INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Learning Cross-lingual Mappings for Data Augmentation to Improve Low-Resource Speech Recognition

Muhammad Umar Farooq, Thomas Hain

Speech and Hearing Research Group, University of Sheffield, UK.
{mufarooq1, t.hain}@sheffield.ac.uk

Abstract

Exploiting cross-lingual resources is an effective way to compensate for data scarcity of low resource languages. Recently, a novel multilingual model fusion technique has been proposed where a model is trained to learn cross-lingual acoustic-phonetic similarities as a mapping function. However, handcrafted lexicons have been used to train hybrid DNN-HMM ASR systems. To remove this dependency, we extend the concept of learnable cross-lingual mappings for end-to-end speech recognition. Furthermore, mapping models are employed to transliterate the source languages to the target language without using parallel data. Finally, the source audio and its transliteration is used for data augmentation to retrain the target language ASR. The results show that any source language ASR model can be used for a low-resource target language recognition followed by proposed mapping model. Furthermore, data augmentation results in a relative gain up to 5% over baseline monolingual model.

Index Terms: automatic speech recognition, low-resource, cross-lingual, multilingual, data augmentation

1. Introduction

End-to-end (e2e) acoustic modelling techniques require a lot of training data for reliable parameters estimation. However, more than half of the world's population speak only 23 languages out of more than 7000 languages being spoken across the globe [1]. Thus only a few languages have sufficient data resources, and a lot of languages are still under resourced to build an ASR system. For such languages, multilingual speech recognition systems have stolen the lime light over the past decade [2, 3, 4, 5, 6, 7] which have been used for feature extraction [8, 9, 10] or directly for transfer learning [11, 12].

Data augmentation is another approach to increase the training data of a low-resource language. Commonly used data augmentation technique includes extending training data by making perturbed copies either by adding noise [13, 14], varying speed and tempo of original speech [15], vocal tract length perturbation (VTLP) [16, 17], SpecAugment [18] and combinations of these methods [19]. All these techniques are based on audio data augmentation.

In the recent past, a few studies have been done to augment data by processing text rather than speech [20, 21, 22]. Transcripts from different languages have been transliterated to Latin script to train a multilingual system [21]. However, it requires paired data (a word in original script and its transliteration in Latin) for each language. Thomas *et al.* [22] have proposed to transliterate a source language data to the target language without using parallel data. Source language audio data has been decoded using the target language ASR to transliterate them into target language which is then used as augmented data to retrain target language ASR. Though this is a novel idea, an out of domain ASR has no knowledge of input language and thus is not expected to generate a good transliteration.

Recently, we have proposed a technique to learn cross-lingual acoustic-phonetic similarities on phoneme level [23] which has been used for multilingual and cross-lingual acoustic model fusion [24]. A model is trained to learn mappings from a source language ASR output posterior distributions to that of the target language ASR. The study has been based on an underlying assumption that these mapping models can learn some language-related relations between phonemic posterior distributions. Though the study proves the concept, the work has been done on phoneme level using DNN-HMM hybrid systems and handcrafted lexicons for each language. In this work, we extend the previous work for cross-lingual e2e speech recognition systems. Then the ASR systems of source languages followed by a source-target mapping model for each source-target pair is used to transliterate source data into the target language script. Though both the components are trained on task specific data and are expected to generate better output labels, transliteration of a source language audio data into the target language is still unintelligible especially for unrelated languages and thus called *ciphered* data. So, the key contribution of this work is two-fold;

- it extends the concept of learning cross-lingual mappings for e2e speech recognition systems and
- generates ciphered text for a target language data augmentation using source languages ASR and <*source-target*> mapping models.

Exploiting mapping models for cross-lingual speech recognition shows that using a source language ASR for a target language gives comparable results. These mapping models are trained on limited data, and using a source language ASR followed by a mapping model enables us to exploit cross-lingual ASR to recognise the target language speech data. Furthermore, the proposed data transliteration and augmentation techniques yield up to 5% and 28.5% relative improvement in character error rate (CER) when compared with monolingual and multilingual ASR systems respectively.

