INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Leveraging non-target language resources to improve ASR performance in a target language

Jayadev Billa

Information Sciences Institute,
University of Southern California, Marina del Rey, CA 90292, USA
jbilla@isi.edu

Abstract

This paper investigates approaches to improving automatic speech recognition (ASR) performance in a target language using resources in other languages. In particular, we assume that we have untranscribed speech in a different language and a well trained ASR system in yet another language. Concretely, we structure this as a multi-task problem, where the primary task is acoustic model training in the target language, and the secondary task is also acoustic model training but using a synthetic data set. The synthetic data set consists of pseudo transcripts generated by decoding the untranscribed speech using a well trained ASR model. We compare and contrast this with using labeled data sets, i.e. matched audio and human-generated transcripts, and show that our approach compares favorably. In most cases, we see performance improvements, and in some cases, depending on the selection of languages and nature of speech data, performance exceeds that of systems using labeled data sets as the secondary task. When extended to larger sets of data, we show that the mismatched data approach performs similarly to in-language semi-supervised training (SST) when the secondary task pseudo transcripts are generated by ASR models trained on large diverse data sets.
Index Terms: acoustic modeling, multi-task learning

1. Introduction

Automatic speech recognition systems, in general, leverage large amounts (100s to 1000s of hours) of labeled, i.e. transcribed, speech to build usable systems. The cost of transcription is prohibitive at the scale needed, leading to multi-decade research efforts to directly leverage untranscribed speech data via semi-supervised training, e.g. [1, 2, 3, 4, 5], as well as the use of available labeled speech in other languages, via such approaches as cross-lingual knowledge transfer [6, 7], multi-task [8, 9], and multilingual training [10, 11, 12].

The use of untranscribed speech started with early efforts to harvest training data from speech transcribed with a bootstrap ASR system, followed by the selection of a subset of transcripts based on a threshold applied to the confidence in the truth of the transcript words [13]. This harvested data is then combined with existing labeled data to create a larger training set with which to build better performing models in a semi-supervised training (SST) framework. Later efforts applied this approach to ever larger data collections, e.g. [1], and most recently to neural network based ASR systems e.g. [14, 2, 3].

In parallel, there has been significant work on using multi-task training, i.e. training models on two or more tasks, where the intuition is that by making the model perform on multiple tasks, it is forced to better model the underlying discriminative characteristics of the data than would be the case with a single task/goal; see [9] for low resource language efforts and [8] for a general treatment. This idea of multi-task training is implicit in multilingual training in that instead of training a model for one language, we train on multiple languages even if we are interested in one language [10, 11, 12].

One drawback of semi-supervised training is the assumption that large quantities of untranscribed speech are available in the language of interest. This is mostly true for languages that have larger speaker populations, but with significant variations in availability. In our work within the IARPA MATERIAL [15] program, we have struggled to gather 2000 hours of data for some languages but found it trivial to collect over 8000 hours of data for other languages. As language speaker sizes reduce we expect the availability of data to drop as well. This raises an interesting question that is not necessarily limited to low resource languages: is it possible to use untranscribed data in other languages to improve performance in the language of interest?

To address this question, we construct a multi-task paradigm where one task, the primary task, is training an acoustic model in the target language, and the other, the secondary task, is also training an acoustic model but with a data set created from untranscribed speech. We generate this data set by decoding the untranscribed speech in a second language using a well trained ASR model in a third language. Concretely, by well trained we mean trained on large amounts of diverse data, reflecting diversity in speakers and environmental noise.

The organization of this paper is as follows: Section 2 describes our approach, the training paradigm, and the various data sets and ASR models used. Section 3 describes the experiments themselves and results. We conclude in Section 4 with a summary of the results and how this may fit in the ASR model development process.

2. Experimental Setup

2.1. Approach

As described briefly earlier, we seek to leverage non-target language resources to improve ASR system performance on a target language by creating and training on a synthetic acoustic modeling secondary task in addition to the primary task of acoustic modeling for the target language. Here we assume that we have access to untranscribed data in a language other than the one we are interested in, and we also have access to a well trained ASR system, which may be in yet another language. The synthetic task is created by decoding the untranscribed data with the ASR system to generate transcripts. These transcripts, albeit in another language, are matched to the untranscribed audio, and thus form a consistent set of labeled data with which to train an acoustic model.

If instead of the synthetic task, we use a labeled data set

http://dx.doi.org/10.21437/Interspeech.2021-1657

in another language for training as the secondary task, this approach reverts to the standard multilingual training paradigm. For the primary task, speech recognition in the target language, we assume that we have a labeled data set with human-generated transcripts or otherwise high quality transcripts with which to train the acoustic models.

2.2. ASR training paradigm

We use the Kaldi toolkit [16] in our experimentation, specifically the end-to-end approach [17] to building acoustic models using 40-dimensional MFCC features on speed perturbed [18] data. In terms of the actual training process, we use Kaldi's BABEL multilingual chain2 recipe[1], modified to use the end-to-end modeling paradigm, and updated to use a 15-layer TDNN-F [19] model architecture. There is no experimental reason to adopt the end-to-end approach beyond that it is a more streamlined process for experimentation.

2.3. Data sets and ASR models used

In the work presented here, we use speech data from IARPA's MATERIAL program; this program's goal is to develop methods to find, retrieve, and summarize speech and text content in response to English queries using minimal human created resources. For the ASR component of the program, a small training set, the **build** set, typically consisting of approximately 40-80 hours of telephony speech is provided, as well as a mixed speech (telephone, broadcast news, and conversation) test set, the **analysis** set. Unless otherwise indicated, we use the build data for training and the analysis set as the test set in our experiments.

We choose Pashto as the target language and train using its corresponding 80 hour build set; the Pashto results in this paper are obtained with a 4-gram language model (LM) trained on the build data and around 50M words from web crawl data. For unlabeled data, we choose Bulgarian with its corresponding 40 hour build set. Note the choice of Pashto and Bulgarian is arbitrary; the goal here is to investigate the potential of our approach rather than to obtain the best performance on Pashto. Our current best Pashto system uses about 3000 hours of Pashto audio downloaded from YouTube and trained with incremental semi-supervised training [20]. We will elaborate on how our current approach fits in with the broader goal of obtaining the best performance in our concluding remarks in Section 4.

To explore the impact of the ASR model language, we used existing models in our inventory on Arabic, English, and Tagalog. The Arabic ASR system was trained on a 800 hour subset of GALE Arabic Phase 2-4 data sets (available from LDC) consisting of broadcast news and broadcast conversation in Arabic. The English ASR system was trained on the WSJ corpus (80 hours from LDC93S6B/LDC94S13B). The Tagalog ASR system was trained on the build data, BABEL Tagalog (LDC2016S13), as well as $\tilde{1}$000 hours of YouTube data in a semi-supervised training paradigm. As regards language modeling, the Arabic system uses a pruned 3-gram LM based on the training transcripts, the WSJ system uses a 3-gram LM trained on the training transcripts, and the Tagalog system uses an RNNLM trained on the training data as well as approximately 200M words from web crawl data. The selection of languages, data sets, and ASR systems was primarily dictated by a desire to have as diverse a set of languages/domains as possible to better understand the dynamics of this paradigm.

In terms of experimentation, the primary task is Pashto phoneme based acoustic modeling, and the secondary task is grapheme based acoustic modeling with audio and true transcripts in one of the available ASR models languages (GALE Arabic, Tagalog, or WSJ), or alternatively Bulgarian audio with corresponding transcripts generated with the three ASR models. The rationale behind phoneme and grapheme was that they represented the best performance on each language – the best Pashto system was phoneme-based, the best systems in all the other languages considered were grapheme based. Table 1 summarizes the various combinations of secondary tasks used to improve modeling on the primary task.

3. Results

3.1. Baselines

In order to establish the comparison baselines, we train all models for 5 epochs. The Bulgarian data set was approximately 40 hours; to eliminate the impact of training set size, we limit all the secondary model training to 40 hours selected at random from its corresponding training data set. In all cases, speed perturbation is applied during feature generation, effectively increasing the training data size to 120 hours. For Pashto only baselines, we consider three cases: first, a phoneme based Pashto system trained for 5 epochs; second, a multi-task trained Pashto model, where the primary task remains a phoneme based Pashto acoustic model but the secondary task is a grapheme based Pashto model; finally, a phoneme based Pashto system that is trained for twice as long, 10 epochs vs. 5 epochs. The latter is to provide a fair comparison with the multi-task Pashto system which would in effect have seen twice the amount of data in the 5 epochs, equivalent to 10 epochs of training if we were training on a single task.

The baselines are summarized in Table 2 and represent the best performance we can obtain using these languages as secondary tasks when limited to 40 hours (120 hours after speed perturbation) of training data, 5 epochs of training, and using the true transcripts. Another common approach to improve ASR performance when working on a new language is to use transfer learning, e.g. see [21], where one takes a well trained model in another language, replaces the output layer(s) with new randomly initialized output layers to reflect the new language, and then retrains the entire model with a lower learning rate. Since the approach we present here has a similar goal, it is useful to also compare against a transfer model baseline. The third row in Table 2 uses an Arabic ASR acoustic model, trained as described in Section 2.3, as the source model, followed by 2 epochs of training with the Pashto build data.

On Pashto, as expected, training for more epochs (5→10) improves performance, and if we train in a multi-task framework with phoneme and grapheme targets we see a further small incremental improvement, similar to other reports [9]. We also see the most improvement, 10.8% relative WER improvement with Arabic as the secondary task. Arabic, while not in the same language family as Pashto, does contribute many loanwords to Pashto. With Bulgarian and WSJ as secondary tasks, we see performance improvements though not of the same magnitude; with Tagalog as the secondary task we see a degradation. Part of this is explained by the nature of the data: WSJ corpus is considered a "clean" corpus, free of disfluencies, and is recorded with a high quality microphone; the Bulgarian corpus, while not as clean from a recording perspective, is a mix of

[1] available at `https://github.com/kaldi-asr/kaldi` as `egs/babel_multilang/s5/local/chain2/run_tdnn.sh`

Table 1: *Secondary task descriptions.*

Secondary Task	Audio	Transcript language	Transcript type/method
Arabic	GALE Arabic Broadcast News/Conversation	Arabic	Manual/Human
Bulgarian	Mixed domain Bulgarian	Bulgarian	Manual/Human
Tagalog	Conversational Tagalog	Tagalog	Manual/Human
WSJ	WSJ Read Speech	English	Manual/Human
Bulgarian/Arabic	Mixed domain Bulgarian	Arabic	Auto/Arabic ASR
Bulgarian/Tagalog	Mixed domain Bulgarian	Tagalog	Auto/Tagalog ASR
Bulgarian/WSJ	Mixed domain Bulgarian	English	Auto/English ASR
Somali/Tagalog	Conversational Somali	Tagalog	Auto/Tagalog ASR

Table 2: *Pashto ASR primary task WERs with secondary task as ASR training on labeled data in other languages. Third row is transfer learning using an Arabic source model and does not involve a secondary task.*

Secondary Task	**Epochs Trained**	**Pashto WER**
-	5	56.1
-	10	55.4
Arabic transfer	2	53.4
Pashto (grapheme)	5	55.1
Arabic	5	49.4
Bulgarian	5	54.3
Tagalog	5	55.6
WSJ	5	52.1

broadcast news/conversation and some telephony speech, and benefits from the fact that Bulgarian is highly standardized and broadcast media is typically well articulated. From this perspective, the Bulgarian build data is much cleaner and more diverse than the Tagalog telephony speech in the build set. As a comparison, to illustrate the differences between the Tagalog and Bulgarian data, our baseline Tagalog system trained using only the build data operates at around 75% WER, whereas a similarly trained Bulgarian system, using its corresponding build data, operates at around 55% WER on a similarly constructed test set.

Also of note is the performance of the transfer learning model; it is clearly better than training with just the build data, but multi-task learning with a 40 hour Arabic data set outperforms the transfer model with a 4% absolute reduction in WER compared to the transfer learning model.

3.2. Synthetic secondary task

For the next set of experiments, we investigate how to leverage a well trained ASR model in one language and untranscribed data in yet another language, both of which are different from the language of interest, to improve performance on the language of interest. To imitate this scenario in an experimental framework we use the Bulgarian build set as untranscribed speech and proceed to decode with each of our three available ASR models, resulting in Arabic, Tagalog, and English transcripts for the Bulgarian audio. In our initial attempt, we found that the resulting transcripts for Tagalog and English contained a very high number of instances of the *[UNK]* token, which represents words not in the vocabulary. Clearly, this is not of value when we are trying to train a speech recognition system with a diversity of speech. To mitigate this, we removed the *[UNK]* from the language model, forcing the ASR models to output transcripts without *[UNK]* and yielding a better distribution of words. Following these decodes, we have speech and corresponding transcripts, albeit in a different language, with which we can train an ASR model. We can ignore the fact that the audio is in Bulgarian and train the model in Arabic, Tagalog, and English as the secondary learning task since the transcripts are in the requisite language. Table 3 details the results when we train with Bulgarian audio and ASR generated pseudo transcripts in the three languages.

Table 3: *Pashto ASR primary task WERs with secondary task as ASR training on Bulgarian audio with pseudo transcripts in other languages. All models trained for 5 epochs.*

Secondary Task	**Pashto WER**
Bulgarian/Arabic	53.6
Bulgarian/Tagalog	54.9
Bulgarian/WSJ	55.6

Comparing the results in Table 2 and Table 3, we observe that training on Bulgarian audio with Arabic transcriptions outperforms a system trained on Bulgarian audio/transcripts as a secondary task. This opens up an interesting approach – if we have relatively clean data in a language other than the one we are interested in, we can transcribe it either with a well trained model in a language closely related to the language of interest, or possibly even a bootstrap model in the language of interest. To test this hypothesis, we transcribed the Bulgarian audio with the Pashto bootstrap model with WER 55.7%, as well as an improved Pashto model (CNN-BLSTM-TDNNF model trained with incremental SST over 3000hrs of Pashto audio) with WER 43.4%, on the same test set and using the same language model. These results, in Table 4, suggest that for the synthetic secondary task, a well trained model is an important requirement. If we consider a well trained model to mean a model that has been trained on a variety of data, this would explain the poor Bulgarian/WSJ results in Table 3; the WSJ model is trained on 80 hours of read speech, which is very different from the telephony Pashto data and Bulgarian mixed telephony and broadcast news/conversation data.

Another interesting result from Table 3 is that training on Bulgarian/Tagalog outperforms a system trained on labeled Tagalog data as a secondary task. One hypothesis is that training with a synthetic secondary task has a regularization effect. To extend this further, if we use untranscribed data in a language that is closer to Pashto, then the combined effect of regularization and language similarity should result in better

Table 4: *Pashto ASR primary task WERs comparing impact of ASR model performance (B–bootstrap, I–improved). All models trained for 5 epochs.*

Secondary Task	Pashto WER
Bulgarian/Arabic	53.6
Bulgarian/Pashto (B)	55.3
Bulgarian/Pashto (I)	54.5

performance than the Bulgarian/Tagalog result in Table 3. One such language is Somali, which, while not in the same language family, has similar sounds and loanwords from Arabic (due to geographical proximity). Table 5 compares the Pashto performance with Tagalog, Bulgarian/Tagalog, and Somali/Tagalog secondary tasks. We use a random 40 hour subset of the Somali build set for the latter experiment. As can be seen from the table, our hypothesis is borne out experimentally – we see a respectable improvement in performance in Pashto with Somali/Tagalog secondary task, for an overall 3.4% relative WER improvement on the Pashto baseline.

Table 5: *Phoneme based Pashto ASR primary task WERs with secondary task as ASR training with variants of audio with Tagalog transcripts. All models trained for 5 epochs.*

Secondary Task	Pashto WER
Tagalog	55.6
Bulgarian/Tagalog	54.9
Somali/Tagalog	53.5

3.3. Transferability to larger data sets

While the results in the previous section appear robust, often the question is whether these gains still exist when trained on significantly more data. To address the transferability of these results to larger data sets, we run similar experiments using around 1000 hours of randomly selected 3-way speed perturbed Bulgarian audio segments downloaded from YouTube. Note that since we select utterances after speed perturbation, this 1000 hours of data will span more than a specific 333 hours of original data. The experiments are set up as before, we decode the 1000 hours of Bulgarian speed perturbed audio with Arabic, Bulgarian, Pashto, Tagalog, and English (WSJ) models, and train the Pashto system with the Bulgarian audio and generated transcripts. In the case of Pashto, we simply pool the data with the Pashto training data and train as a single task. To provide a comparison with the more traditional SST approach, we train with 1000 hours of untranscribed speed perturbed Pashto audio (nominally 333 hours) downloaded from YouTube using one-best transcripts generated with the best bootstrap Pashto model trained on the build set (55.4% WER). The training process is identical to the earlier experiments but with the larger sized secondary task and the number of epochs reduced from 5 to 3.

Table 6 details the results from using these larger data sets as the secondary task. First, we see that SST on Pashto is quite effective, reducing absolute WER by 3.7%. Furthermore, we see that across three of the five Bulgarian audio variants, performance is on par with Pashto SST results. The worst performance is seen with the Bulgarian/WSJ model – this is the same behavior we noted while discussing results in Table 3, albeit amplified when using more data. A bit counterintuitive is the Bulgarian/Pashto experiment, where one perhaps expects better performance since the transcripts are in the primary language. The common thread between this and the WSJ result is that the models used to generate the transcripts were trained on a smaller and narrower, i.e. less diverse, set of data, Pashto on conversational speech and WSJ on read news. In comparison, the other models used to generate transcripts were trained on significantly more data (1000-3000 hours) as well as much more diverse data (YouTube & Broadcast News/Conversation).

Table 6: *Phoneme based Pashto ASR primary task WERs with secondary task as ASR training with 1000 hours (after speed perturbation) secondary task audio.*

Secondary Task		Pashto WER
Audio	Transcripts	
-	-	55.4
Pashto	Pashto	51.7
Bulgarian	Arabic	50.8
Bulgarian	Bulgarian	51.9
Bulgarian	Pashto	54.9
Bulgarian	Tagalog	51.9
Bulgarian	English/WSJ	60.5

4. Discussion

In this paper, we have demonstrated, initially with relatively small data sets and then with larger data sets, how using resources in different languages can be used to improve performance in a target language. We show that the performance of such systems is on par with in-language semi-supervised training at least for the languages and combinations considered in this paper. There are several ways in which can leverage this approach: first, as outlined in Section 1, if data in the target language is not available in significant quantities, i.e. languages with a smaller online footprint; second, when building models for a new language, when data is still being collected or crawled, we can use existing data and existing models to improve the initial models even if we intend to use in-language semi-supervised training at a later stage. In both scenarios we envision this approach to be a component of the overall ASR model development process by providing a good starting point for other techniques such as SST. We consistently observe that the better the source model for generating SST transcripts, the better the subsequent trained model.

5. Acknowledgments

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract FA8650-17-C-9116 and by the United States Air Force Research Laboratory (AFRL) under contract FA8650-16-C-6697. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, USAF, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

6. References

[1] J. Z. Ma, S. Matsoukas, O. Kimball, and R. M. Schwartz, "Unsupervised training on large amounts of broadcast news data," in *ICASSP 2006, Toulouse, France, May 14-19*. IEEE, 2006, pp. 1056–1059. [Online]. Available: https://doi.org/10.1109/ICASSP.2006.1660839

[2] K. Veselý, M. Hannemann, and L. Burget, "Semi-supervised training of deep neural networks," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, December 8-12*. IEEE, 2013, pp. 267–272. [Online]. Available: https://doi.org/10.1109/ASRU.2013.6707741

[3] S. Thomas, M. L. Seltzer, K. Church, and H. Hermansky, "Deep neural network features and semi-supervised training for low resource speech recognition," in *ICASSP 2013, Vancouver, BC, Canada, May 26-31*. IEEE, 2013, pp. 6704–6708. [Online]. Available: https://doi.org/10.1109/ICASSP.2013.6638959

[4] V. Manohar, H. Hadian, D. Povey, and S. Khudanpur, "Semi-supervised training of acoustic models using lattice-free MMI," in *ICASSP 2018, Calgary, AB, Canada, April 15-20*. IEEE, 2018, pp. 4844–4848. [Online]. Available: https://doi.org/10.1109/ICASSP.2018.8462331

[5] K. Yu, M. J. F. Gales, L. Wang, and P. C. Woodland, "Unsupervised training and directed manual transcription for LVCSR," *Speech Communication*, vol. 52, no. 7-8, pp. 652–663, 2010. [Online]. Available: https://doi.org/10.1016/j.specom.2010.02.014

[6] P. Swietojanski, A. Ghoshal, and S. Renals, "Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR," in *2012 IEEE Spoken Language Technology Workshop (SLT), Miami, FL, USA, December 2-5*. IEEE, 2012, pp. 246–251. [Online]. Available: https://doi.org/10.1109/SLT.2012.6424230

[7] J. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *ICASSP 2013, Vancouver, BC, Canada, May 26-31*, 2013, pp. 7304–7308. [Online]. Available: https://doi.org/10.1109/ICASSP.2013.6639081

[8] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, