paper used the WSJ0-2mix dataset to compare the performance of two-speaker speech separation. It contained 30 h of training, 10 h of validation, and 5 h of test data. Mixed speech in WSJ0-2mix was generated by randomly selecting different speakers and sentences in the Wall Street Journal (WSJ0) training set si_tr_s and combining them with a random signal-to-noise ratio (SNR) between −5 dB and 5dB. Cor-

rections in the test set were derived from 16 speakers in the WSJ0 dataset si_dt_05 and si_et_05 that were not used in training. All speech in WSJ0-2mix was resampled to 8000 Hz.

Language identification: The language identification model was trained and tested on the AP20_OLR [38] Oriental Language Dataset provided by Speech Ocean (China). This dataset was provided by the AP20_OLR competition. In this paper, five datasets, including Mandarin Chinese (zh-cn), Vietnamese (vi-vn), Indonesian (id-id), Japanese (ja-jp), and Korean (ko-kr) were used. One thousand eight hundred speech data were extracted from each language and divided according to the ratio of 7:2:1 (training set: validation set: test set) to construct a dataset. Moreover, we assigned numeric labels to each language in the language recognition task, for example, "0": "id-id", "1": "ja-jp", "2": "ko-kr", "3": "vi-vn", "4": "zh-cn". The structure of this dataset is presented in Table 1.

Table 1. Oriental Language Dataset Structure.

Language	Train	Validation	Test	Total
zh-cn	1260	360	180	1800
id-id	1260	360	180	1800
ja-jp	1260	360	180	1800
ko-kr	1260	360	180	1800
vi-vn	1260	360	180	1800

4.2. Evaluation Metrics

4.2.1. Speech Separation

In this experiment, the scale invariant signal-to-noise ratio (SI-SNR) and SDR were used to evaluate the separation results, with larger values being better. The SI-SNR equation is defined as:

$$\begin{cases} s_{target} = \frac{\langle \hat{s}, s \rangle s}{\|s\|^2} \\ e_{noise} = \hat{s} - s_{target} \\ SI - SNR = 10 \log_{10} \frac{\|s_{target}\|^2}{\|e_{noise}\|^2}. \end{cases} \tag{8}$$

where $\hat{s} \in R^{1*T}$ and $s \in R^{1*T} s$ r$\hat{s}$ represent the estimated and original clean sources, respectively, <,> represents the inner product of these two signals, and $\|s\|^2 =< s, s >$ indicates signal power. Scale invariance can be ensured by normalizing and to zero mean prior to calculation.

Scale-invariant signal-to-noise ratio improvement (SI-SNRi) and signal-to-noise ratio improvement (SDRi) were used as objective metrics to assess separation accuracy. Theorem-type environments (including propositions, lemmas, corollaries, etc.) can be formatted as follows:

$$SDR = 10 \log_{10} \frac{\|s\|^2}{\|\hat{s} - s\|^2}. \tag{9}$$

among them $\hat{s}$ is the estimated speech, s is the clean component of the estimated speech, and $\hat{s} - s$ is the noise component.

4.2.2. Language Identification

For a general language identification system, the most important thing is the identification rate, and the accuracy of the results must be guaranteed; this is the starting point for all evaluation metrics. Some of the more commonly used performance evaluation metrics are *Acc*, *Precision*, *Recall*, *Specificity*, and *F*1. Before we classify a language, there are P target languages and N non-target languages in the test set. After the language identification system classifies the test set, T samples are identified as target languages, and F samples are identified as non-target languages. Based on the actual number of target languages in the test set and the classification of the language identification system, we can obtain the confusion matrix for the target and non-target languages, as shown in Table 2.

Table 2. Results of Statistical Identification.

Confusion Matrix for Target and Non-Target Languages		Predicted Value		Total
		Target Languages	Non-Target Languages	
True value	Target languages	TP	FP	P
	Non-target languages	TN	FN	N
Total		T	F	-

According to Table 2, the calculation formulas of Accuracy, Precision, Recall, Specificity and F1 value can be obtained, as shown below:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{13}$$

$$Specificity = \frac{TN}{FP + TN} \tag{14}$$

4.3. Network Parameters

This paper utilized the Pytorch framework to conduct experiments on an NVIDIA GeForce GTX 3090 GPU.

