Multilingual Graphemic Hybrid ASR with Massive Data Augmentation

Chunxi Liu, Qiaochu Zhang, Xiaohui Zhang, Kritika Singh, Yatharth Saraf, Geoffrey Zweig

Facebook AI
New York, NY, and Menlo Park, CA, USA
{chunxiliu,frankz,xiaohuizhang,skritika,ysaraf,gzweig}@fb.com

Abstract

Towards developing high-performing ASR for low-resource languages, approaches to address the lack of resources are to make use of data from multiple languages, and to augment the training data by creating acoustic variations. In this work we present a single grapheme-based ASR model learned on 7 geographically proximal languages, using standard hybrid BLSTM-HMM acoustic models with lattice-free MMI objective. We build the single ASR grapheme set via taking the union over each language-specific grapheme set, and we find such multilingual graphemic hybrid ASR model can perform language-independent recognition on all 7 languages, and substantially outperform each monolingual ASR model. Secondly, we evaluate the efficacy of multiple data augmentation alternatives within language, as well as their complementarity with multilingual modeling. Overall, we show that the proposed multilingual graphemic hybrid ASR with various data augmentation can not only recognize any within training set languages, but also provide large ASR performance improvements.

Keywords: Multilingual graphemic acoustic models, hybrid speech recognition, data augmentation

1. Introduction

It can be challenging to build high-accuracy automatic speech recognition (ASR) systems in the real world due to the vast language diversity and the requirement of extensive manual annotations on which the ASR algorithms are typically built. Series of research efforts have thus far been focused on guiding the ASR of a target language by using the supervised data from multiple languages.

Consider the standard hidden Markov models (HMM) based hybrid ASR system with a phonemic lexicon, where the vocabulary is specified by a pronunciation lexicon. One popular strategy is to make all languages share the same phonemic representations through a universal phonetic alphabet such as International Phonetic Alphabet (IPA) phone set (Lin et al., 2009; Liu et al., 2016; Pulugundla et al., 2018; Tong et al., 2019), or X-SAMPA phone set (Wells, 1995; Knill et al., 2013; Knill et al., 2014; Wiesner et al., 2018). In this case, multilingual joint training can be directly applied. Given the effective neural network based acoustic modeling, another line of research is to share the hidden layers across multiple languages while the softmax layers are language dependent (Huang et al., 2013; Heigold et al., 2013); such multitask learning procedure can improve ASR accuracies for both within training set languages, and also unseen languages after language-specific adaptation, i.e., cross-lingual transfer learning. Different nodes in hidden layers have been shown in response to distinct phonetic features (Nagamine et al., 2015), and hidden layers can be potentially transferable across languages. Note that the above works all assume the test language identity to be known at decoding time, and the language specific lexicon and language model applied.

In the absence of a phonetic lexicon, building graphemic systems has shown comparable performance to phonetic lexicon-based approaches in extensive monolingual evaluations (Kanthak and Ney, 2002; Gales et al., 2015; Trmal et al., 2017). Recent advances in end-to-end or sequence-to-sequence ASR models have attempted to take the union of multiple language-specific grapheme (i.e. orthographic character) sets, and use such union as a universal grapheme set for a single sequence-to-sequence ASR model (Watanabe et al., 2017; Toshniwal et al., 2018; Kim and Seltzer, 2018; Kannan et al., 2019). It allows for learning a grapheme-based model jointly on data from multiple languages, and performing ASR on within training set languages. In various cases it can produce performance gains over monolingual modeling that uses in-language data only.

Since HMM-based hybrid model remains a competitive ASR approach especially in low/medium-resource settings (Lüscher et al., 2019; Wang et al., 2020), in our work, we aim to examine the same approach above of building a multilingual graphemic lexicon, while using a hybrid ASR system – based on Bidirectional Long Short-Term Memory (BLSTM) and HMM – learned with lattice-free maximum mutual information (MMI) objective (Povey et al., 2016). Our initial attempt is on building a single cascade of an acoustic model, a phonetic decision tree, a graphemic lexicon and a language model – for 7 geographically proximal languages that have little overlap in their character sets. We evaluate it in a low resource context where each language has around 160 hours training data. We find that, despite the lack of explicit language identification (ID) guidance, our multilingual graphemic hybrid ASR model can accurately produce ASR transcripts in the correct test language scripts, and provide higher ASR accuracies than each language-specific ASR model. We further examine if using a subset of closely related languages – along language family or orthography – can achieve the same performance improvements as using all 7 languages.

Though extensive end-to-end or sequence-to-sequence ASR works have been built on multilingual graphemic models, to the best of our knowledge, there is no prior work in hybrid ASR that uses a single multilingual graphemic lexicon (rather than an IPA or X-SAMPA based phonetic lexicon) for multiple training languages. In this work, we show for the first time that multilingual graphemic hybrid ASR can provide large improvements across all training languages, even though almost each training language has distinct graphemic set.

We proceed with our investigation on various data augmentation techniques to overcome the lack of training data in

the above low-resource setting. Given the highly scalable neural network acoustic modeling, extensive alternatives to increasing the amount or diversity of existing training data have been explored in prior works, e.g., applying vocal tract length perturbation and speed perturbation (Ko et al., 2015), volume perturbation and normalization (Peddinti et al., 2015), additive noises (Amodei et al., 2016), reverberation (Peddinti et al., 2015; Ko et al., 2017; Kim et al., 2017), and SpecAugment (Park et al., 2019). In this work we focus particularly on techniques that mostly apply to our wildly collected video datasets. In comparing their individual and complementary effects, we aim to answer: (i) if there is benefit in scaling the model training to significantly larger quantities, e.g., up to 9 times greater than the original training set size, and (ii) if any, is the data augmentation efficacy comparable or complementary with the above multilingual modeling.

Improving accessibility to videos "in the wild" such as automatic captioning on YouTube has been studied in (Liao et al., 2013; Soltau et al., 2017). While allowing for applications like video captions, indexing and retrieval, transcribing the heterogeneous social media videos of extensively diverse languages is highly challenging for ASR systems. On the whole, we present empirical studies in building a single multilingual graphemic hybrid ASR model capable of language-independent decoding on multiple languages, and in effective data augmentation techniques for video datasets.

2. Multilingual Graphemic Hybrid ASR

In this section we first briefly describe our deployed hybrid ASR architecture based on the weighted finite-state transducers (WFSTs) outlined in (Mohri et al., 2008). Then we present its extension to multilingual training. Lastly, we discuss its language-independent decoding and language-specific decoding.

2.1. Graphemic ASR with WFST

In the ASR framework of a hybrid BLSTM-HMM, the decoding graph can be interpreted as a composed WFST of cascade $H \circ C \circ L \circ G$. Acoustic models, i.e. BLSTMs, produce acoustic scores over context-dependent HMM (i.e. triphone) states. A WFST H, which represents the HMM set, maps the triphone states to context-dependent phones. While in graphemic ASR, the notion of phone is turned to grapheme, and we typically create the grapheme set via modeling each orthographic character as a separate grapheme. Then a WFST C maps each context-dependent grapheme, i.e. tri-grapheme, to an orthographic character. The lexicon L is specified where each word is mapped to a sequence of characters forming that word. G encodes either the transcript during training, or a language model during decoding.

2.2. A Single Multilingual ASR Model Using Lattice-Free MMI

To build a single grapheme-based acoustic model for multiple languages, a multilingual graphemic set is obtained by taking a union of each grapheme set from each language considered, each of which can be either overlapping or non-overlapping. In the multilingual graphemic lexicon, each

word in any language is mapped to a sequence of characters in that language.

A context-dependent acoustic model is constructed using the decision tree clustering of tri-grapheme states, in the same fashion as the context dependent triphone state tying (Young et al., 1994). The graphemic-context decision tree is constructed over all the multilingual acoustic data including each language of interest. The optimal number of leaves for the multilingual model tends to be larger than for a monolingual neural network.

The acoustic model is a BLSTM network, using sequence discriminative training with lattice-free MMI objective (Povey et al., 2016). The BLSTM model is bootstrapped from a standard Gaussian mixture model (GMM)-HMM system. A multilingual n-gram language model is learned over the combined transcripts including each language considered.

2.3. Language-Independent and Language-Specific Decoding in the WFST Framework

Given the multilingual lexicon and language model, the multilingual ASR above can decode any within training set language, even though not explicitly given any information about language identity. We refer to it as language-independent decoding or multilingual decoding. Note that such ASR can thus far produce any word in the multilingual lexicon, and the hypothesized word can either be in the vocabulary of the considered test language, or out of test language vocabulary as a mismatched-language error.

We further consider applying language-specific decoding, assuming the test language identity to be known at decoding time. Again consider the decoding graph $H \circ C \circ L \circ G$, and H & C are thus multilingual while the lexicon L and language model G can include only the words in test language vocabulary. The multilingual acoustic model can therefore make use of multilingual training data, while its language-specific decoding operation only produces monolingual words matched with test language identity.

3. Data Augmentation

In this section, we consider 3 categories of data augmentation techniques that are effectively applicable to video datasets.

3.1. Speed and Volume Perturbation

Both speed and volume perturbation emulate mean shifts in spectrum (Ko et al., 2015; Peddinti et al., 2015). To perform speed perturbation of the training data, we produce three versions of each audio with speed factors 0.9, 1.0, and 1.1. The training data size is thus tripled. For volume perturbation, each audio is scaled with a random variable drawn from a uniform distribution $[0.125, 2]$.

3.2. Additive Noise

To further increase training data size and diversity, we can create new audios via superimposing each original audio with additional noisy audios in time domain. To obtain diverse noisy audios, we use AudioSet, which consists of 632 audio event classes and a collection of over 2 million

manually-annotated 10-second sound clips from YouTube videos (Gemmeke et al., 2017).

Note that in our video datasets, video lengths vary between 10 seconds and 5 minutes, with an average duration of about 2 minutes. Rather than constantly repeating the 10-second sound clip to match the original minute-long audio, we superimpose each sound clip on the short utterances via audio segmentation. Specifically, we first use an initial bootstrap model to align each original long audio, and segment each audio into around 10-second utterances via word boundaries. Then for each utterance in the original train set, we can create a new noisy utterance by the steps:

1. Sample a sound clip from AudioSet.

2. Trim or repeat the sound clip as necessary to match the duration of the original utterance.

3. Sample a signal-to-noise ratio (SNR) from a Gaussian distribution with mean 10, and round the SNR up to 0 or down to 20 if the sample is beyond 0-20dB. Then scale the sound clip signal to obtain the target SNR.

4. Superimpose the original utterance signal with the scaled sound clip signal in time domain to create the resulting utterance.

Thus for each original utterance, we can create a variable number of new noisy utterances via sampling sound clips. We use a 3-fold augmentation that combines the original train set with two noisy copies.

3.3. SpecAugment

We consider applying the frequency and time masking techniques – which are shown to greatly improve the performance of end-to-end ASR models (Park et al., 2019) – to our hybrid systems. Similarly, they can be applied online during each epoch of LF-MMI training, while time warping requires the need for realignment and thus does not fit hybrid model training.

Consider each utterance (i.e. after the audio segmentation in Section 3.2.), and we compute its log mel spectrogram with ν dimension and τ time steps:

1. Frequency masking is applied m_F times, and each time the frequency bands $[f_0, f_0 + f)$ are masked, where f is sampled from $[0, F]$ and f_0 is sampled from $[0, \nu - f)$.

2. Time masking is optionally applied m_T times, and each time the time steps $[t_0, t_0 + t)$ are masked, where t is sampled from $[0, T]$ and t_0 is sampled from $[0, \tau - t)$.

As in (Park et al., 2019), we increase the training schedule accordingly, i.e., number of epochs.

4. Experiments

4.1. Data

Our multilingual ASR attempt was on 7 geographically proximal languages: Kannada, Malayalam, Sinhala, Tamil, Bengali, Hindi and Marathi. The datasets were a set of public social media videos, which were wildly collected and anonymized. We categorized them into four sets: clean, noisy, extremeI (xtrmI) and extremeII

Language	Train	Test			
		clean	noisy	xtrmI	xtrmII
Kannada	125.5	1.5	9.9	0.8	2.7
Malayalam	127.7	4.5	9.2	0.7	1.0
Sinhala	160.0	13.9	25.0	8.6	8.8
Tamil	176.9	2.8	16.4	0.5	0.7
Bengali	160.0	7.4	24.9	25.0	16.4
Hindi	160.0	22.2	21.5	19.4	19.8
Marathi	148.6	2.7	13.7	0.3	0.5

Table 1: The amounts of audio data in hours.

(xtrmII). xtrmI differed from xtrmII in chronological order, and were both more acoustically challenging than clean and noisy categories.

For each language, the train and test set size are described in Table 1, and most training data were of noisy category. On each language we also had a small validation set for model parameter tuning. Each monolingual ASR baseline was trained on language-specific data only.

To create the grapheme set, we consult the unicode character ranges of each language, and also include apostrophe, hyphen and zero width joiner in the final character sets. The character sets of these 7 languages have little overlap except that (i) they all include common basic Latin alphabet, and (ii) both Hindi and Marathi use Devanagari script. We took the union of 7 character sets therein as the multilingual grapheme set (Section 2.2.), which contained 432 characters. In addition, we deliberately split 7 languages into two groups, such that the languages within each group were more closely related in terms of language family, orthography or phonology. We thus built 3 multilingual ASR models trained on:

(i) all 7 languages, for 1059 training hours in total,

(ii) 4 languages – Kannada, Malayalam, Sinhala and Tamil – for 590 training hours,

(iii) 3 languages – Bengali, Hindi and Marathi – for 469 training hours,

which are referred to as *7lang*, *4lang*, and *3lang* respectively. Note that Kannada, Malayalam and Tamil are Dravidian languages, which have rich agglutinative inflectional morphology (Pulugundla et al., 2018) and resulted in around 10% OOV token rates on test sets (Hindi had the lowest OOV rate as 2-3%). Such experimental setup was designed to answer the questions:

(i) If a single graphemic ASR model could scale its language-independent recognition up to all 7 languages.

(ii) If including all 7 languages could yield better ASR performance than using a small subset of closely related languages.

4.2. Model Configurations

Each bootstrap model was a GMM-HMM based system with speaker adaptive training, implemented with Kaldi

48

Language	Model	clean	noisy	xtrmI	xtrmII	Average	% Gain
Kannada	monolingual	56.9	56.6	58.7	57.6	57.5	–
	monolingual + fm	53.3	54.8	56.9	56.4	55.4	3.7
	monolingual + sp	53.1	54.7	56.4	55.2	54.9	4.5
	monolingual + fm + sp	50.3	53.1	**54.8**	53.9	53.0	7.8
	monolingual + sp + noise	50.7	53.3	54.8	53.6	53.1	7.7
	monolingual + fm + sp + noise	**49.7**	**52.5**	54.9	**52.7**	52.5	8.7
	multilingual, *4lang*	50.2	**53.4**	55.7	**53.4**	**53.2**	7.5
	multilingual, *7lang*	**49.7**	53.5	**54.9**	55.6	53.4	7.1
	multilingual, *7lang + lang-specific decoding*	*49.4*	*52.5*	*54.6*	*53.7*	*52.5*	*8.7*
	multilingual, *7lang* + fm + sp + noise	**46.6**	52.0	53.0	53.3	51.2	**11.0**
Malayalam	monolingual	56.5	53.2	70.3	55.9	59.0	–
	multilingual, *4lang*	52.8	**51.6**	**65.8**	53.4	**55.9**	**5.3**
	multilingual, *7lang*	**52.1**	51.9	66.3	54.0	56.1	5.0
Sinhala	monolingual	45.4	39.5	62.7	51.8	49.9	–
	multilingual, *4lang*	**42.1**	38.4	59.7	50.3	47.6	4.6
	multilingual, *7lang*	42.9	**38.3**	**59.3**	**49.9**	47.6	**4.6**
Tamil	monolingual	44.2	44.4	49.0	52.7	47.6	–
	multilingual, *4lang*	40.7	42.8	46.6	**50.9**	45.2	5.0
	multilingual, *7lang*	**40.1**	**42.7**	**46.1**	51.7	**45.2**	**5.0**
Bengali	monolingual	53.4	50.8	68.2	58.0	57.6	–
	multilingual, *3lang*	**45.5**	**47.0**	**62.6**	**53.3**	**52.1**	**9.5**
	multilingual, *7lang*	45.7	48.1	63.9	54.7	53.1	7.8
Hindi	monolingual	36.9	38.2	58.4	45.0	44.6	–
	monolingual + fm	33.2	34.8	54.1	40.9	40.8	8.5
	monolingual + sp	33.6	34.9	55.0	41.1	41.2	7.6
	monolingual + fm + sp	32.1	33.4	52.7	39.5	39.4	11.7
	monolingual + sp + noise	32.0	33.5	52.6	39.5	39.4	11.7
	monolingual + fm + sp + noise	**30.9**	**32.2**	**50.7**	**38.2**	**38.0**	**14.8**
	multilingual, *3lang*	32.2	33.9	**53.5**	40.3	40.0	10.3
	multilingual, *7lang*	**31.9**	33.8	53.6	40.8	**40.0**	**10.3**
	multilingual, *7lang + lang-specific decoding*	*31.8*	*33.4*	*52.7*	*40.1*	*39.5*	*11.4*
	multilingual, *7lang* + fm + sp + noise	**28.5**	**30.8**	**49.6**	**36.7**	**36.4**	**18.4**
Marathi	monolingual	38.2	39.8	63.2	49.0	47.6	–
	multilingual, *3lang*	**34.9**	**37.4**	**56.4**	46.3	**43.7**	**8.2**
	multilingual, *7lang*	35.2	38.1	56.5	**46.1**	44.0	7.6

Table 2: WER results on each video dataset. Frequency masking is denoted by fm, speed perturbation by sp, and additive noise (Section 3.2.) by noise. 3lang, 4lang and 7lang denote the multilingual ASR models trained on 3, 4 and 7 languages, respectively, as in Section 4.1.. Lang-specific decoding denotes using multilingual acoustic model with language-specific lexicon and language model, as in Section 2.3.. Average is unweighted average WER across 4 video types. Gain (%) is the relative reduction in the Average WER over each monolingual baseline.

(Povey et al., 2011). Each neural network acoustic model was a latency-controlled BLSTM (Zhang et al., 2016), learned with lattice-free MMI objective and Adam optimizer (Kingma and Ba, 2015). All neural networks were implemented with Caffe2 (Hazelwood et al., 2018). Due to the production real time factor (RTF) requirements, we used the same model size in all cases – a 4 layer latency-controlled BLSTM network with 600 cells in each layer and direction – except that, the softmax dimensions, i.e. the optimal decision tree leaves, were determined through experiments on validation sets, varying within 7-30k. Input acoustic features were 80-dimensional log-mel filterbank coefficients. After lattice-free MMI training, the model with the best accuracy on validation set was used for evaluation on test set. We used standard 5-gram language models in all cases. Each multilingual 5-gram language model is learned simply via combining transcripts of each language.

