electronics

Article

Domain-Adversarial Based Model with Phonological Knowledge for Cross-Lingual Speech Recognition

Qingran Zhan [1], Xiang Xie [1,2,*], Chenguang Hu [1], Juan Zuluaga-Gomez [3,4,5], Jing Wang [1] and Haobo Cheng [1,2]

1 School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China; qingran.zhan@gmail.com (Q.Z.); 3220200551@bit.edu.cn (C.H.); wangjing@bit.edu.cn (J.W.); chb@bit.deu.cn (H.C.)
2 Shenzhen Research Institute, Beijing Institute of Technology, Shenzhen 518063, China
3 Idiap Research Institute, 1920 Martigny, Switzerland; juan-pablo.zuluaga@idiap.ch
4 Ecole Polytechnique Federale de Lausanne (EPFL), 1015 Lausanne, Switzerland
5 Department of Mechatronics Engineering, Universidad Autonoma del Caribe, Barranquilla 080020, Colombia
* Correspondence: xiexiang@bit.edu.cn

Abstract: Phonological-based features (articulatory features, AFs) describe the movements of the vocal organ which are shared across languages. This paper investigates a domain-adversarial neural network (DANN) to extract reliable AFs, and different multi-stream techniques are used for cross-lingual speech recognition. First, a novel universal phonological attributes definition is proposed for Mandarin, English, German and French. Then a DANN-based AFs detector is trained using source languages (English, German and French). When doing the cross-lingual speech recognition, the AFs detectors are used to transfer the phonological knowledge from source languages (English, German and French) to the target language (Mandarin). Two multi-stream approaches are introduced to fuse the acoustic features and cross-lingual AFs. In addition, the monolingual AFs system (i.e., the AFs are directly extracted from the target language) is also investigated. Experiments show that the performance of the AFs detector can be improved by using convolutional neural networks (CNN) with a domain-adversarial learning method. The multi-head attention (MHA) based multi-stream can reach the best performance compared to the baseline, cross-lingual adaptation approach, and other approaches. More specifically, the MHA-mode with cross-lingual AFs yields significant improvements over monolingual AFs with the restriction of training data size and, which can be easily extended to other low-resource languages.

Keywords: cross-lingual automatic speech recognition (ASR); articulatory features; domain-adversarial neural network; multi-stream learning

Citation: Zhan, Q.; Xie, X.; Hu, C.; Zuluaga-Gomez, J.; Wang, J.; Cheng, H. Domain-Adversarial Based Model with Phonological Knowledge for Cross-Lingual Speech Recognition. *Electronics* **2021**, *10*, 3172. https://doi.org/10.3390/electronics10243172

Academic Editors: Daniel Hládek, Matúš Pleva, Piotr Szczuko and Andrej Zgank

Received: 28 October 2021
Accepted: 15 December 2021
Published: 20 December 2021

1. Introduction

Automatic speech recognition (ASR) systems have been improved greatly in recent years due to deep neural networks (DNNs). However, there are more than 7000 living languages in the world, where only about 125 different languages have access to ASR technologies [1], so it is still a big challenge to develop a reliable ASR system for low-resourced languages. The phonological attribute modeling, also known as "acoustic-to-articulatory(-attribute) modeling", is widely used to describe the movement of the organ during speech production and can be shared among all languages. Articulatory information has been proved useful in many related areas, such as pathological speech recognition [2], pronunciation prediction [3] and multilingual speech recognition [4]. There are mainly three methods to derive phonological-based features: (i) using an X-ray radiometer to measure movements of vocal organs [5], (ii) acoustic-articulatory mapping using filtering techniques [6] and, (iii) statistical model based speech attribute detectors [7]. The first approach has a high initial setup cost, thus it is an unfeasible approach, while the second only detects some

of the attributes and not for all the phonological attributes [8]. This research explores the third approach due to its feasibility and reliability. The main advantage of phonological attributes based on cross-lingual ASR is that the phonological knowledge can be shared across different languages.

In low-resource languages, the lack of linguistic knowledge causes a scarcity in transcribed speech data for the training of ASR systems. Therefore, International Phonetic Alphabet (IPA) [9] explores an approach for cross-lingual phonological attributes. The languages tackled in this paper are English, German, French, and Mandarin. In practice, Mandarin is a well-resourced language, however, to verify our proposed framework on cross-lingual speech recognition task, we take Mandarin as a low-resourced language by using a limited dataset.

As shown in Figure 1, our system has two key modules: (i) a domain adversarial based multi-task learning model to extract phonological knowledge-based features (abbreviated as AFs) and, (ii) a framework that fuses the AFs with conventional acoustic features e.g., Mel Frequency Cepstral Coefficients (MFCCs). Firstly, the AFs detector is modeled into a domain adversarial-based multitask learning system (abbreviated as DANN). The DANN model contains a gradient reversal layer that can prevent the model from learning domain information (languages and speakers in this paper). In the proposed method, the domain classifier of DANN is modified as multi-task supervised learning with speaker and language classification, the classification of the AFs is the main task. To combine the MFCCs and AFs, different fusion approaches are considered.

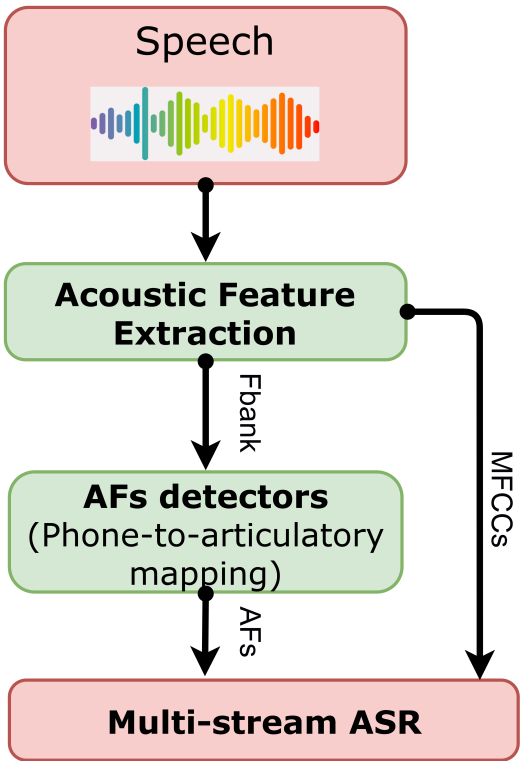

Figure 1. The overview diagram in this paper.