Our speech separation network was trained on a 4-s segment of 100 epochs. The initial learning rate was set to 1e-3, and if the accuracy of the validation set did not improve within three consecutive periods, the learning rate was halved. The optimizer was used with the Adam optimizer. There was a 50% overlap between consecutive frames in the convolutional autoencoder, because the convolution step set in this 1D convolution was 1/2 the size of the convolutional kernel. During training, the gradient clipping used a maximum L2 norm of 5.

The input to our language identification network was a 224 × 224 speech spectrogram with a batch size of 32 × 32, an initial learning rate of 0.0001, an epoch of 30 iterations, an Adam optimizer, and a cross-entropy loss function.

5. Results and Analysis

5.1. Retraining the Language Identification Network

To test the performance of the separation network in separating single speaker speech signals in a complex overlapping speech environment, we trained the language identification network with a mixed multi-person multi-speech spectrogram. For mixing audio in different languages, first, we made each audio have the same length before mixing, so that the duration of the audio was 4 s. Then, we set the weight of the target language to 1.2 and the weight of the non-target language to 1, so that the target language had a higher energy to be recognized more accurately. Finally, mixing was performed with different overlap rates.

(1) We retrained and tested the language identification network with a multi-person, multi-lingual mixed audio speech spectrum. That is, we replaced the single-person single-speech audio frequency spectrogram in the original dataset with the multi-person multi-language hybrid audio speech spectrogram. We replaced the id-id speech frequency spectrogram in the original dataset with the hybrid audio spectrogram obtained by mixing with ja-jp when the target language was id-id. At the same time,

we replaced the ja-jp audio speech spectrogram in the original dataset with the hybrid audio spectrogram obtained by mixing with id-id when the target language was ja-jp. Similarly, when the target language was ko-kr, we replaced the ko-kr audio spectrogram in the original dataset with the hybrid audio spectrogram obtained by mixing with vi-vn. Likewise, when the target language was vi-vn, we replaced the vi-vn audio spectrogram in the original dataset with the hybrid speech spectrogram obtained by mixing with zh-cn. When the target language was zh-cn, the hybrid audio spectrum obtained by mixing with vi-vn was used to replace the zh-cn speech spectrogram in the original dataset. The accuracy and loss curves of this dataset are shown in Figure 6.

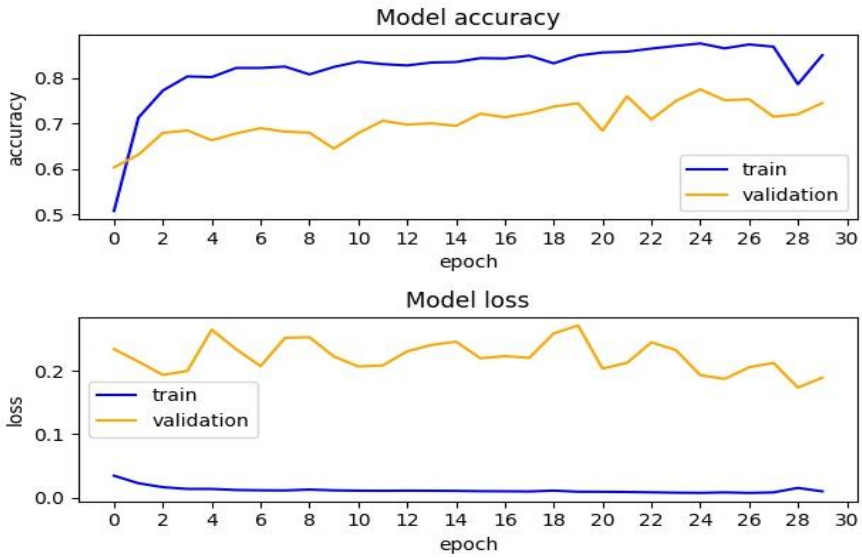

Figure 6. Accuracy and loss curves of the model after training with input mixed speech spectrograms.