4.3. Results with Multilingual ASR

ASR word error rate (WER%) results are shown in Table 2. We found that, although not explicitly given any information on test language identities, multilingual ASR with language-independent decoding (Section 2.3.) - trained on 3, 4, or 7 languages - substantially outperformed each monolingual ASR in all cases, and on average led to relative WER reductions between 4.6% (Sinhala) and 10.3% (Hindi).

Note that, in contrast to the multilingual phonetic hybrid ASR (i.e. using phonetic lexicons), it is intuitive to see ASR performance improve when different languages share the same phone set via IPA or X-SAMPA , since each phonetic

modeling can use more training data than monolingual training. However, in our multilingual graphemic ASR, only 2 of 7 training languages overlapped in character sets; for the first time, we show that, such multilingual graphemic-context decision tree based hybrid ASR can still improve performance for all languages.

Also, the word hypotheses from language-independent decoding could be language mismatched, e.g., part of a Kannada utterance was decoded into Marathi words. So we counted how many word tokens in the decoding transcripts were not in the lexicon of corresponding test language. We found in general only 1-3% word tokens are language mismatched, indicating that the multilingual model was very effective in identifying the language implicitly and jointly recognizing the speech.

Consider the scenario that, test language identities are known likewise in each monolingual ASR, and we proceed with language-specific decoding (Section 2.3.) on Kannada and Hindi, via language-specific lexicon and language model at decoding time. We found that, the language-specific decoding provided only moderate gains, presumably as discussed above, the language-independent decoding had given the mismatched-language word token rates as sufficiently low as 1-3%.

Additionally, the multilingual ASR of *4lang* and *3lang* (Section 4.1.) achieved the same, or even slightly better performance as compared to the ASR of *7lang*, suggesting that incorporating closely related languages into multilingual training is most useful for improving ASR performance. However, the *7lang* ASR by itself still yields the advantage in language-independent recognition of more languages.

4.4. Results with Data Augmentation

First, we experimented with monolingual ASR on Kannada and Hindi, and performed comprehensive evaluations of the data augmentation techniques described in Section 3.. As in Table 2, the performance gains of using frequency masking were substantial and comparable to those of using speed perturbation, where $m_F = 2$ and $F = 15$ (Section 3.3.) worked best. In addition, combining both frequency masking and speed perturbation could provide further improvements. However, applying additional volume perturbation (Section 3.1.) or time masking (Section 3.3.) was not helpful in our monolingual experiments, and we omit showing the results in the table.

Note that after speed perturbation, the training data tripled, to which we could apply another 3-fold augmentation based on additive noise (Section 3.2.), and the final train set was thus 9 times the size of original train set. We found that all 3 techniques were complementary, and in combination led to large fusion gains over each monolingual baseline – relative WER reductions of 8.7% on Kannada, and 14.8% on Hindi. Secondly, we applied the 3 data augmentation techniques to the multilingual ASR of *7lang*, and tested their additive effects. We show the resulting WERs on Kannada and Hindi in Table 2. Note that on Kannada, we found around 7% OOV token rate on `clean` but around 10-11% on other 3 test sets, and we observed more gains on `clean`; presumably because the improved acoustic model could only correct the in-vocabulary word errors, lower OOV rates therefore

left more room for improvements. Hindi had around 2.5% OOV rates on each test set, and we found incorporating data augmentation into multilingual ASR led to on average 9.0% relative WER reductions.

Overall, we demonstrated the multilingual hybrid ASR with massive data augmentation – via a single graphemic model even without the use of explicit language ID – allowed for relative WER reductions of 11.0% on Kannada and 18.4% on Hindi.

5. Conclusion

Multilingual training have been extensively studied in conventional phonetic hybrid ASR (Lin et al., 2009; Knill et al., 2013) and the recent end-to-end ASR (Watanabe et al., 2017; Toshniwal et al., 2018). In our work, for the first time, we demonstrate that a multilingual grapheme-based hybrid ASR model can effectively perform language-independent recognition on any within training set languages, and substantially outperform each monolingual ASR alternative. Various data augmentation techniques can yield further complementary improvements. Such single multilingual model can not only provide better ASR performance, but also serves as an alternative to a typical production deployment, which typically includes extensive monolingual ASR systems and a separate language ID model. The proposed approach of building a single multilingual graphemic hybrid ASR model without requiring individual language ID - while being especially competitive in low-resource settings - can greatly simplify the productionizing and maintenance process.

Additionally, as compared to the multilingual multitask learning plus monolingual fine-tuning methods in (Huang et al., 2013; Heigold et al., 2013), our preliminary experimentation shows that our proposed approach above can give comparable performance without requiring separate language ID guidance during decoding. We leave the detailed studies to the future work. Also, future work will expand the language coverage to include both geographically proximal and distant languages.

6. Acknowledgements

7. Bibliographical References

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., et al. (2016). Deep speech 2: End-to-end speech recognition in English and Mandarin. In *International conference on machine learning*, pages 173–182.

Gales, M. J., Knill, K. M., and Ragni, A. (2015). Unicode-based graphemic systems for limited resource languages. In *Proc. ICASSP*.

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. (2017). Audio set: An ontology and human-labeled dataset for audio events. In *Proc. ICASSP*.

Hazelwood, K., Bird, S., Brooks, D., Chintala, S., Diril, U., Dzhulgakov, D., Fawzy, M., Jia, B., Jia, Y., Kalro, A., et al. (2018). Applied machine learning at facebook: A data-center infrastructure perspective. In *IEEE International Symposium on High Performance Computer Architecture (HPCA)*.

Heigold, G., Vanhoucke, V., Senior, A., Nguyen, P., Ranzato, M., Devin, M., and Dean, J. (2013). Multilingual acoustic models using distributed deep neural networks. In *Proc. ICASSP*.

Huang, J.-T., Li, J., Yu, D., Deng, L., and Gong, Y. (2013). Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In *Proc. ICASSP*.

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

Kanthak, S. and Ney, H. (2002). Context-dependent acoustic modeling using graphemes for large vocabulary speech recognition. In *Proc. ICASSP*.

Kim, S. and Seltzer, M. L. (2018). Towards language-universal end-to-end speech recognition. In *Proc. ICASSP*.

Kim, C., Misra, A., Chin, K., Hughes, T., Narayanan, A., Sainath, T., and Bacchiani, M. (2017). Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in google home. In *Proc. Interspeech*.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *The International Conference on Learning Representations (ICLR)*.

Knill, K. M., Gales, M. J., Rath, S. P., Woodland, P. C., Zhang, C., and Zhang, S.-X. (2013). Investigation of multilingual deep neural networks for spoken term detection. In *Proc. ASRU*.

Knill, K. M., Gales, M. J., Ragni, A., and Rath, S. P. (2014). Language independent and unsupervised acoustic models for speech recognition and keyword spotting. In *Proc. Interspeech*.

Ko, T., Peddinti, V., Povey, D., and Khudanpur, S. (2015). Audio augmentation for speech recognition. In *Proc. Interspeech*.

Ko, T., Peddinti, V., Povey, D., Seltzer, M. L., and Khudanpur, S. (2017). A study on data augmentation of reverberant speech for robust speech recognition. In *Proc. ICASSP*.

Liao, H., McDermott, E., and Senior, A. (2013). Large scale deep neural network acoustic modeling with semi-supervised training data for YouTube video transcription. In *Proc. ASRU*.

Lin, H., Deng, L., Yu, D., Gong, Y.-f., Acero, A., and Lee, C.-H. (2009). A study on multilingual acoustic modeling for large vocabulary ASR. In *Proc. ICASSP*.