2. Mapping models

Let M_A and M_{S_i} be the monolingual acoustic models of the target and i^{th} source language respectively, a mapping model N_{S_iA} is trained to translate posteriors P_{S_i} of dimension d_{S_i} from M_{S_i} to the posteriors P_{S_iA} of dimension d_A where d_A is the dimension of posteriors from M_A. Given a set of observa-

This work was partly supported by LivePerson Inc. at the Liveperson Research Centre.

10.21437/Interspeech.2023-1613

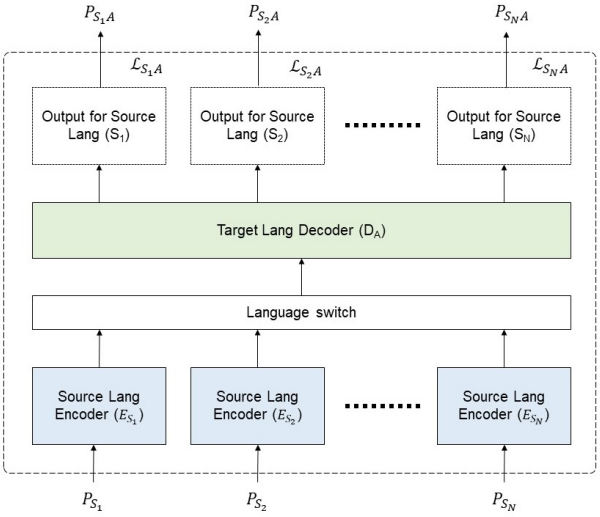

Figure 1: *Architecture of the MESD mapping model*

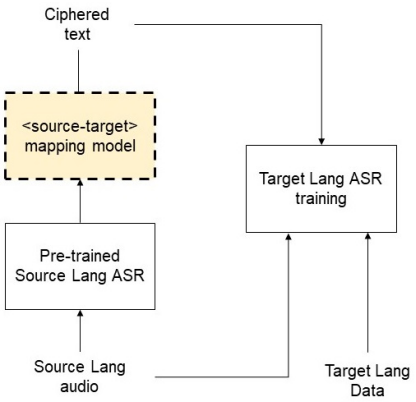

Figure 2: *Flow of generating data for augmentation and retraining of target language ASR*

tions $X = \{x_1, x_2, \ldots, x_T\}$ of the target language, posterior distributions ($P^Z = \{p_1, p_2, \ldots, p_T\}$ where $Z \in \{A, S_i\}$) are attained from the target and the i^{th} source acoustic models. A mapping model is trained using KL divergence loss to map posteriors from i^{th} source acoustic model (P^{S_i}) to the target language posteriors (P^{S_iA}). The loss function is given as;

$$\mathcal{L}_{S_iA}(\theta) = \sum_{n=1}^{B} p_n^A \cdot (\log p_n^A - \log p_n^{S_iA}) \tag{1}$$

where B is the number of frames in one batch for training a mapping model N_{S_iA} to map posteriors from i^{th} source language to the target language.

Mapping models in the previous work [24] have been trained on frame level without considering the contextual information but connected speech is a continuous signal which poses co-articulation and temporal smearing. Furthermore, a separate model has been trained for each source-target language pair rising a requirement of $N(N-1)$ mapping models. So, the architecture of mapping model is modified in this work to a sequence-to-sequence model with Multi Encoder Single Decoder (MESD) architecture. Thus, it incorporates contextual information and reduces the required number of mapping models to just N. The architecture of MESD is shown in the Figure 1.

During the training of MESD model, outputs from all the source acoustic models for a given utterance u are fed to source-language dependent encoders successively. Embeddings from the final layer of the encoders are then passed to a single target-language dependent decoder. Loss is calculated as mean of the losses for all encoder-decoder pairs.

$$\mathcal{L}_A(\theta) = \sum_{K} w_k \cdot \mathcal{L}_{S_kA} \tag{2}$$

where K is the number of total source languages ($N-1$), $w_k = \frac{1}{K}$ in the case of mean average and $\mathcal{L}_{S_kA}$ is given in Equation 1 which is still frame based. It allows mapping models training to converge in low-resource setting as a small amount of data provides millions of examples. However, this causes unbalanced training across languages as mean value can be continuously decreasing when loss for one of the languages is decreasing monotonically but increasing in same fashion for the other one. This can cause model to learn mappings for one language way better than the other. To cope with this issue, a dynamic weighting scheme is applied to weight the losses for each encoder-decoder loss. For the experimentation here, rank sum weighting [25] is used to assign the weights. In this scheme, weights are assigned based on their normalised ranks. So, w in Equation 2 now becomes

$$w_r = \frac{2(K+1-r)}{K(K+1)} \tag{3}$$

where r is rank of the language when the languages are sorted on decreasing values of their losses. It restricts model from biasing towards a specific language or a group of languages.

Though a mapping model contains multiple encoders, any encoder can be used with decoder during decoding and MESD does not require data stream from all the encoders for a given utterance. It implies that mappings can be obtained having input even from only one source language at a time. Training of these mapping models allows to use any source language ASR for decoding the data of a target language followed by the source-target mapping model.

3. Ciphering text

In the previous work [22], target language ASR has been used for transliteration of source language audio data for data augmentation and retraining of target language ASR. However, an ASR does not have any source language information and is not expected to generate a rationale transliterated transcriptions.

In this work on the contrary, source language audio data is decoded using in-domain ASR (M_{S_i}) and then the output posterior distributions (P^{S_i}) are transformed to the target language posterior distributions (P^{S_iA}) using the source-target mapping model (N_{S_iA}). Mapped posteriors from the mapping models are then used to generate transliterated transcriptions (alternatively referred as *ciphered* text or transcriptions) using greedy decoding. Though the transliterations still might not be exact transliterations (thus called *ciphered* text), both the components involved in the process are trained using the task-specific data and are expected to perform better.

Source language audios and their ciphered transcriptions are then used as augmented data for retraining of the target language ASR. The flow is shown in the Figure 2.

4. Experimental setup

4.1. Data set

As this work extends the previous work, experiments here are done on same data set as used in [24]. Full Language Packs (FLP) of four low-resource languages from IARPA BABEL speech corpus [26] (Tamil (*tam*), Telugu (*tel*), Cebuano (*ceb*) and Javanese (*jav*)) are used for baseline ASR training and evaluation. BABEL data set mostly consist of conversational telephone speech with real-time background noises and is quite challenging because of conversation styles, limited bandwidth, environment conditions and channel. All the utterances without any speech are discarded. The details of the data sets are tabulated in Table 1.

For training of the mapping models, a subset of 30 hours is randomly selected from each language pack. This data is further split into 29 hours of train set and 1 hour of dev set.

4.2. Speech recognition systems

Hybrid CTC/attention architecture [27] is used to train all speech recognition models which consists of three modules that are; a shared encoder, an attention decoder and a CTC module. The training process jointly optimises the weighted sum of CTC and attention model.

$$\mathcal{L}_{ASR}(\theta) = \alpha\mathcal{L}_{CTC} + (1-\alpha)\mathcal{L}_{att} \tag{4}$$

The input to the model is 40 filterbanks and the output of the model is the byte-pair encoded (BPE) tokens. Monolingual ASRs are trained for 100 BPE tokens for each language while the output of multilingual ASR is 400 tokens. SentencePiece library [28] is used for tokenisation. During decoding, the final prediction is made based on a weighted sum of log probabilities from both the CTC and attention components. Given a speech input X, the final prediction $\hat{Y}$ is given by;

$$\hat{Y} = \arg\max_{Y\in\mathcal{Y}}\{\lambda \log P_{CTC}(Y|X) + (1-\lambda)\log P_{att}(Y|X)\} \tag{5}$$

where λ is a hyper-parameter. The values of α and λ are kept same for all ASR systems. SpeechBrain toolkit [29] is used for training of all ASR systems.

4.3. Mapping models

A multi encoder single decoder model is trained for each target language. In an MESD model, there are three encoders and only one attention decoder. Each encoder and single decoder consists of one bidirectional RNN layer. For each target language, mapping model size is only 2.59 million parameters.

Table 1: *Details of BABEL data sets used for the experimentation*

Lang	Train		Eval	
	# hours	# spks	# hours	# spks
Tamil *(tam)*	59.11	372	7.8	61
Telugu *(tel)*	32.94	243	4.97	60
Cebuano *(ceb)*	37.44	239	6.59	60
Javanese *(jav)*	41.15	242	7.96	60

4.4. Performance metric

Accuracy of a mapping model is measured as the ratio of number of correctly mapped frames to the total number of frames as given in Equation 6. Correctly mapped frames are defined as the frames where the values of $\arg\max(mapped\_posteriors)$ and $\arg\max(targetAM\_posteriors)$ are the same.

$$\arg\max_k(p^A_{t,k}) == \arg\max_k(p^{S_iA}_{t,k}))) \Rightarrow CMF++$$

$$Accuracy = \frac{CMF}{T} \tag{6}$$

where k is the index of classes in the output vector p_t, CMF is the number of correctly mapped frames and T is the total number of frames.

For downstream speech recognition task, results are reported in terms of percent character error rate.

5. Results and Discussion

5.1. Mapping models

Accuracies of mapping models, trained to map posterior distribution from a source language ASR to the target language ASR, are tabulated in Table 2. Analysis shows that correct target class is still among top n mapped classes if not the most probable one. So, the mapping models accuracy is calculated for different values of n where n represents the number of most probable classes. Though the accuracy increases with increasing value of n, rate of change is not as much as observed in case of phonemes by [24] which implies that the performance of mapping model in case of phoneme based hybrid DNN-HMM systems has been better than that for e2e systems. Since the mapping models are trained using posterior distributions of ASR outputs, one potential reason could be the detrimental affect of speech recognition systems on the training of mapping models. However, the joint analysis of amount of training data (Table 1), performance of monolingual speech recognition systems (Table 3) and performance of mapping models (Table 2) rules out this reason.

Amount of mapping model training data is same for all the languages but the mappings for *ceb* and *jav* target language is better than *tam* and *tel*. Even for *ceb* and *jav* target languages, accuracy of mappings from *tel* source language is very low in comparison to other source languages. The investigation reveals that as the number of BPE tokens are restricted to 100 for all the languages, *ceb* and *jav* having only 19 and 26 characters respectively have good context coverage in 100 BPE tokens. But the BPE tokens extracted for *tel*, which have more than 52 characters, do not cover context very well. Furthermore, both *ceb* and *jav* are written in Latin script and thus have a full overlap of characters and are even acoustically close. While on the other hand though both *tam* and *tel* belong to same Dravidian family, their writing scripts are different which makes it difficult for model to learn mappings with limited number of BPE tokens.

5.2. Ciphering text

For a given target language, audio data of all the source languages is decoded using language dependent ASR systems and the output posterior distributions are then mapped to target language distributions using the mapping models N_{S_iA}. Greedy decoding is carried out on these output posterior distributions to

Actual *(ceb)*	hello
Ciphered *(tam)*	இஹலோ
Ciphered *(tel)*	ఎలో న్ ్ో
Ciphered *(jav)*	hlo
Actual *(ceb)*	singkuwinta
Ciphered *(tam)*	சுயலாா
Ciphered *(tel)*	ర సిం ్ాం
Ciphered *(jav)*	sing pull ka

Figure 3: *Examples of ciphered transcriptions*

generate ciphered transcriptions for the target language. Language model (LM) is not integrated at this stage to avoid LM affect on transliterations. As this stage solely depends on mapping models, the quality of ciphered text depends on mapping models accuracy for $n = 1$. The analysis of ciphered transcriptions shows that the transliteration is fairly good for shorter utterances but gets worse for longer utterances. A few examples of ciphered transcriptions are shown in Figure 3.

5.3. ASR

Monolingual systems (*mono*) are the language dependent acoustic and language models which are trained on target language specific data sets. The train sets of all the languages are then mixed to train a multilingual system (*multi*). Language model for a multilingual system is also trained using mix corpora of individual languages. The results of speech recognition systems are shown in Table 3. The first row contains the monolingual ASR result without using LM for a later comparison while rest of the results are ASR decoding with LM.

For a given target language test set, speech recognition results are also computed on top of mapping models after decoding target language data using source language acoustic models. Greedy decoding is applied on mapped posteriors and the results are shown in Table 4. CER on diagonal is the same as the first row of Table 3. Though these results are from source language ASR followed by a source-target mapping model and does not use language dependent ASR, it preforms better than monolingual ASR in case of *jav*. Results are comparable for other languages but fairly depend on mapping models performance. It is evident from these results that a source language acoustic model can be used for decoding of a target language followed by a mapping model trained on limited amount of data.

Table 2: *Accuracy of the mapping models considering top n mapped classes*

Target Lang	Source Lang	*Mapping model* accuracy			
		n=1	*n=2*	*n=5*	*n=10*
tam	tel	47.46	54.58	66.31	77.06
	ceb	45.98	52.88	64.25	74.65
	jav	46.97	54.02	65.63	76.26
tel	tam	48.88	56.20	67.80	78.28
	ceb	46.22	53.27	64.97	75.96
	jav	47.40	54.78	66.76	77.54
ceb	tam	60.53	66.32	74.79	82.31
	tel	48.32	51.43	56.49	62.53
	jav	65.04	71.39	80.06	86.58
jav	tam	62.24	68.40	77.00	83.76
	tel	54.64	57.92	62.29	67.69
	ceb	65.51	71.85	80.30	86.65

Table 3: *ASR performance in terms of %CER*

Lang	tam	tel	ceb	jav
mono	44.6	58.24	39.40	42.42
+ LM	39.25	52.68	31.25	32.11
multi	41.15	54.38	38.91	42.65
augAll	41.90	56.10	32.30	32.86
augTwo	**38.83**	**52.06**	**29.94**	**30.47**

5.3.1. Data augmentation

Ciphered transcriptions are generated from all the source languages for a target language using mapping models as described in Section 3. Then the audio data of source languages and the ciphered transcriptions are used together as augmented data for retraining of target language ASR (*augAll*). As described earlier, the quality of ciphered transcriptions depends on performance of mapping models, using ciphered transcriptions data augmentation from all the source languages include very low quality transcriptions and have detrimental effect on retraining of target language ASR. So, the augmentation is then restricted to use ciphered data from only closest language (*augTwo*). For a target language, the source language with highest mapping model accuracy is chosen as the closest language. By augmenting this data for retraining of a target language, an relative gain of up to 5% is achieved in terms of CER (*augTwo*).

6. Conclusion

In this work, the technique of mapping models is extended for e2e speech recognition systems. For a given target language, a mapping model is trained on limited amount of data to transform output posterior distributions from a source language ASR model to that of the target language. A source language ASR followed by a mapping model is then used for cross-lingual speech recognition in low-resource setting. Mapping models are further exploited to transliterate data of a source language to the target language for data augmentation. Retraining of target language ASR after data augmentation results in a relative CER reduction of up to 5% and 28.5% in comparison to monolingual and multilingual ASR systems respectively.

Table 4: *Cross-lingual ASR performance in terms of %CER*

Target Lang	*Source Languages* accuracy			
	tam	*tel*	*ceb*	*jav*
tam	*44.60*	49.34	49.21	49.03
tel	63.19	*58.24*	64.33	63.65
ceb	48.10	65.31	*39.40*	40.94
jav	46.72	56.92	40.88	*42.42*

7. References

[1] "Languages of the world," https://www.ethnologue.com/guides/how-many-languagesm, accessed: 2022-10-20.

[2] S. T. Abate, M. Y. Tachbelie, and T. Schultz, "Multilingual acoustic and language modeling for ethio-semitic languages," in *Proc. Interspeech 2020*, 2020, pp. 1047–1051.

[3] M. Y. Tachbelie, S. T. Abate, and T. Schultz, "Development of multilingual asr using globalphone for less-resourced languages: The case of ethiopian languages," in *Proc. Interspeech 2020*, 2020, pp. 1032–1036.

[4] M. Karafiát, M. K. Baskar, P. Matějka, K. Veselý, F. Grézl, and J. Černocky, "Multilingual blstm and speaker-specific vector adaptation in 2016 but babel system," in *IEEE SLT*, 2016, pp. 637–643.

[5] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Communication*, vol. 56, pp. 85–100, 2014.

[6] D. Imseng, P. Motlicek, H. Bourlard, and P. Garner, "Using out-of-language data to improve an under-resourced speech recognizer," *Speech Communication*, vol. 56, p. 142–151, 01 2014.

[7] N. T. Vu and T. Schultz, "Multilingual multilayer perceptron for rapid language adaptation between and across language families," in *Proc. Interspeech 2013*, 2013, pp. 515–519.

[8] F. Grézl, M. Karafiát, and K. Veselý, "Adaptation of multilingual stacked bottle-neck neural network structure for new language," in *ICASSP*, 2014, pp. 7654–7658.

[9] K. Veselý, M. Karafiát, F. Grézl, M. Janda, and E. Egorova, "The language-independent bottleneck features," in *IEEE SLT*, 2012, pp. 336–341.

[10] A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual training of deep neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7319–7323.

[11] S. Tong, P. N. Garner, and H. Bourlard, "Cross-lingual adaptation of a ctc-based multilingual acoustic model," *Speech Communication*, vol. 104, pp. 39–46, 2018.

[12] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7304–7308.

[13] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, "Deep speech: Scaling up end-to-end speech recognition," 2014.

[14] M. J. F. Gales, A. Ragni, H. AlDamarki, and C. Gautier, "Support vector machines for noise robust asr," in *2009 IEEE Workshop on Automatic Speech Recognition and Understanding*, 2009, pp. 205–210.

[15] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. Interspeech 2015*, 2015, pp. 3586–3589.

[16] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (vtlp) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, vol. 117, 2013, p. 21.

[17] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1469–1477, 2015.

[18] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 2613–2617. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2680

[19] A. Ragni, K. M. Knill, S. P. Rath, and M. J. F. Gales, "Data augmentation for low resource languages," in *Proc. Interspeech 2014*, 2014, pp. 810–814.

[20] J. Emond, B. Ramabhadran, B. Roark, P. Moreno, and M. Ma, "Transliteration based approaches to improve code-switched speech recognition performance," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 448–455.

[21] A. Datta, B. Ramabhadran, J. Emond, A. Kannan, and B. Roark, "Language-agnostic multilingual modeling," in *ICASSP*, 2020, pp. 8239–8243.

[22] S. Thomas, K. Audhkhasi, and B. Kingsbury, "Transliteration Based Data Augmentation for Training Multilingual ASR Acoustic Models in Low Resource Settings," in *Proc. Interspeech 2020*, 2020, pp. 4736–4740.

[23] M. U. Farooq and T. Hain, "Investigating the Impact of Crosslingual Acoustic-Phonetic Similarities on Multilingual Speech Recognition," in *Proc. Interspeech 2022*, 2022, pp. 3849–3853.

[24] M. U. Farooq, D. A. H. Narayana, and T. Hain, "Non-Linear Pairwise Language Mappings for Low-Resource Multilingual Acoustic Model Fusion," in *Proc. Interspeech 2022*, 2022, pp. 4850–4854.

[25] E. Roszkowska, "Rank ordering criteria weighting methods – a comparative overview," *Optimum. Studia Ekonomiczne*, no. 5(65), p. 14–33, 2013.

[26] M. J. F. Gales, K. M. Knill, A. Ragni, and S. P. Rath, "Speech recognition and keyword spotting for low-resource languages: Babel project research at CUED," in *Proc. 4th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2014)*, 2014, pp. 16–23.

[27] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 4835–4839.

[28] T. Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," 2018.

[29] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "Speechbrain: A general-purpose speech toolkit," 2021.