(2) We took the mixed multi-lingual audio in (1) above, separated it with a single-channel speech separation network, and later retrained the network with the separated single-person single-speech frequency spectrograms, and through training, the accuracy and loss curves on this dataset are shown in Figure 7.

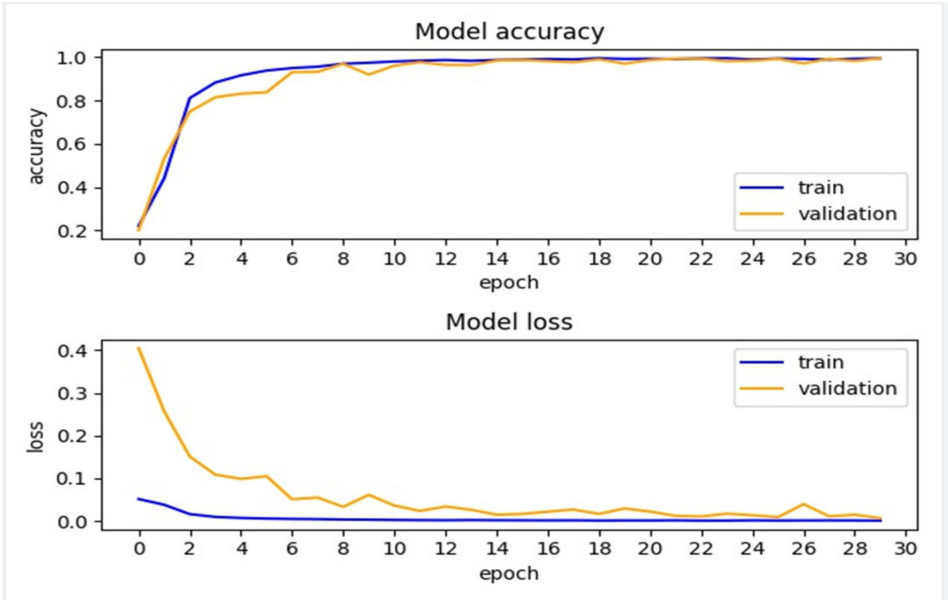

Figure 7. Accuracy and loss curves of the trained model after input separated speech spectrograms.

The experimental results are shown in Figures 6 and 7, which reflect the accuracy and loss variations on the training and validation sets, respectively. The accuracy and loss curves on the overlapping multi-person multi-lingual dataset are shown in Figure 6. Moreover, the accuracy and loss curves on the dataset after separation through the separation network are shown in Figure 7. The blue curves represent the accuracy and loss curves on the training set, and the red curves represent the accuracy and loss curves on the validation set. From the figures, it can be seen that on the overlapping multi-person multi-lingual dataset, the accuracy and loss curves fluctuated and vibrated, the model accuracy was relatively low, and the loss was relatively high, while on the dataset after separation through the separation network, the accuracy and loss curves were smoother, the model accuracy also improved, and the loss also decreased, which also reflects that the CNN-BLSTM model based on the attention mechanism through the separation network separation obtained better results for the dataset.

5.2. Training Language Identification Network with the Original Dataset

(1) To further analyze the performance of the single-channel speech separation network in separating the speech signals of the target language speakers in an overlapping multi-person multi-lingual environment, we evaluated the model trained with the original data. We retained the training and validation sets in the original dataset and replaced only the test set with a multi-person multi-lingual mixed speech spectrogram. The single-person monolingual speech frequency spectrograms of id-id, ja-jp, ko-kr, vi-vn, and zh-cn in the test set were replaced with the corresponding mixed speech spectrograms described in 5.1 The language identification system compared and matched the features of the input language with the already trained language feature information, gave the corresponding labels for the language, realized the classification of the language by the softmax classifier, and then calculated the accuracy, recall and other information for each language.

In the course of the experiment, we set the weight of the target language to 1.2 and the weight of the non-target language to 1 for mixed speech, because we considered realistic multi-lingual overlapping speech scenarios, where a person may be suddenly interrupted while vocalizing, resulting in two or more people speaking at the same time and generating background noise of the human voice. At this point, we gave the person who was vocalizing a higher weight, i.e., a higher energy than the other speaker as noise, so that the linguistic information could be recognized more easily and correctly. The results of the language identification network tests are shown in the table below when the speech of different languages was mixed at different overlap rates and the mixed audio was fed into the language recognition network.

① In this section, our paper replaced all the original datasets in the language identification test set by replacing the speech spectrograms of id-id, ja-jp, ko-kr, vi-vn, zh-cn in the test set with a mixture of id-id and ja-jp, ko-kr and vi-vn, vi-vn and zh-cn. The target language weight of the replacement was 1.2, when the overlap ratio was set to 1, the model accuracy rate was 0.64, and the identification results of the language identification network are shown in Table 3:

Table 3. Experimental results of mixed language audio with an input overlap rate of 1.

Language	Precision	Recall	Specificity
zh-cn	0.637	0.606	0.667
id-id	0.584	0.594	0.718
ja-jp	0.663	0.700	0.677
ko-kr	0.438	0.733	0.702
vi-vn	0.684	0.644	0.791

② When the overlap ratio was set to 0.6, the model accuracy rate was 0.72, and the identification results of the language identification network are shown in Table 4:

Table 4. Experimental results of mixed language audio with an input overlap rate of 0.6.

Language	Precision	Recall	Specificity
zh-cn	0.837	0.683	0.967
id-id	0.992	0.694	0.818
ja-jp	0.982	0.710	0.997
ko-kr	0.438	0.933	0.7
vi-vn	0.991	0.644	0.994

③ When the overlap ratio was set to 0.2, the model accuracy rate was 0.81, and the identification results of the language identification network are shown in Table 5:

Table 5. Experimental results of mixed language audio with an input overlap rate of 0.2.

Language	Precision	Recall	Specificity
zh-cn	0.91	0.672	0.983
id-id	0.993	0.761	0.999
ja-jp	0.995	0.711	0.997
ko-kr	0.462	0.972	0.717
vi-vn	0.959	0.653	0.993

(2) Firstly, the performance of the speech separation network is shown in Table 6.

Table 6. Performance of the speech separation network.

Model	SDRi
Conv-TasNet	15.3

This part separated the mixed audios of different languages combined according to different overlap rates in the above (1) with a single-channel speech separation network and then obtained the corresponding monolingual single-speaker audios, respectively. In the multi-speaker mixed language spectrum in (1), the graphs are replaced with the spectrograms of the related single-person monolingual audio after separation in turn. At this time, the test results of the language identification network were as follows:

① When the overlap ratio was set to 1, the model accuracy rate was 0.72, and the identification results of the language identification network are shown in Table 7:

Table 7. Experimental results of the separated speech audio with an input overlap of 1.

Language	Precision	Recall	Specificity
zh-cn	0.837	0.710	0.967
id-id	0.992	0.994	0.818
ja-jp	0.995	0.721	0.997
ko-kr	0.534	0.943	0.802
vi-vn	0.991	0.655	0.996

② When the overlap ratio was set to 0.6, the model accuracy rate was 0.81, and the identification results of the language identification network are shown in Table 8:

Table 8. Experimental results of the separated speech audio with an input overlap of 0.6.

Language	Precision	Recall	Specificity
zh-cn	0.923	0.825	0.986
id-id	0.994	0.786	0.994
ja-jp	0.996	0.821	0.997
ko-kr	0.543	0.946	0.796
vi-vn	0.976	0.774	0.994

③ When the overlap ratio was set to 0.2, the model accuracy rate was 0.93, and the identification results of the language identification network are shown in Table 9:

Table 9. Experimental results of the separated speech audio with an input overlap of 0.2.

