
Proceedings Paper:
Farooq, M.U., Haniya Narayana, D.A. and Hain, T. (2022) Non-linear pairwise language mappings for low-resource multilingual acoustic model fusion. In: Interspeech 2022 - 23rd Annual Conference of the International Speech Communication Association. Interspeech 2022 - Human and Humanizing Speech Technology, 18-22 Sep 2022, Incheon, Korea. International Speech Communication Association, pp. 4850-4854. ISSN: 1990-9772.

https://doi.org/10.21437/Interspeech.2022-11449

Interspeech 2022
18-22 September 2022, Incheon, Korea

Non-Linear Pairwise Language Mappings for Low-Resource Multilingual Acoustic Model Fusion

Muhammad Umar Farooq, Darshan Adiga Haniya Narayana, Thomas Hain

Speech and Hearing Research Group, University of Sheffield, UK.

mufarooq1@sheffield.ac.uk

Abstract

Multilingual speech recognition has drawn significant attention as an effective way to compensate data scarcity for low-resource languages. End-to-end (e2e) modelling is preferred over conventional hybrid systems, mainly because of no lexicon requirement. However, hybrid DNN-HMMs still outperform e2e models in limited data scenarios. Furthermore, the problem of manual lexicon creation has been alleviated by publicly available trained models of grapheme-to-phoneme (G2P) and text to IPA transliteration for a lot of languages. In this paper, a novel approach of hybrid DNN-HMM acoustic models fusion is proposed in a multilingual setup for the low-resource languages. Posterior distributions from different monolingual acoustic models, against a target language speech signal, are fused together. A separate regression neural network is trained for each source-target language pair to transform posteriors from source acoustic model to the target language. These networks require very limited data as compared to the ASR training. Posterior fusion yields a relative gain of 14.65% and 6.5% when compared with multilingual and monolingual baselines respectively. Cross-lingual model fusion shows that the comparable results can be achieved without using posteriors from the language dependent ASR.

Index Terms: automatic speech recognition, low-resource, model fusion, multilingual, cross-lingual

1. Introduction

With the advancements of the computational resources, many Deep Neural Networks (DNNs) architectures and networks have been proposed to make speech recognition more efficient and accurate. DNN-HMM hybrid systems [1] outperform conventional GMM-HMM systems. For end-to-end (e2e) speech recognition, sequence-to-sequence models [2], RNN transducers (RNN-T) [3], transformers [4] and unsupervised learning [5] are being used. These systems can be further improved with coupling of various techniques such as multi-task learning (MTL) [6], mixture of experts (MOE) [7] and learning hidden unit contributions (LHUC) [8] depending on the task. All these statistical modelling techniques require a lot of data for reliable parameters estimation. However, out of nearly 7000 languages being spoken around the world, just 23 languages are spoken by more than half of the world's population [9]. So, sufficient data resources are available for few languages.

Over the past decade, multilingual automatic speech recognition systems have stolen the limelight being an effective way to compensate the data scarcity for low-resource languages [10, 11, 12, 13, 14, 15]. DNN based multilingual acoustic models (AM) can be used to extract features to train a monolingual model [16, 17, 18] or multilingual models can directly be adapted to target language [19, 20]. Though e2e multilingual speech recognition systems are preferred over conventional ASR to avoid lexicon creations, DNN-HMMs still outperform e2e models in limited data scenarios such as low-resource languages. Furthermore, the advancement of G2P and text to IPA transliteration approaches such as Phonetisarus [21], Epitran [22] and open source LanguageNet G2P models [23] for many languages have alleviated the problem of manual creation of lexicons.

Previous work on e2e multilingual speech recognition systems has shown that a multilingual setup does not guarantee the reduction in Word Error Rate (WER) for target languages [5, 24, 25]. Recent efforts to interpret the learning of multilingual speech recognition systems [26, 27] observe that Phoneme Error Rate (PER) of an overlapped phoneme is not reduced with the growing number of sharing languages. The number of shared phonemes is not a reliable metric to measure language similarities and each participating language in a multilingual system has a different similarity with the target language. Even the balanced language data sampling can cause degradation or improvement due to internal acoustic-phonetic unbalancing [28]. It demands very controlled language mixing for a target language ASR.

To that end, a novel technique is proposed to fuse outputs of different monolingual models against the target language speech. Various previous studies on monolingual speech recognition have fused outputs from different models for acoustic [29, 30, 31] and language models [32, 33]. However, monolingual models have never been fused in a multilingual setup because it can not be done straightforwardly due to different phonetic decision trees of monolingual models. In this work, a separate regression neural network (*mapping network*) is trained for each <*source, target*> pair to map posteriors from a source language AM to the posteriors of the target language AM. The mapped posteriors are then fused in multilingual and cross-lingual fashion for phoneme recognition of the target language. The intuition is that a *mapping network* is able to learn some language related relationships between posterior distributions of source and target acoustic models. The proposed approach is helpful especially for low-resource languages because;

- the *mapping networks* can be trained with very limited amounts of data since a few hours can provide sufficient examples for phonetic level training.
- controlled fusion of posteriors based on language similarity will allow to control contribution of different source languages.

The mapped posteriors from the monolingual AMs are fused in a multilingual setup which not only outperforms the classical multilingual systems, but also the monolingual ASRs.

10.21437/Interspeech.2022-11449

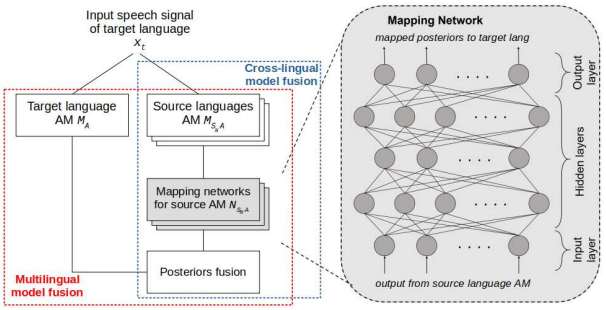


Figure 1: *Proposed system architecture*

2. Acoustic Model Fusion

Hybrid DNN-HMM systems outperform e2e ASRs where the amount of training data is limited [34]. Though the proposed fusion technique can be applied to e2e systems, the approach is described for DNN-HMM based ASRs here as a proof of concept.

In hybrid speech recognition systems, a deep neural network is trained to produce a posterior distribution of tied states of HMM models. Theoretically, the total number of tied states for a language with N number of phonemes and S number of states per HMM model is given by $N^n \times S$, where n is the context width. However, many polyphonemes never occur in a language and many are quite similar to the others. The total number of states is reduced by clustering many polyphonemes together. Each language yields a different phonetic decision tree in its monolingual ASR. Thus the number of tied states differs for each language and the posterior distributions are not directly comparable across the languages and thus not fusible.

Let M_A and M_{S_i} be the monolingual acoustic models of target and source languages respectively. A regression neural network 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. An underlying assumption is that this *mapping network* is able to learn some language-related relationships between posterior distributions of source and target acoustic models. For example, the network could learn phonemes of the target language which are more amenable to cross-lingual transfer than the others. Furthermore, a few hours of speech data can give thousands of examples that provide sufficient training data for *mapping network*. The proposed system architecture is shown in Fig. 1.

Let $X = \{x_1, x_2, \ldots, x_T\}$ be a set of observations of target language, for which posterior distributions ($P^Z = \{p_1, p_2, \ldots, p_T\}$ where $Z \in (A, S_iA)$) are attained from all monolingual acoustic models. A mapping network is trained using KL divergence loss to map posteriors from source acoustic models (P^{S_i}) to the target language posteriors (P^{S_iA}). The loss function for a batch is given as;

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

where N is the batch size for training a mapping network N_{S_iA} to map posteriors from i^{th} source language to the target language.

Posterior distributions from target AM and mapped distributions from source AMs are fused together for phoneme recognition of a target language. For a given observation at time t, the final posterior vector is given as;

$$p_t^F = w_T \cdot p_t^A + \sum_{i=1}^{K} w_i \cdot p_t^{S_iA} \tag{2}$$

where w_i are the scalar weights assigned to each posterior vector such that $\sum w_i = 1$ and K is the number of source languages.

In the experimentation, model fusion is done in multilingual and cross-lingual settings. In cross-lingual settings, only the mapped posteriors from source models are fused (the term $w_T \cdot p_t^A$ is omitted from Equation 2). The cross-lingual setting avoids using target AM which is helpful for low-resource languages. The weights are assigned to the fusing languages on the basis of similarity of the source and the target language. The study on cross-lingual acoustic-phonetic similarities using the same *mapping network* approach observes that the entropy of a <*source, target*> mapping network shows the language similarities [28]. The same similarity measure is used along with mapping network accuracy to assign the weights.