The paper is organized as follows: Section 2 reviews the latest related work on AFs and multi-stream framework. Section 3 gives detailed information regarding AFs detectors and Section 4 provides further elements on multi-stream frameworks. Section 5 reviews the configuration of the proposed experiments. Section 6 presents the results and analysis. Finally, Section 7 gives the conclusions and directions to future research.

2. Related Work

Phonological research demonstrated that each sound unit of a language can be split into smaller phonological units based on articulators used to produce the respective sound. To generate reliable AFs, it is critical to design stable AFs detectors. Several studies found that CNNs have a better ability to capture the articulator's information [7,10]. The combination of articulatory features and conventional acoustic features has been shown useful in many speech tasks. In [11], researchers show that the AFs can improve the ASR performance by combining the MFCCs and AFs at the lattice level. Similarly, the results by [12] work indicates that the combination of MFCCs and AFs at lattice level can improve the performance of pronunciation error detection.

It is also proved that the phonological features can be shared across languages. Chin-Hui Lee et al., trained three attribute detectors on Mandarin speech for three "manner" (articulatory class) features, and further used these detectors to process an English utterance spoken by a non-native Mandarin speaker [13]. In those experiments, both *stops* and *nasals* attributes were correctly detected, which can prove that the speech attribute can be used in cross-lingual speech recognition in English and Mandarin. There are few studies on multilingual speech recognition integrating AFs; Hari Krishna et al., trained a bank of AFs detectors using source language to predict the articulatory features for the target languages, which showed that the combination of AFs using AF-Tandem method performs better than the lattice-rescoring approach [14].

Because English, German, French, and Mandarin are from two different language families, it is not straightforward to implement cross-lingual speech recognition. In [15], the researchers used the English-trained MLPs for AF extraction for Mandarin, after applying PCA (Principal Components Analysis), the AFs tandem features were directly concatenated with MFCCs. Results indicated that the English-trained AFs detector slightly degraded the performance. Li et al. [16] use bottleneck features from a system trained with English and Mandarin, only achieving 1.6% relative improvement compared to a Mandarin baseline system built using conventional acoustic Perceptual Linear Predictive (PLP) features.

The AFs and acoustic features have different numerical ranges, therefore the multi-stream framework can relate better both features by fusing them. The multi-stream framework has been proved to improve the performance of the ASR system. In [17], researchers proposed a multi-stream set up to combine the M-vector features (Sub-band Based Energy Modulation Features) and MFCCs, which improves the ASR performance. In [18], the authors implement a 5 sub-band multi-stream system, with a proposed fusion network in a noise-robust ASR task. Considering the previous works, multi-stream is an effective way to boost ASR systems, especially in challenging tasks (i.e., noisy environment and low-resource ASR).

3. AFs Detectors

3.1. Phonological Attributes

We define the phonological attributes and their corresponding phone set, which are listed in Table 1. For Mandarin, English, German, and French we define the symbol *sil* to represent the silence.

Adapted from previous work [19,20] and IPA [21], we define a universal phonological class definition. Unlike other previous works [22,23], they defined the phonological attributes for only one language (only English or Mandarin). However, our delineation can be shared by multiple languages and can be easily extended to other languages. As shown in Table 1, these attributes of speech can be comprehended by a collection of information from fundamental speech sounds. There are six attributes for each phone: *place*, *manner*, *backness*, *height*, *roundness* and, *voice*. Every phone has a one-hot encoding in each attribute, so after combining all the 6 attributes, there is a 32-dimensional AF vector. Each phone has a unique AF definition. In the Table 1, the *nil* means "not specified". For example,

the phonological class *Place* does not exist in consonants, thus, in consonant phones, this class is defined as *nil*.

Although the phones of these languages do not share the same phone set, we could describe the phones by these attributes. Thus all phones can share phonological knowledge at the phonological level (AFs).

Table 1. Phonological attributes definition.

		Phone			
		Mandarin	**English**	**German**	**French**
Place	alveolar	n l d t	d t s l ng n	b d s z	h l d n z
	bilabial	b p m	b p m	p	b m
	dental	c iy	th dh	t pf ts	N.A
	labiodental	f	f v	f v m	f
	palatal	aa o u j q a oo i iz ei uu g ee x v e vv ii	y	C	j J
	pos-alveolar	zh r sh ix ch	jh z sh ch zh	S Z tS	S Z
	retroflection	er	r er	r	N/A
	velar	h	k g w	x N k g	w N g R
	glottal	N/A	hh	Q	N/A
	nil	all_vowel sil	all_vowel sil	all_vowel sil	all_vowel sil
Manner	approximant	N/A	r y w l	r j I	j w h l
	fricative	f s sh r x h	sh f jh s th z ch zh v dh	S Z	S Z f
	lateral	i	N/A	ts tS pf	N/A
	nasal	m ng n	ng n m	m n	N/A
	stop	t q j b d ch zh c g k p z	d p k t g b hh	p b t d k g Q	g b d p t
	nil	sil all_vowels	sil all_vowels	sil all_vowels	sil all_vowels
Voiced	voiced	oo uu o n ng ei ix a er i vv ee ii iz r m e u iy aa i v	d t g at eh ao uh r z l er uw ow iy ah dh aw aa ey ih m v n w ae jh y s oy ng	all_vowels	2 6 g b @ E J O N θ 9 9~ Z a A~ e d k j m l o u w Λ U~ ε~ y v z
	unvoiced	s ch p zh z x sh b t g q k c h d j f	sh p f k th b ch zh hh	all_constants	S f h p s t
	nil	sil	sil	sil	nil
Height	height	iz vv i iy ix v u uu ii	ih uh uw iy	i i: y y: u: u L l Y U	6 a
	mid	N/A	ey aw	2 2: @	@
	low	aa a	ae aa oy ay ao ow	a~ a~: 6 aU aI	i u y
	mid-height	ee e o o	er eh	e e: o o: o~ o~: E E: 9	O E 9 ε ~ U~
	mid-low	er ei	ah	0	θ Λ o e
	nil	all_constant sil	all_constant sil	all_constant sil	all_constant sil
Round	round	u v uu vv o oo	aw oy ao uh uw iw	y y: u u: Y 2 2: o o o~o~: 9 0 OY aU	2 6 @ θ 9 9~o u U~y
	unround	ix iy a e ee iz i ii aa er ei	ae ih aa ay eh iy er ah	a a: a~a~: 6 aI	E a A~e i ε~o U~
	nil	all_constant sil	all_constant sil	all_constant	all_constant sil
Front	front	i ei v iy vv iz ii	ae ih ey ay ei iy	i i: y y: u u: l Y e	2 6 E θ 9 9~a e i y
	central	a aa er ix	er ah	E OY	@
	back	uu e ee oo o u	aw aa oy ao uh uw ow	u u: U o o: o~ o~: O aU aI	O Λ U~u o A~
	nil	all_constant sil	all_constant sil	all_constant sil	all_constant sil

3.2. Domain-Adversarial Modeling: Integrating Phonological Knowledge

To make the model learn phonological knowledge, we propose a novel representation extraction model, which combines a deep CNN-based model and domain adversarial learning.