Language	Precision	Recall	Specificity
zh-cn	0.945	0.806	0.993
id-id	0.996	0.894	0.997
ja-jp	0.998	0.889	1.0
ko-kr	0.713	0.989	0.991
vi-vn	0.977	0.921	0.997

Table 3, Table 4, Table 5, Table 7, Table 8 and Table 9, show the recognition results for the input mixed audio spectrum and the input audio spectrogram through the separation network in the language identification network, respectively. From these six tables, it can be seen that when the audio spectrograms separated through the separation network were inputted into the language identification network, the overall language identification results were much better than when the mixed speech spectrograms are input. The model accuracy of the language recognition network varied when the overlap rate of the mixed audio varied. Specifically, when the overlap rate was 20%, the recognition of the language identification network was better and the model accuracy was 12% better when inputting the separated speech spectrograms than when inputting the mixed speech spectrograms; when the overlap rate was 60%, the model accuracy was 9% better when inputting the separated speech spectrograms than when inputting the mixed speech spectrograms. When the overlap rate was 100%, the model accuracy was 6% better when inputting the separated speech spectrogram than when inputting the mixed speech spectrogram. Thus, it can be seen that the speech separation network can better separate the audio information of a single speaker in a mixed language in the case of multiple speakers, which also illustrates the value of our separation network in complex language environments.

6. Conclusions and Future Work

The speech separation task is associated with the language recognition task. In this paper, speech separation networks were used in complex multi-lingual and multi-speaker overlapping speech scenarios to more efficiently and accurately extract monolingual information corresponding to each speaker and to distinguish between language categories. In this paper, a single-channel speech separation network was trained using the WSJ0-2mix dataset. Then the language identification network was trained and tested using the Oriental Language Dataset and a dataset manually mixed at different overlap rates. Finally, the multi-person multi-lingual frequency speech spectrum mixed at different overlap rates and the single-person monolingual frequency speech spectrum separated by the Conv-TasNet network were sequentially fed into the language recognition system in the testing phase. The experimental results showed that the recognition results corresponding to each language were significantly higher in terms of accuracy and recall when the single-person monolingual spectrum separated by the separation network was fed into the language recognition network than when it was fed into the mixed audio spectrum. With this paper, the experimental results obtained from the language recognition network can be used as

an intermediate reference metric for optimizing our final single-channel speech separation network and can be used in complex overlapping speech scenarios.

In future work, we will consider trying to train the Conv-TasNet single channel speech separation and the language identification network based on the attention mechanism and BLSTM in cascade, constructing the language identification dataset in the same form as the speech separation dataset. Moreover, we will use this new dataset to train the two networks together and input the feature spectrograms of the separated speech output from the speech separation network directly into the language identification network to construct an end-to-end deep learning system. This end-to-end system can be better applied to multilingual overlapping speech scenarios while improving the performance of the separation and recognition networks.

Author Contributions: Conceptualization, Z.A. and M.A.; methodology, Z.A., H.Y. and A.H.; software, Z.A. and H.Y.; validation, M.A., H.Y. and A.H.; formal analysis, Z.A. and H.Y.; investigation, Z.A. and M.A.; resources, M.A., H.Y. and A.H.; data curation, H.Y.; writing—original draft preparation, Z.A.; writing—review and editing, M.A. and A.H.; visualization, Z.A., M.A. and H.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Strengthening Plan of National Defense Science and Technology Foundation of China (grant number 2021-JCJQ-JJ-0059) and Natural Science Foundation of China (grant number U2003207).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The mixed speech in WSJ0-2mix used in this paper for the single-channel speech separation task was generated by mixing in the Wall Street Journal (WSJ0) training set. This publicly available dataset can be found here: https://catalog.ldc.upenn.edu/LDC93S6A (accessed on 20 July 2022). The language recognition model was trained and tested on the AP20 OLR Oriental language dataset provided by Speech Ocean (China). This publicly available dataset can be found here: http://index.cslt.org/mediawiki/index.php/OLR_Challenge_2020 (accessed on 20 July 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yurong, H. Research on Speech Language Recognition Technology Based on Deep Learning Network. Master's Thesis, Northwestern University, Xi'an, China, 2021.
2. Bell, A.J.; Sejnowski, T.J. An information-maximization approach to blind separation and blind deconvolution. *Neural Comput.* **1995**, *7*, 1129–1159. [CrossRef]
3. Wang, X.; Jiang, Z.; Zhang, Y. Deep clustering of speaker speech separation based on temporal convolutional networks. *Comput. Eng. Des.* **2020**, *41*, 2630–2635.
4. Luo, Y.; Wang, J.; Xu, L.; Yang, L. Multi-Stream Gated and Pyramidal Temporal Convolutional Neural Networks for Audio-Visual Speech Separation in Multi-Talker Environments. In Proceedings of the Interspeech 2021, Brno, Czechia, 30 August–3 September 2021; ISCA: Beijing, China, 2021; pp. 1104–1108.
5. Allen, F.; Ambikairajah, E.; Epps, J. Warped magnitude and phase-based features for language identification. In Proceedings of the 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, Toulouse, France, 14–19 May 2006; Volume 1, p. I.
6. Wang, S.-J.; Meng, M.; Liang, J.L.; Xu, B. Multilingual-based phoneme identification and its application to language recognition. *J. Tsinghua Univ. Nat. Sci. Ed.* **2008**, *S1*, 678–682.
7. Qi, J.; Du, J.; Siniscalchi, S.M.; Ma, X.; Lee, C.H. Analyzing upper bounds on mean absolute errors for deep neural network-based vector-to-vector regression. *IEEE Trans. Signal Process.* **2020**, *68*, 3411–3422. [CrossRef]
8. Qi, J.; Du, J.; Siniscalchi, S.M.; Ma, X.; Lee, C.H. On mean absolute error for deep neural network-based vector-to-vector regression. *IEEE Signal Process. Lett.* **2020**, *27*, 1485–1489. [CrossRef]
9. Siniscalchi, S.M. Vector-to-vector regression via distributional loss for speech enhancement. *IEEE Signal Process. Lett.* **2021**, *28*, 254–258. [CrossRef]
10. Wang, Y.; Narayanan, A.; Wang, D. On training targets for supervised speech separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 1849–1858. [CrossRef]

11. Hershey, J.R.; Chen, Z.; Le Roux, J.; Watanabe, S. Deep clustering: Discriminative embeddings for segmentation and separation. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 31–35.
12. Yu, D.; Kolbæk, M.; Tan, Z.H.; Jensen, J. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 241–245.
13. Chen, Z.; Luo, Y.; Mesgarani, N. Deep attractor network for single-microphone speaker separation. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 246–250.
14. Liu, Y.; Wang, D.L. Divide and conquer: A deep CASA approach to talker-independent monaural speaker separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 2092–2102. [CrossRef]
15. Luo, Y.; Mesgarani, N. Tasnet: Time-domain audio separation network for real-time, single-channel speech separation. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 696–700.
16. Luo, Y.; Mesgarani, N. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 1256–1266. [CrossRef] [PubMed]
17. Luo, Y.; Chen, Z.; Yoshioka, T. Dual-path rnn: Efficient long sequence modeling for time-domain single-channel speech separation. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 46–50.
18. Subakan, C.; Ravanelli, M.; Cornell, S.; Bronzi, M.; Zhong, J. Attention is all you need in speech separation. In Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 21–25.
19. Lam MW, Y.; Wang, J.; Su, D.; Yu, D. Sandglasset: A light multi-granularity self-attentive network for time-domain speech separation. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 5759–5763.
20. Lutati, S.; Nachmani, E.; Wolf, L. SepIt Approaching a Single Channel Speech Separation Bound. *arXiv* **2022**, arXiv:2205.11801.
21. Torres-Carrasquillo, P.A.; Singer, E.; Kohler, M.A.; Greene, R.J. Approaches to language identification using Gaussian mixture models and shifted delta cepstral features. In Proceedings of the 7th International Conference on Spoken Language Processing, ICSLP2002-INTERSPEECH 2002, Denver, CO, USA, 16–20 September 2002.
22. Campbell, W.M.; Singer, E.; Torres-Carrasquillo, P.A.; Reynolds, D.A. Language recognition with support vector machines. In Proceedings of the ODYSSEY04-The Speaker and Language Recognition Workshop, Toledo, Spain, 31 May–3 June 2004.
23. Dehak, N.; Torres-Carrasquillo, P.A.; Reynolds, D.; Dehak, R. Language recognition via i-vectors and dimensionality reduction. In Proceedings of the Twelfth Annual Conference of the INTERNATIONAL speech Communication Association, Florence, Italy, 28–31 August 2011.
24. Ramoji, S.; Ganapathy, S. Supervised I-vector modeling for language and accent recognition. *Comput. Speech Lang.* **2020**, *60*, 101030. [CrossRef]
25. Lopez-Moreno, I.; Gonzalez-Dominguez, J.; Plchot, O.; Martinez, D.; Gonzalez-Rodriguez, J.; Moreno, P. Automatic language identification using deep neural networks. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 5337–5341.
26. Gelly, G.; Gauvain, J.L. Spoken Language Identification Using LSTM-Based Angular Proximity. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017; pp. 2566–2570.
27. Zazo, R.; Lozano-Diez, A.; Gonzalez-Dominguez, J.; Toledano, D.T.; Gonzalez-Rodriguez, J. Language identification in short utterances using long short-term memory (LSTM) recurrent neural networks. *PLoS ONE* **2016**, *11*, e0146917. [CrossRef] [PubMed]
28. Lopez-Moreno, I.; Gonzalez-Dominguez, J.; Martinez, D.; Plchot, O.; Gonzalez-Rodriguez, J.; Moreno, P.J. On the use of deep feedforward neural networks for automatic language identification. *Comput. Speech Lang.* **2016**, *40*, 46–59. [CrossRef]
29. Chowdhury, S.A.; Ali, A.; Shon, S.; Glass, J. What does an end-to-end dialect identification model learn about non-dialectal information? In Proceedings of the INTERSPEECH 2020, Shanghai, China, 25–29 October 2020; pp. 462–466.
30. Geng, W.; Wang, W.; Zhao, Y.; Cai, X.; Xu, B. End-to-End Language Identification Using Attention-Based Recurrent Neural Networks. In Proceedings of the Interspeech 2016, San Francisco, CA, USA, 8–12 September 2016; pp. 2944–2948.
31. Raffel, C.; Ellis, D.P.W. Feed-forward networks with attention can solve some long-term memory problems. *arXiv* **2015**, arXiv:1512.08756.
32. Mounika, K.V.; Achanta, S.; Lakshmi, H.R.; Gangashetty, S.; Vuppala, A. An investigation of deep neural network architectures for language recognition in Indian languages. In Proceedings of the INTERSPEECH 2016, San Francisco, CA, USA, 8–12 September; pp. 2930–2933.
33. Snyder, D.; Garcia-Romero, D.; McCree, A.; Sell, G.; Povey, D.; Khudanpur, S. Spoken language recognition using x-vectors. In Proceedings of the Odyssey 2018, Les Sables d'Olonne, France, 26–29 June 2018; pp. 105–111.
34. Cai, W.; Cai, D.; Huang, S.; Li, M. Utterance-level end-to-end language identification using attention-based CNN-BLSTM. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 5991–5995.

35. Alvarez, A.A.; Issa, E.S.A. Learning Intonation Pattern Embeddings for Arabic Dialect Identification. In Proceedings of the Interspeech 2020, Shanghai, China, 25–29 October 2020; pp. 472–476.
36. Pandey, A.; Wang, D.L. TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain. In Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6875–6879.
37. Bizzoni, Y.; Ghanimifard, M. Bigrams and BLSTMs two neural networks for sequential metaphor detection. In Proceedings of the Workshop on Figurative Language Processing, New Orleans, LA, USA, 6 June 2018.
38. Li, Z.; Zhao, M.; Hong, Q.; Li, L.; Tang, Z.; Wang, D.; Song, L.; Yang, C. AP20-OLR Challenge: Three Tasks and Their Baselines. *arXiv* **2020**, arXiv:2006.03473.