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

Lang	Train		Eval	
	# hours	# spks	# hours	# spks
Tamil *(tam)*	110.67	372	16.08	61
Telugu *(tel)*	67.27	243	13.92	60
Cebuano *(ceb)*	74.26	239	15.51	60
Javanese *(jav)*	76.39	242	16.25	60

3. Experimental Setup

3.1. Data set

In this work, experiments are done using four low-resource languages from IARPA BABEL speech corpus [35]. Full Language Packs (FLP) of Tamil (*tam*), Telugu (*tel*), Cebuano (*ceb*) and Javanese (*jav*) are used for baseline ASR training and evaluation. Since the eval data of BABEL is not publicly available, train and dev sets of BABEL data sets are used as train and eval sets respectively for the experiments. The details of the data sets are tabulated in Table 1. These data consist of conversational telephone speech and are quite challenging because of limited bandwidth, conversational styles, channel and background environment conditions. A limited amount of scripted read speech is also included in each language pack.

Full amounts are used for the training of baseline monolingual and multilingual speech recognition models. Multilingual ASR is trained by mixing data from all the languages. However, for training of the mapping networks, a subset of 30 hours

Table 2: *Examples (in millions) for training of mapping networks for each target language. Train set: 29 hours; Dev set: 1 hour; Eval set is same as for the ASR*

Language	Train	Dev	Eval
Tamil *(tam)*	3.234	0.358	1.664
Telugu *(tel)*	3.232	0.356	1.915
Cebuano *(ceb)*	3.241	0.348	1.943
Javanese *(jav)*	3.225	0.365	1.854

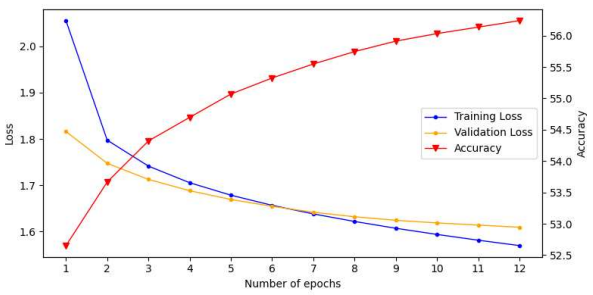


Figure 2: *Training curve of $N_{ceb-tel}$ mapping network*

is chosen from each language pack. Utterances containing only non-speech or silence are discarded while randomly sampling the 30 hours. This data is further divided randomly into 29 and 1 hour portions as train and dev sets to train the mapping networks. Since the mapping networks are trained on phonetic level, 30 hours provide millions of examples for the sufficient training of these models. The examples, used for building the mapping networks, are given in Table 2.

3.2. Baseline ASRs

Baseline monolingual and multilingual acoustic models are hybrid DNN-HMM models. 40 Mel-Frequency Cepstral Coefficients (MFCCs) are extracted for each frame of the speech signals using a window size of 25ms and a shift of 10ms. These features are then fed to DNN which is consisted of 12 factorised TDNN (TDNN-F) layers [36]. Each TDNN-F hidden layer is of dimension 1024, factorised with a linear *'bottleneck'* dimension of 128. The acoustic model is trained using lattice-free MMI criterion (LF-MMI) [34]. Neural network outputs posteriors of the clustered monophone classes. Clustering in each monolingual ASR training is different and thus the outputs from different acoustic models against an identical speech signal are not directly comparable. The experiments are done using Kaldi toolkit [37].