Ganin et al. [24] proposed the domain adversarial neural network (DANN). The network learns two classifiers: the main classification task and the domain classifier. The latter determines whether the input sample is from the source or target domain. Both classifiers share the same hidden layers which learn hidden representations for each specific task. The DANN model has a gradient reversal layer (GRL) between the domain classifier and the hidden layers. This layer passes the data during forward propagation, while inverting the sign of the gradient during backward propagation. The network attempts to minimize the task classification error and find a representation that maximizes the error of the domain classifier. The goal of the DANN is to reduce the distribution differences between the source and target domain. With the help of GRL, the model receives the reversed gradient

lamdba. Thus, the network will maximize the error of the domain classifier. Meanwhile, the network attempts to minimize the task classification as usual. By considering these two goals. The model can learn a discriminative representation for the main classification task while making the samples from either domain indistinguishable. As shown in Figure 2, we apply three supervised classification tasks in the DANN model. We include articulatory and phoneme classifiers as main tasks and speaker and language identification as domain classifiers. The objective function of the DANN is defined as follows:

$$L_{DAAE} = \lambda_{phn} L_{phn} + \lambda_i L_{pf-i} - (\lambda_{lid} L_{lid} + \lambda_{spk} L_{spk}). \quad (1)$$

L_{phn}, L_{af},L_{lid} and L_{spk} are the loss functions of the phoneme, articulatory, language and speaker classification, respectively. λ is a trade-off weight parameter to control each loss term (i.e., there is one weight term for each loss). For articulatory classes, there are 6 classes so the 'i' means the i_{th} articulatory class.

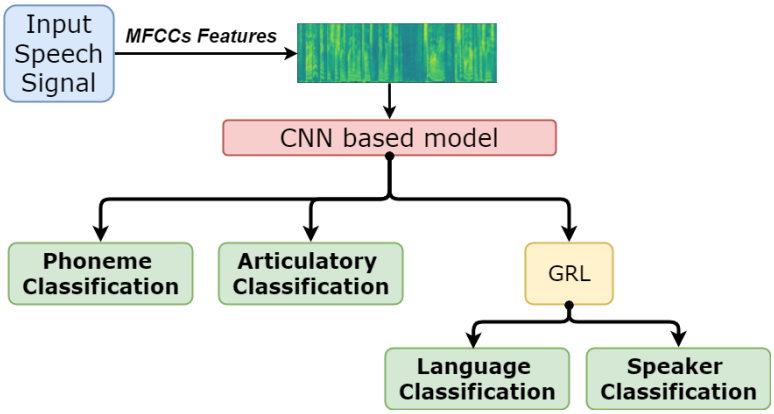

Figure 2. Overview of a DANN based model. **GRL** means *Gradient Reversal Layer* .

The CNN block is a U-Net [25] like CNN structure which is used to obtain features with different time and frequency scales, as depicted in Figure 3.

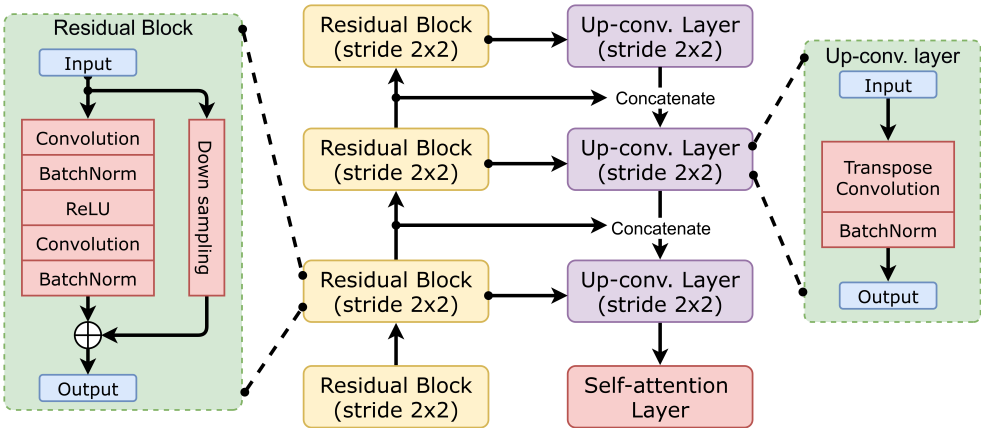

Figure 3. Diagram of the CNN block in DANN based model.

When using the multi-task learning, the performance of the model can be sensitive to the weight between different tasks and finding optimal values can be expensive. To better train the model, we propose to use the adaptative loss function, to automatically tune task-specific weight on the loss functions [26].

$$L_{ada}(\sigma_1, \sigma_2, \sigma_i) = \frac{1}{\sigma_2^2} L_{phn} + \frac{1}{\sigma_2^2} L_{phn-ctc} + \frac{1}{\sigma_i^2} L_{af} + \frac{1}{\sigma_3^2} L_{lid} + \frac{1}{\sigma_4^2} L_{spk} + log\sigma_1\sigma_2\sigma_i\sigma_3\sigma_4 \quad (2)$$

The resulting tuned task-specific equation is presented in Equation (2). In this equation, σ means the coefficient (weight) of different tasks.

4. Multi-Stream ASR Framework

MFCCs and AFs have different numerical ranges, thus simply concatenating them together and training the hybrid system will bring bias towards one feature stream. Therefore, the feature combination would even harm the performance [15]. We overcome this problem by implementing different modes of multi-stream training, where AFs and MFCCs are integrated. For instance, parallel-mode, joint-mode, and MHA-mode:

- **Parallel-mode**: the parallel-mode multi-stream ASR framework is shown in Figure 4. Following the standard Kaldi recipe, The Time Delayed Neutral Network-Hidden Markov Model (TDNN-HMM) is considered the ASR model. First, two TDNN-HMM networks with different features (i.e., MFCCs and AFs) are trained. Then, the bottleneck features from the single TDNN-HMM are taken to concatenate into a later feature vector. This feature vector is used to train the final TDNN-HMM network. The bottleneck features (BNF1 and BNF2) are extracted from the last batch-norm layer following the approach from [27], and the bottleneck dimension is set to a 100 dimension. All the layers in parallel mode are standard TDNN layers.
- **Joint-mode**: the joint-mode approach is described in Figure 5. The joint-mode involves two separate layers for two individual feature streams, and one combination layer to integrate the medium representation of the individual stream. The configuration of the joint-mode is shown in Table 2.
- **MHA-mode**: Multihead attention (MHA) based fusion method is also used in this paper, which is shown in Figure 6. The attention mechanism allows a neural network to capture speech representation from different inputs. The attention score for each $Head_i$ is calculated as:

$$\text{Head}_i = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_i}}\right)V \tag{3}$$

In our experiments, the Q is represented using MFCCs, the K is represented using AFs and the concatenated features are used as V. After fusing those features, a 6-layers TDNN-HMM model is used to train the ASR.

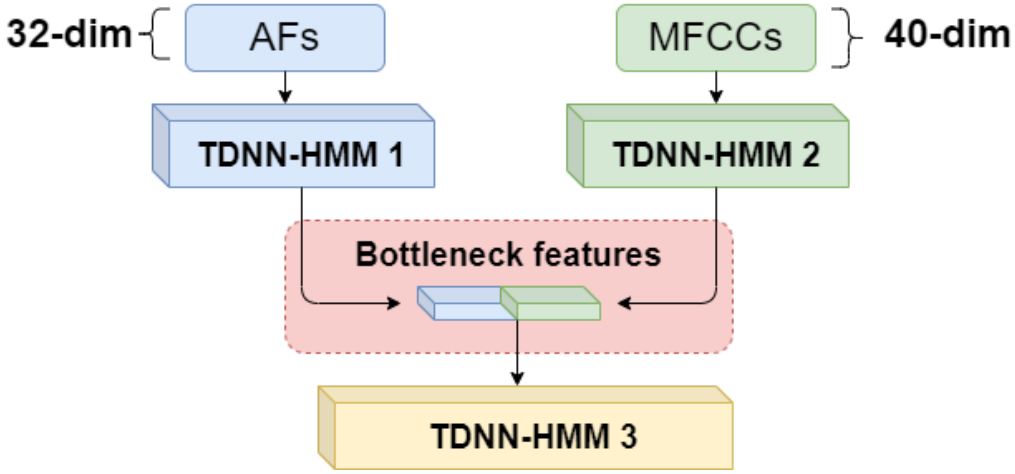

Figure 4. Architecture of the proposed **parallel-mode ASR**. Outputs (bottleneck features) of both single-stream hybrid systems are concatenated. The new concatenated set of features is set as input for the multi-stream hybrid system i.e., TDNN-HMM 3.

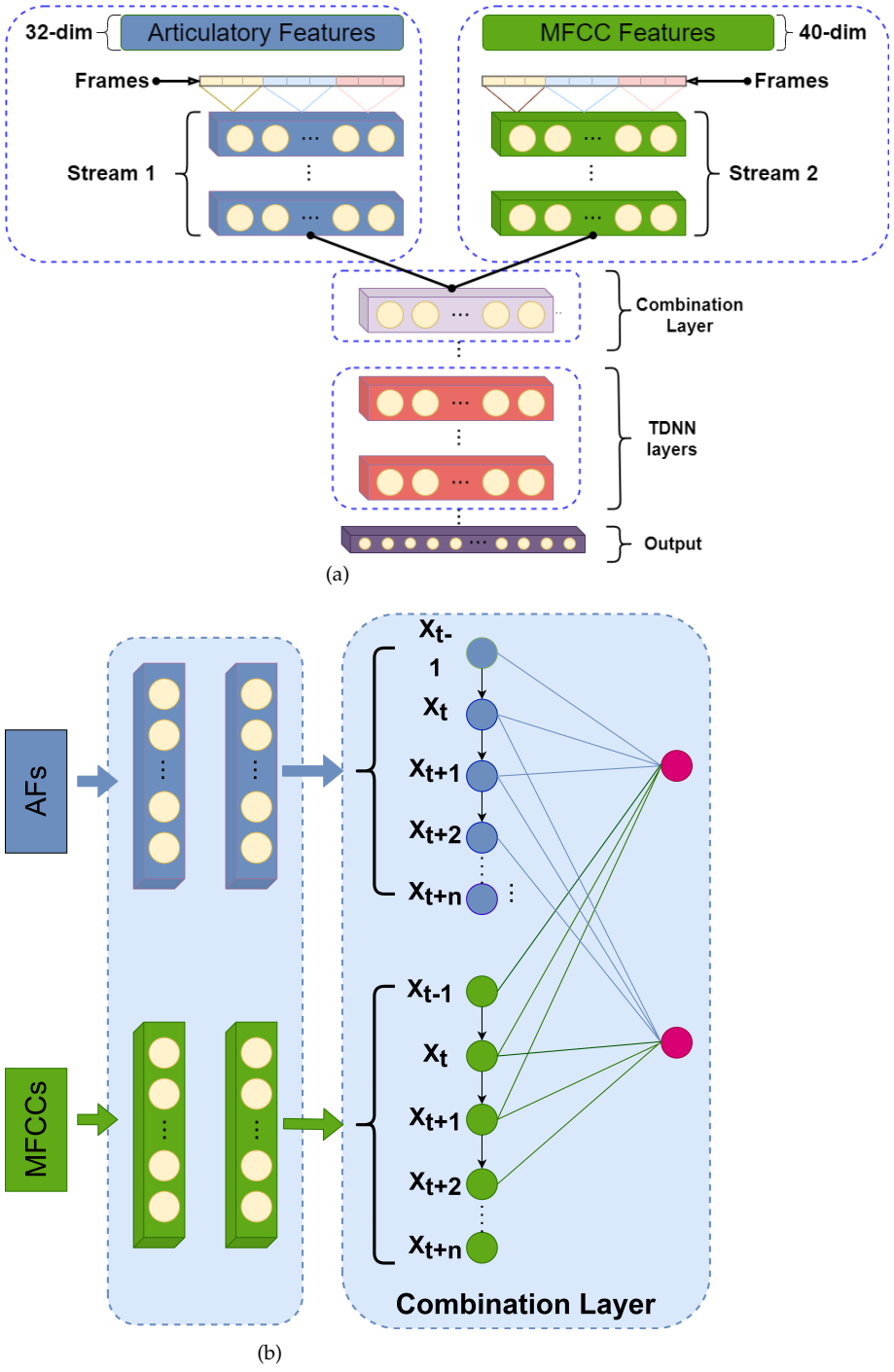

Figure 5. Architecture of the proposed joint-mode multi-stream ASR. Two feature streams are input to the TDNN system and a combination layer is used to combine two feature streams followed by a 3-layers TDNN. (**a**) shows the overview of the joint-mode multi-stream framework while (**b**) is the details of the combination layer.

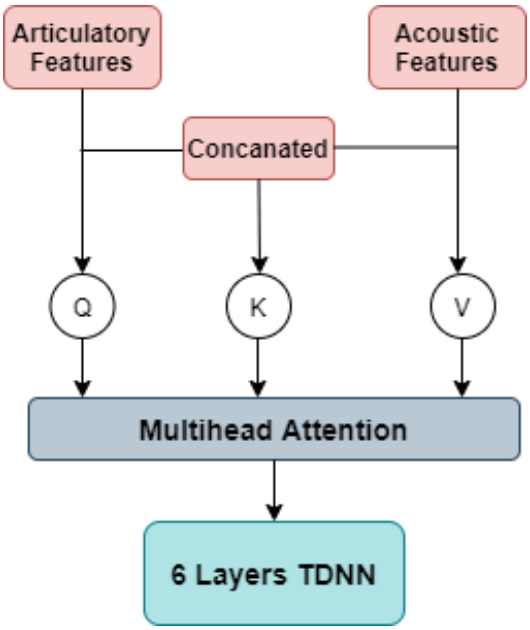

Figure 6. Architecture of the proposed **mha-mode ASR**.

Table 2. Layer-wise context configuration for joint-mode multi-stream framework.

Layer	1	2	Combination Layers	4	5	6
Context	{−1,1}	{−1,1}	{Stream-MFCC(−1,0,1),Stream-AFs(−1,0,1)}	{−3,3}	{−3,3}	{−3,3}

5. Experimental Setup

5.1. Train and Test Data Sets

Because Mandarin, English, German and French are well-researched languages, the experiments are conducted on these languages. For English, we take 100 h subset from Librispeech [28] dataset for training. The French and German dataset is randomly selected from MLS multilingual dataset [29]. For Mandarin, we use THCHS30 [30], which consists of 27 h of data for training and 5.4 h of data for testing. The language model for Mandarin is trained using a text collection that is randomly selected from the Chinese Gigaword corpus (https://catalog.ldc.upenn.edu/LDC2003T09 (accessed on 27 October 2021)). The detailed statistics for these languages are shown in Table 3. The quality of both speech corpora can be considered acoustically similar (i.e., relatively high signal-to-noise ratio, reading speech under similar acoustic conditions, 16 kHz of sampling frequency, etc).

Table 3. Statistics of speech corpora. Target language is Mandarin and English is the source language.

	Utterances		Duration (hours)	
Language	Train	Test	Train	Test
Mandarin	10,893	2496	27.2	6.2
English	28,539	-	100.6	-
French	32,278	-	122.5	-
German	27,148	-	104.6	-

With well-studied linguistic knowledge, English, German, French, and Mandarin are considered in our experiments. The Mandarin here is taken to play the role of low-resourced languages by using limited datasets. By studying our proposed method on those languages using a limited dataset, we can apply this method to other low-resourced languages.

5.2. AFs Detectors

In this paper, the phonological attribute-based features (AFs) extraction model is represented by a DANN model, which is shown in Figure 3. Kernel sizes of all convolutional layers are set to 3 and strides are specified to conserve sequence lengths. A self-attention layer with 8 heads of multi-head attention is stacked on top of the convolutional layers. A learning rate of 0.08 is used and the dropout rate is set to 0.2. All the weights in these models are randomly initialized and are trained using stochastic gradient descent with momentum. The input of the AFs detector is 40-dimensional log mel-filterbank coefficients together with their first and second-order derivatives, derived from 25 ms frames with a 10 ms frame shift.

5.3. Comparison Approaches

To compare the proposed multi-stream systems, we study different approaches. The details of the comparison approaches are described in the following parts. All the TDNN-HMM models have the same architecture (otherwise, it is stated). We use the Lattice-free MMI (LF-MMI) loss function [31] with one-third frame sub-sampling.

- **Baseline**: the TDNN-HMM with LF-MMI loss function using MFCCs is trained as the baseline.
- **TDNN**$_{doublesize}$: The parallel-mode uses the feature stream from two TDNN-HMMs. By training a new TDNN-HMM with double parameters (i.e., double number of parameters in hidden layers and units), we want to verify that the improvement of the parallel-mode is not because of the increasing of parameters. To avoid over-fitting, dropout and L1 regularization are applied.
- **TDNN-adapted**: As shown in Figure 7, a transfer learning-based cross-lingual approach is developed, which is denoted as "*TDNN-adapted*". Firstly, an English TDNN-HMM is trained using the LF-MMI loss function, then the output layer is chopped out and replaced by one corresponding to the Mandarin target units. The whole model is retrained by Mandarin while the transferred layer has a smaller learning rate [32].
- **Feature concatenated approach**: The MFCCs and AFs are concatenated into one feature vector directly. Then the concatenated features are used to train the TDNN-HMM ASR model.
- **Lattice combination approach**: Two word-based lattices from two TDNN-HMM systems (i.e., MFCCs ASR system and AFs ASR system) are combined. Then the combined lattices are used to compute the final results.
- **Bottleneck features**: To better illustrate our AFs, we take the bottleneck features for comparison. First, we train a TDNN-HMM model with LF-MMI loss function using source languages and finally we extracted frame-level embeddings from the well-trained TDNN-HMM model on target language.

5.4. Experiments Configuration

All experiments are conducted on Pkwrap toolkit [33]. The ASR models are trained during 7 epochs. The batch size is 32, whereas the learning rate is reduced gradually from 0.01 to 0.001 (e.g., epoch 1: 0.01, epoch 7: 0.001).

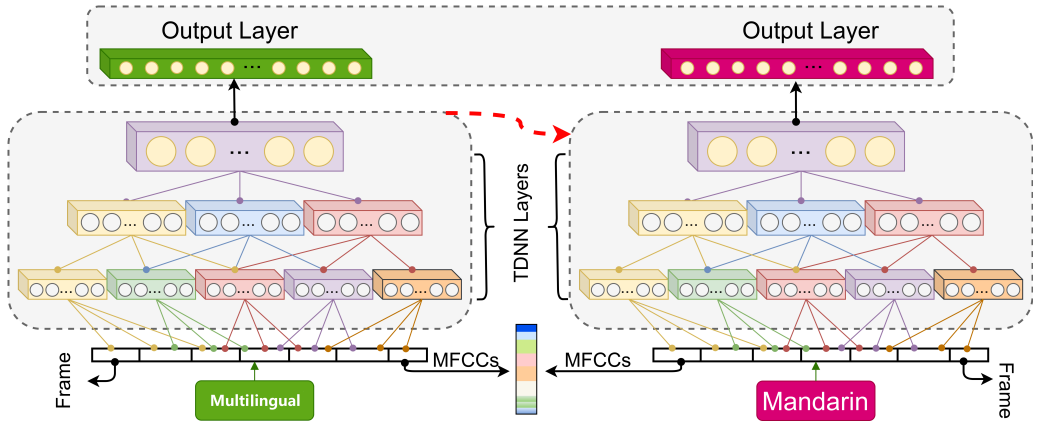

Figure 7. Cross-lingual adaptation from multilingual to Mandarin. The chain TDNN system is trained with concatenated AFs and MFCCs for multilingual speech recognition. The network learns similar acoustic representation (both languages, Mandarin and English) by transferring knowledge on the TDNN layer's weights. The Mandarin system (also chain TDNN) is initialized with the same multilingual system's weights (pointed by the red arrow). The input and output layers are different between both languages.

6. Results and Analysis

6.1. Performance on AFs Detectors

The AF detector is the key part of our framework. The first step is to train a reliable AFs detector. The frame-level performance is listed in Table 4. DANN AFs detector can produce over 82.9% frame-level average accuracy on Mandarin as listed in Table 4.

Table 4. Frame-level accuracy [%] for different AF detectors evaluated on the Mandarin test set for six articulatory classes.

AF Classes	DANN-AF
Place	80.3
Manner	81.4
Voice	87.6
Round	85.5
Height	80.8
Front	82.0
Average accuracy	82.9

6.2. Effectiveness of the Cross-Lingual AFs on ASR

We experimented with different systems to get the best configuration for the joint-mode approach. Table 5 indicates that the configuration that uses $\{-1_{s1}, 0_{s1}, 1_{s1}, -1_{s2}, 0_{s2}, 1_{s2}\}$ (see Table 5) has the best performance.

Table 5. Joint-mode multi-stream framework using different configuration of combination layer evaluated on the Mandarin test set. The subscript **s1** means this frame is from stream-MFCCs and the **s2** means this frame is from stream AFs corresponding to Figure 5 (i.e., -1_{s1} means the $t-1$ frame is from stream MFCCs, 0_{s1} means the current t frame is from stream MFCCs).

Combination Layer Configuration	WER[%]
-1_{s1},0_{s1},1_{s1},0_{s2},-1_{s2},1_{s2}	**24.0**
-1_{s1},-1_{s2},0_{s1},0_{s2},1_{s1},1_{s2}	24.3
Fully connected layer	24.7

Then different results on ASR are listed in Table 6. When comparing the baseline (24.8% WER) with the parallel-mode (23.4% WER), joint-mode (24.0% WER) and MHA-mode (22.5% WER) our model achieves 5.6%, 3.2% and 9.2% relative improvement in WERs, respectively. Meanwhile, the results of $\text{TDNN}_{doublesize}$ (24.6% WER) show that the improvement of parallel-mode is not because of the increase in the number of parameters. Similarly, the parallel-mode (23.4% WER), joint-mode (24.0% WER) and MHA-mode (22.5% WER) still outperform the TDNN-adapted approach (24.2% WER). This indicates that the improvements here are not only due to the source language, but because our proposed approaches have a better ability to integrate cross-lingual AFs. However, the feature concatenated approach (24.7% WER) and lattice combination approach (24.6% WER) have slight improvement compared to the baseline (24.8% WER). Finally, it also proves that simply combining the AFs and MFCCs cannot improve the performance considerably [15].

Table 6. Comparison of word error rates % (WER) for different approaches and different multi-stream approaches evaluated on target-language (Mandarin) test set. The lattice-combination approach is described in [12], Feature-combination approach is used in [14,34].

System	Features	Source Languages	WER[%]
Baseline	MFCCs	-	24.8
$\text{TDNN}_{doublesize}$	MFCCs	-	24.6
TDNN-adapted	MFCCs	-	24.3
Parallel-mode	MFCCs+multi-lingual AFs	Multilingual	23.4
Joint-mode	MFCCs+multi-lingual AFs	Multilingual	24.0
mha-mode	MFCCs+multi-lingual AFs	Multilingual	**22.5**
mha-mode	MFCCs + Bottleneck-features	Multilingual	23.3
Feature concatenated	MFCCs+cross-lingual AFs	Multilingual	24.7
Lattice combination	MFCCs+cross-lingual AFs	Multilingual	24.6

Considering all the results above, all the three feature fusion approaches perform better than the baseline (see Table 6). The MHA-based approach is more generalized when exploiting cross-lingual AFs and gets the best performance (22.5% WER)

We also compare the results using AFs and bottleneck features (BNF) on mha-mode, the results indicate that the AFs still outperform BNF. It is reasonable because from Figure 3 we can see the DANN already contains phoneme information.

6.3. Performance on Extremely Low-Resource Training Data

To further study the performance of MHA, we also consider the approach by training the AFs detectors with Mandarin data directly (so-called *mono-lingual AFs*). The sizes of training data varied from 1 h to 27 h, which are randomly selected from the *THCHS30* dataset. The system trained with the 1-hour data set means the extremely low-resourced case.

The results are presented in Table 7 and Figure 8. In Table 7, the baseline is the same as that in Table 6. From the Figure 8, it can be found that both monolingual AFs and cross-lingual AFs can improve the performance in all cases. Again, it indicates that the AFs from the DANN-AFs detector can boost the ASR performance.

As expected, the more training data employed during training, the better performance the system yielded. In the case of low-resources (i.e., very limited data, less than 5 h) the cross-lingual AFs outperform the system trained with monolingual AFs. More specifically, in the condition of extremely low-resourced training data (i.e., 1 h train subset), the cross-lingual AFs (45.0% WER) have a significant improvement compared to MFCCs baseline (55.4% WER) and monolingual AFs (47.8% WER). Therefore, the less training data, the better improvement can be reached using cross-lingual AFs.

Table 7. WERs(%) on different sizes of Mandarin training data using different versions of AFs, which are evaluated on THCHS30 test set. All the results are based on MHA-mode framework. The Baseline system is trained with Mandarin MFCCs.

	Train Set Size				
System	**1 h**	**3 h**	**5 h**	**10 h**	**27 h**
Baseline	55.4	37.6	32.3	26.3	**24.8**
Monolingual AFs	47.8	35.1	30.3	23.5	**22.2**
Cross-lingual AFs	45.0	34.5	30.2	24.4	**22.5**

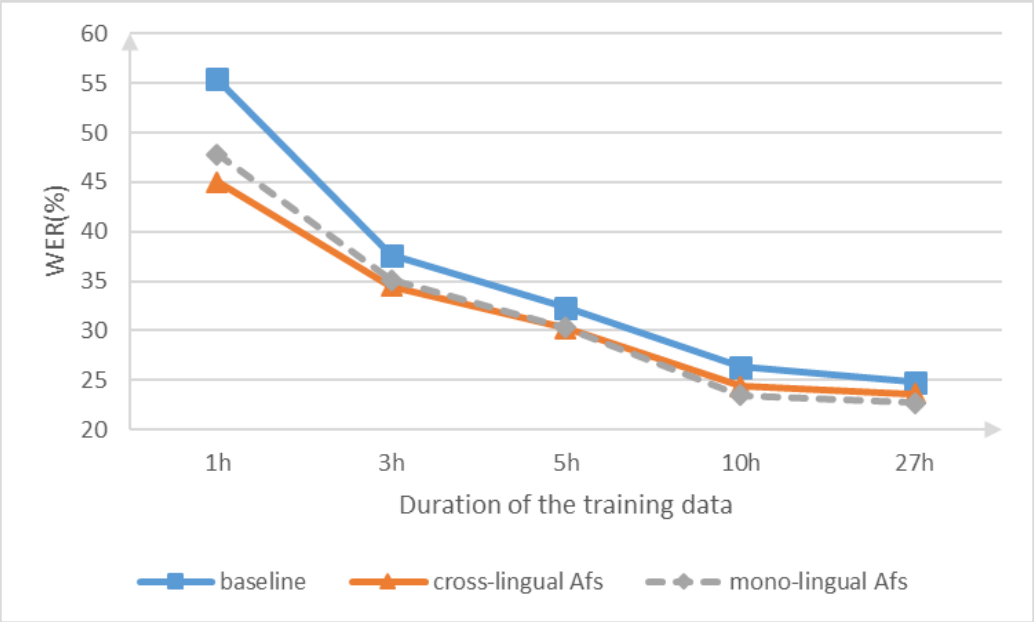

Figure 8. WER(%) on parallel-mode approach for different size of training Mandarin data. Cross-lingual AFs means this parallel-mode is trained with cross-lingual AFs and the mono-lingual AFs means the parallel-mode is trained with mono-lingual AFs.

To better illustrate our proposed method, the Cantonese is also introduced because of the language similarity with Mandarin. The Cantonese dataset is taken from OLR 2021 challenge [35], which only has 13 h for training and 0.4 h for testing. Thus, Cantonese is performed as a low-resourced language as well. We conduct the experiments using mha-mode and the same DANN AFs model is used. The results are shown in Table 8. The same conclusion can be found that the cross-lingual AFs from DANN can boost the Cantonese ASR through MHA-mode multi-stream framework.

Table 8. WERs(%) on Cantonese test set. All the results are based on MHA-mode framework. The Baseline system is trained with Cantonese MFCCs.

System	**WER(%)**
Baseline	32.4
Cross-lingual AFs	29.5

7. Conclusions

This research demonstrates that the AFs detectors can perform phone decomposer tasks, which inject phones into AFs space. However, different languages do not have the same phone set, still, they share the phonological knowledge at AFs level by using AFs detectors.

The languages chosen in our experiments are English, German, French, and Mandarin where Mandarin is presented as a low-resourced language by using limited training data. Because those languages come from different language families, the knowledge transferring method is more challenging; thus this approach can be easily extended to other language pairs. We first propose a universal phone-to-articulatory mapping, where the different language phones can share the same articulatory space. Meanwhile, this mapping also can be extended to other languages easily. Experimental results indicate that the DANN-based AFs can improve the ASR results by using different feature fusion approaches and the multi-head attention method can reach the best performance. On extremely low-resourced conditions, our proposed approach has significant improvements when compared with standard ASR systems. Our future work will extend our proposed approach to some other languages and different downstream tasks.

Author Contributions: Conceptualization, Q.Z., X.X., C.H. and J.W.; methodology, Q.Z.; implementation, Q.Z. and C.H.; validation, Q.Z., X.X., C.H. and H.C.; writing—original draft preparation, Q.Z. and J.Z.-G.; supervision, X.X. and J.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been supported by National Nature Science Foundation of China (No.11590772, No. 62071039), Science and Technology Innovation Foundation of Shenzhen (JCYJ20180504165826861).

Data Availability Statement: MLS multilingual speech dataset: Pratap, V.; Xu, Q.; Sriram, A. ; Synnaeve, G.; Collobert, R. (2020). Mls: a large-scale multilingual dataset for speech research. THCHS30 dataset: Wang, D.; Zhang, X. 2015, THCHS-30: A Free Chinese Speech Corpus. arXiv:cs.CL/1512.01882.

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

Abbreviations

The following abbreviations are used in this manuscript:

ASR	Automatic Speech Recognition
AFs	Articulatory Features
CNN	Convolutional Neural Networks
DNN	Deep Neural Networks
DANN	Domain adversarial Neural Networks
MHA	Multi head Attention
WER	Word Error Rate
Fbank	Filterbank

References

1. Baevski, A.; Hsu, W.N.; Conneau, A.; Auli, M. Unsupervised Speech Recognition. *arXiv* **2021**, arXiv:2105.11084.
2. Mitra, V.; Sivaraman, G.; Nam, H.; Espy-Wilson, C.; Saltzman, E. Articulatory features from deep neural networks and their role in speech recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014.
3. Wang, L.; Chen, H.; Li, S.; Meng, H.M. Phoneme-level articulatory animation in pronunciation training. *Speech Commun.* **2012**, *54*, 845–856. [CrossRef]
4. Lin, Y.; Wang, L.; Dang, J.; Li, S.; Ding, C. End-to-End Articulatory Modeling for Dysarthric Articulatory Attribute Detection. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 7349–7353.
5. Papcun, G.; Hochberg, J.; Thomas, T.R.; Laroche, F.; Zacks, J.; Levy, S. Inferring articulation and recognizing gestures from acoustics with a neural network trained on X-ray microbeam data. *J. Acoust. Soc. Am.* **1992**, *92*, 688–700. [CrossRef] [PubMed]
6. Schroeter, J.; Sondhi, M.M. Techniques for estimating vocal-tract shapes from the speech signal. *IEEE Trans. Speech Audio Process.* **1994**, *2*, 133–150. [CrossRef]
7. Merkx, D.; Scharenborg, O. Articulatory Feature Classification Using Convolutional Neural Networks. In Proceedings of the Eighth Annual Conference of the International Speech Communication Association (Interspeech), Hyderabad, India, 2–6 September 2018; pp. 2142–2146. [CrossRef]
8. Manjunath, K.; Rao, K.S. Improvement of phone recognition accuracy using articulatory features. *Circuits Syst. Signal Process.* **2018**, *37*, 704–728. [CrossRef]

9. Wikipedia Contributors. International Phonetic Alphabet—Wikipedia, The Free Encyclopedia. 2020. Available online: https://en.wikipedia.org/w/index.php?title=International_Phonetic_Alphabet&oldid=1060663021(accessed on 28 October 2020).
10. Mitra, V.; Wang, W.; Bartels, C.; Franco, H.; Vergyri, D. Articulatory information and multiview features for large vocabulary continuous speech recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5634–5638.
11. Siniscalchi, S.M.; Reed, J.; Svendsen, T.; Lee, C.H. Universal attribute characterization of spoken languages for automatic spoken language recognition. *Comput. Speech Lang.* **2013**, *27*, 209–227. [CrossRef]
12. Duan, R.; Kawahara, T.; Dantsuji, M.; Zhang, J. Articulatory modeling for pronunciation error detection without non-native training data based on DNN transfer learning. *IEICE Trans. Inf. Syst.* **2017**, *100*, 2174–2182. [CrossRef]
13. Lee, C.H.; Clements, M.A.; Dusan, S.; Fosler-Lussier, E.; Johnson, K.; Juang, B.H.; Rabiner, L.R. An overview on automatic speech attribute transcription (ASAT). In Proceedings of the Eighth Annual Conference of the International Speech Communication Association, Antwerp, Belgium, 27–31 August 2007.
14. Krishna, H.; Gurugubelli, K.; V, V.V.R.; Vuppala, A.K. An Exploration towards Joint Acoustic Modeling for Indian Languages: IIIT-H Submission for Low Resource Speech Recognition Challenge for Indian Languages. In Proceedings of the Interspeech 2018, Hyderabad, India, 2–6 September 2018; pp. 3192–3196. [CrossRef]
15. Cetin, O.; Kantor, A.; King, S.; Bartels, C.; Magimai-Doss, M.; Frankel, J.; Livescu, K. An articulatory feature-based tandem approach and factored observation modeling. In Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing, Honolulu, HI, USA, 15–20 April 2007; Volume 4, pp. 4–645.
16. Li, J.; Zheng, R.; Xu, B. Investigation of cross-lingual bottleneck features in hybrid ASR systems. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Singapore, 14–18 September 2014; pp. 1395–1399.
17. Sadhu, S.; Li, R.; Hermansky, H. M-vectors: Sub-band Based Energy Modulation Features for Multi-stream Automatic Speech Recognition. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6545–6549. [CrossRef]
18. Mallidi, S.H.; Hermansky, H. Novel neural network based fusion for multistream ASR. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 5680–5684.
19. Yuen, I.; Davis, M.H.; Brysbaert, M.; Rastle, K. Activation of articulatory information in speech perception. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 592–597. [CrossRef]
20. Browman, C.P.; Goldstein, L. Articulatory phonology: An overview. *Phonetica* **1992**, *49*, 155–180. [CrossRef] [PubMed]
21. Association, I.P.; Staff, I.P.A.; Press, C. *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*; Cambridge University Press: Cambridge, UK, 1999.
22. Qu, L.; Weber, C.; Lakomkin, E.; Twiefel, J.; Wermter, S. Combining Articulatory Features with End-to-End Learning in Speech Recognition. In Proceedings of the International Conference on Artificial Neural Networks, Rhodes, Greece, 4–7 October 2018; pp. 500–510.
23. Lin, J.; Xie, Y.; Gao, Y.; Zhang, J. Improving Mandarin tone recognition based on DNN by combining acoustic and articulatory features. In Proceedings of the 2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP), Tianjin, China, 17–20 October 2016; pp. 1–5. [CrossRef]
24. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; Lempitsky, V. Domain-Adversarial Training of Neural Networks. *J. Mach. Learn. Res.* **2016**, *17*, 1–35.
25. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015.
26. Kendall, A.; Gal, Y.; Cipolla, R. Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2018, Salt Lake City, UT, USA, 18–23 June 2018.
27. Veselỳ, K.; Karafiát, M.; Grézl, F.; Janda, M.; Egorova, E. The language-independent bottleneck features. In Proceedings of the 2012 IEEE Spoken Language Technology Workshop (SLT), Miami, FL, USA, 2–5 December 2012; pp. 336–341.
28. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: An ASR corpus based on public domain audio books. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 19–24 April 2015; pp. 5206–5210.
29. Pratap, V.; Xu, Q.; Sriram, A.; Synnaeve, G.; Collobert, R. MLS: A Large-Scale Multilingual Dataset for Speech Research. In Proceedings of the Interspeech 2020, Shanghai, China, 25–29 October 2020. [CrossRef]
30. Wang, D.; Zhang, X. Thchs-30: A free chinese speech corpus. *arXiv* **2015**, arXiv:1512.01882.
31. Povey, D.; Peddinti, V.; Galvez, D.; Ghahremani, P.; Manohar, V.; Na, X.; Wang, Y.; Khudanpur, S. Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI. In Proceedings of the Interspeech 2016, San Francisco, CA, USA, 8–12 September 2016; pp. 2751–2755. [CrossRef]
32. Ghahremani, P.; Manohar, V.; Hadian, H.; Povey, D.; Khudanpur, S. Investigation of transfer learning for ASR using LF-MMI trained neural networks. In Proceedings of the 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Okinawa, Japan, 16–20 December 2017; pp. 279–286.

33. Madikeri, S.; Tong, S.; Zuluaga-Gomez, J.; Vyas, A.; Motlicek, P.; Bourlard, H. Pkwrap: A PyTorch Package for LF-MMI Training of Acoustic Models. *arXiv* **2020**, arXiv:2010.03466.
34. Kirchhoff, K.; Fink, G.A.; Sagerer, G. Combining acoustic and articulatory feature information for robust speech recognition. *Speech Commun.* **2002**, *37*, 303–319. [CrossRef]
35. Wang, B.; Hu, W.; Li, J.; Zhi, Y.; Li, Z.; Hong, Q.; Li, L.; Wang, D.; Song, L.; Yang, C. OLR 2021 Challenge: Datasets, Rules and Baselines. *arXiv* **2021**, arXiv:2107.11113