Liu, C., Jyothi, P., Tang, H., Manohar, V., Sloan, R., Kekona, T., Hasegawa-Johnson, M., and Khudanpur, S. (2016). Adapting ASR for under-resourced languages using mismatched transcriptions. In *Proc. ICASSP*.

Lüscher, C., Beck, E., Irie, K., Kitza, M., Michel, W., Zeyer,

A., Schlüter, R., and Ney, H. (2019). RWTH ASR systems for librispeech: Hybrid vs attention-w/o data augmentation. In *Proc. Interspeech*.

Mohri, M., Pereira, F., and Riley, M. (2008). Speech recognition with weighted finite-state transducers. In *Springer Handbook of Speech Processing*, pages 559–584. Springer.

Nagamine, T., Seltzer, M. L., and Mesgarani, N. (2015). Exploring how deep neural networks form phonemic categories. In *Proc. Interspeech*.

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V. (2019). SpecAugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Peddinti, V., Chen, G., Manohar, V., Ko, T., Povey, D., and Khudanpur, S. (2015). JHU ASPIRE system: Robust LVCSR with tdnns, ivector adaptation and RNN-LMs. In *Proc. ASRU*.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlíček, P., Qian, Y., Schwarz, P., et al. (2011). The Kaldi speech recognition toolkit. In *Proc. ASRU*.

Povey, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., Wang, Y., and Khudanpur, S. (2016). Purely sequence-trained neural networks for ASR based on lattice-free MMI. In *Proc. Interspeech*.

Pulugundla, B., Baskar, M. K., Kesiraju, S., Egorova, E., Karafiát, M., Burget, L., and Černocký, J. (2018). BUT system for low resource Indian language ASR. *Proc. Interspeech*.

Soltau, H., Liao, H., and Sak, H. (2017). Neural speech recognizer: Acoustic-to-word LSTM model for large vocabulary speech recognition. *Proc. Interspeech*.

Tong, S., Garner, P. N., and Bourlard, H. (2019). An investigation of multilingual ASR using end-to-end LF-MMI. In *Proc. ICASSP*.

Toshniwal, S., Sainath, T. N., Weiss, R. J., Li, B., Moreno, P., Weinstein, E., and Rao, K. (2018). Multilingual speech recognition with a single end-to-end model. In *Proc. ICASSP*.

Trmal, J., Wiesner, M., Peddinti, V., Zhang, X., Ghahremani, P., Wang, Y., Manohar, V., Xu, H., Povey, D., and Khudanpur, S. (2017). The Kaldi OpenKWS system: Improving low resource keyword search. In *Proc. Interspeech*.

Wang, Y., Mohamed, A., Le, D., Liu, C., Xiao, A., Mahadeokar, J., Huang, H., Tjandra, A., Zhang, X., Zhang, F., et al. (2020). Transformer-based acoustic modeling for hybrid speech recognition. In *Proc. ICASSP*.

Watanabe, S., Hori, T., and Hershey, J. R. (2017). Language independent end-to-end architecture for joint language identification and speech recognition. In *Proc. ASRU*.

Wells, J. C. (1995). Computer-coding the IPA: a proposed extension of SAMPA. *Revised draft*, 4(28):1995.

Wiesner, M., Liu, C., Ondel, L., Harman, C., Manohar, V., Trmal, J., Huang, Z., Dehak, N., and Khudanpur, S. (2018). Automatic speech recognition and topic identification for almost-zero-resource languages. In *Proc. Interspeech*.

Young, S. J., Odell, J. J., and Woodland, P. C. (1994). Tree-

based state tying for high accuracy acoustic modelling. In *Proceedings of the workshop on Human Language Technology*, pages 307–312. Association for Computational Linguistics.

Zhang, Y., Chen, G., Yu, D., Yaco, K., Khudanpur, S., and Glass, J. (2016). Highway long short-term memory rnns for distant speech recognition. In *Proc. ICASSP*.