3.3. Mapping networks

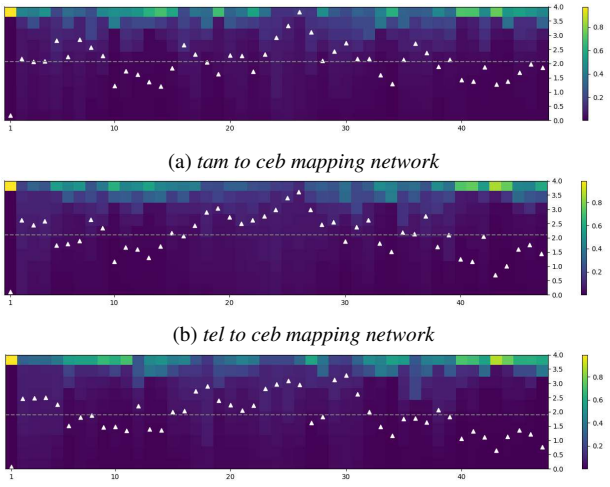
A regression neural network (*mapping network*) is trained for each source-target language pair ($N_{src-tgt}$). The neural network consists of three fully connected hidden layers to map the posterior distributions from dimensions of M_{S_i} to that of M_A. KL divergence loss is used for the training. As an accuracy measure of the mapping network, the number of correctly mapped frames is divided by the total number of frames as given in Equation 3. Correctly mapped frames are defined as the frames where the index of $max(mapped\_posteriors)$ is same as the index of $max(targetAM\_posteriors)$ which means that the output of the mapping network is mapped to the correct clustered phone class of the target AM.

$$(index(max(p_t^A)) == index(max(p_t^{S_i A}))) \Rightarrow CMF++ \quad (3)$$

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

where $index(x)$ returns the index of class x in the output vector, CMF is the number of correctly mapped frames and T is the total number of frames. The training curve with accuracy measure for one of the mapping networks is shown in Fig 2. Having millions of examples, training converges in early epochs for all the networks with accuracy of nearly 50%. However, analysis reveals that most of the times when mapping network is not accurate according to the aforementioned criteria, still the correct target AM class is usually among a few most probable classes of the mapped distribution. So, the network accuracy is recalculated considering top *n* classes of the *mapping network* output. The results in Table 3 show that the accuracy for most of the networks is dramatically increased to nearly 90% from less than 50% by considering top two most probable classes of the *mapping network* output rather than only one.

For weighted fusion of the posteriors from different acoustic models, entropy of each <*source, target*> mapping network is measured as a similarity measure. Entropy and posteriorgrams from mapping networks are shown for a sample *ceb* target file in Fig 3. Each box represents a frame on horizontal axis and probability of an output class on vertical axis. Output classes on vertical axis are sorted (from top to bottom) for each frame. Entropies of frames are plotted on right vertical



(a) *tam to ceb mapping network*

(b) *tel to ceb mapping network*

(c) *jav to ceb mapping network*

Figure 3: *Posteriorgram and entropy plot of $N_{tam-ceb}$, $N_{tel-ceb}$ and $N_{jav-ceb}$ for a sample ceb target speech utterance. Average entropy over the utterance is plotted in grey dashed line. Each box represents a frame on horizontal axis and probability of an output class on vertical axis*

Table 3: *Accuracy of the mapping networks considering top n mapped classes*

Target Lang	Source Lang	*Mapping network* accuracy			
		n=1	*n=2*	*n=5*	*n=10*
tam	tel	42.91	88.16	94.91	97.91
	ceb	44.43	84.63	91.82	96.13
	jav	41.89	85.82	92.82	96.69
tel	tam	54.44	92.08	96.87	98.58
	ceb	35.51	90.26	95.40	97.91
	jav	50.54	90.71	95.70	98.10
ceb	tam	45.73	85.50	93.71	97.23
	tel	46.17	87.87	93.98	97.51
	jav	47.04	88.50	94.67	98.03
jav	tam	47.81	85.63	93.36	96.58
	tel	48.29	86.31	93.74	97.03
	ceb	48.05	86.28	93.61	96.95

Table 4: *Entropy of the source-target mapping networks on eval set*

Target Lang.	Source Lang.			
	tam	tel	ceb	jav
tam	0	1.292	1.286	1.383
tel	1.109	0	1.161	1.139
ceb	1.214	1.235	0	1.098
jav	1.335	1.460	1.279	0

axis. The figure shows that *jav* to *ceb* mapping network has lower average entropy than mappings from *tam* and *tel* AMs. This is comprehensible due to the fact that *ceb* and *jav* are from same language family and thus *jav* is closer to *ceb* than *tam* and *tel*. For model fusion, entropies and accuracy of the mapping networks are considered while assigning the weights. The entropies for all the trained mapping networks are tabulated in the Table 4.

4. Results and Discussion

4.1. Baseline ASRs

Monolingual baseline systems (*mono*) are the language dependent acoustic and pronunciation models which are trained on a language specific data set. The train sets of all the languages are then mixed to train the multilingual models (*multi*). The results of the baseline systems are given in the Table 5 (in terms of PER). The results show that the error for all the languages is slightly increased in the baseline multilingual setup.

4.2. Model fusion

A multilingual acoustic model is imitated by fusing the target language and mapped source language posteriors. The fusion is the linear weighted sum of all these posterior distributions. In Table 5, the results of multilingual and cross-lingual model fusion settings (*multi-mf* and *cross-mf* respectively) are compared with *mono* and *multi* baseline ASRs. Multilingual model fusion yields a maximum gain of 6.5% over monolingual and 14.65% when compared with multilingual baseline systems.

Results of cross-lingual model fusion shows that without using the language dependent ASR, a comparable phoneme error rate for a target language can be achieved. For cross-lingual fusion, mapped posteriors from all the source language AMs are fused. However, the computation cost for fusing large number of languages incites us to minimise the number of fusing languages. For a given target language, further experiments are carried out using the mapped posteriors from only one source language at a time. Table 6 shows that nearly similar results as *cross-mf* can be achieved using mapped posteriors form the closest source language AM only. In the case of Telugu language, mapped posteriors from single AM model of Tamil perform even better than the cross-lingual model fusion. The first row for each language is same as *cross-mf* of Table 5 and following rows are the cross-lingual mapped posteriors from only one of the source language AM. For a target language, the change in results using different source language AMs can be seen in relation with mapping networks entropy of Table 4. For example in the case of Javanese target language, the entropy is highest for *tel-jav* mapping network and so the WER is highest for jav when using mapped posteriors from tel AM and so on.

For the model fusion, the weights are manually assigned to the posteriors which pose an issue of sub-optimal output. However, these weights could be learnt with the training of the mapping networks. Our next steps will include expanding the work for e2e ASRs and learning the weights jointly.

Table 5: *Baseline ASR performance in terms of % PER*

Lang	*mono*	*multi*	*multi-mf*	*cross-mf*
tam	43.96	43.67	**41.96**	55.47
tel	43.66	46.36	**42.05**	52.76
ceb	36.67	41.02	**35.54**	43.04
jav	41.60	45.54	**38.87**	47.79

Table 6: *Performance of model fusion in cross-lingual setting. 'Y' represents the source languages being fused together*

Target Language	Fused languages				% PER
	tam	*tel*	*ceb*	*jav*	
tam	N	Y	Y	Y	**55.47**
	N	Y	N	N	55.65
	N	N	Y	N	57.69
	N	N	N	Y	57.33
tel	Y	N	Y	Y	52.76
	Y	N	N	N	**52.37**
	N	N	Y	N	55.68
	N	N	N	Y	53.58
ceb	Y	Y	N	Y	**43.04**
	Y	N	N	N	45.94
	N	Y	N	N	45.28
	N	N	N	Y	43.91
jav	Y	Y	Y	N	**47.79**
	Y	N	N	N	48.40
	N	Y	N	N	48.90
	N	N	Y	N	48.25

5. Conclusion

In this work, a novel monolingual acoustic model fusion technique is proposed for low resource languages in a multilingual setup. Posterior distributions from different monolingual acoustic models against a target language speech signal are fused together. A separate regression neural network is trained for each source-target language pair to map posteriors from source acoustic model to the target language acoustic model. The mapping networks need very limited amount of data for training as compared to the ASR building. Multilingual model fusion yields a relative gain of 14.65% and 6.5% when compared with multilingual and monolingual baselines for the target language. Cross-lingual model fusion shows that the comparable results can be achieved without using the target language ASR.

6. Acknowledgements

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

7. References

[1] S. Madikeri, B. K. Khonglah, S. Tong, P. Motlicek, H. Bourlard, and D. Povey, "Lattice-Free Maximum Mutual Information Training of Multilingual Speech Recognition Systems," in *Proc. Interspeech 2020*, 2020, pp. 4746–4750.

[2] J. Cho, M. K. Baskar, R. Li, M. Wiesner, S. H. Mallidi, N. Yalta,

M. Karafiát, S. Watanabe, and T. Hori, "Multilingual sequence-to-sequence speech recognition: Architecture, transfer learning, and language modeling," in *IEEE SLT*, 2018, pp. 521–527.

[3] A. Kannan, A. Datta, T. N. Sainath, E. Weinstein, B. Ramabhadran, Y. Wu, A. Bapna, Z. Chen, and S. Lee, "Large-scale multilingual speech recognition with a streaming end-to-end model," 2019.

[4] V. M. Shetty and M. Sagaya Mary N.J., "Improving the performance of transformer based low resource speech recognition for indian languages," in *ICASSP*, 2020, pp. 8279–8283.

[5] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised Cross-Lingual Representation Learning for Speech Recognition," in *Proc. Interspeech 2021*, 2021, pp. 2426–2430.

[6] H. B. Sailor and T. Hain, "Multilingual Speech Recognition Using Language-Specific Phoneme Recognition as Auxiliary Task for Indian Languages," in *Proc. Interspeech 2020*, 2020, pp. 4756–4760.

[7] N. Gaur, B. Farris, P. Haghani, I. Leal, P. J. Moreno, M. Prasad, B. Ramabhadran, and Y. Zhu, "Mixture of informed experts for multilingual speech recognition," in *ICASSP*, 2021, pp. 6234–6238.

[8] S. Tong, "Multilingual training and adaptation in speech recognition," Ph.D. dissertation, Lausanne, 2020.

[9] "Languages of the worldn," https://www.ethnologue.com/guides/how-many-languagesm, accessed: 2022-03-27.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] J. R. Novak, N. Minematsu, and K. Hirose, "WFST-based grapheme-to-phoneme conversion: Open source tools for alignment, model-building and decoding," in *Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing*, 2012, pp. 45–49.

[22] D. R. Mortensen, S. Dalmia, and P. Littell, "Epitran: Precision G2P for many languages," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, May 2018.

[23] M. Hasegawa-Johnson, L. Rolston, C. Goudeseune, G.-A. Levow, and K. Kirchhoff, "Grapheme-to-phoneme transduction for cross-language asr," in *Statistical Language and Speech Processing*, L. Espinosa-Anke, C. Martín-Vide, and I. Spasić, Eds. Springer International Publishing, 2020, pp. 3–19.

[24] V. Pratap, A. Sriram, P. Tomasello, A. Hannun, V. Liptchinsky, G. Synnaeve, and R. Collobert, "Massively Multilingual ASR: 50 Languages, 1 Model, 1 Billion Parameters," in *Proc. Interspeech 2020*, 2020, pp. 4751–4755.

[25] W. Hou, Y. Dong, B. Zhuang, L. Yang, J. Shi, and T. Shinozaki, "Large-Scale End-to-End Multilingual Speech Recognition and Language Identification with Multi-Task Learning," in *Proc. Interspeech 2020*, 2020, pp. 1037–1041.

[26] P. Żelasko, L. Moro-Velázquez, M. Hasegawa-Johnson, O. Scharenborg, and N. Dehak, "That Sounds Familiar: An Analysis of Phonetic Representations Transfer Across Languages," in *Proc. Interspeech 2020*, 2020, pp. 3705–3709.

[27] S. Feng, P. Żelasko, L. Moro-Velázquez, A. Abavisani, M. Hasegawa-Johnson, O. Scharenborg, and N. Dehak, "How phonotactics affect multilingual and zero-shot asr performance," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7238–7242.

[28] M. U. Farooq and T. Hain, "Investigating the Impact of Cross-lingual Acoustic-Phonetic Similarities on Multilingual Speech Recognition," in *Submitted to Interspeech*, 2022.

[29] A. H. Abdelaziz, "Comparing fusion models for dnn-based audio-visual continuous speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 475–484, 2018.

[30] I. Rebai, Y. BenAyed, W. Mahdi, and J.-P. Lorré, "Improving speech recognition using data augmentation and acoustic model fusion," *Procedia Computer Science*, vol. 112, pp. 316–322, 2017.

[31] S. H. Mallidi and H. Hermansky, "Novel neural network based fusion for multistream asr," in *ICASSP*, 2016, pp. 5680–5684.

[32] C. Shan, C. Weng, G. Wang, D. Su, M. Luo, D. Yu, and L. Xie, "Component fusion: Learning replaceable language model component for end-to-end speech recognition system," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5361–5635.

[33] E. McDermott, H. Sak, and E. Variani, "A density ratio approach to language model fusion in end-to-end automatic speech recognition," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 434–441.

[34] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI," in *Proc. Interspeech 2016*, 2016, pp. 2751–2755.

[35] 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.

[36] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks," in *Proc. Interspeech 2018*, 2018, pp. 3743–3747.

[37] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, 2011.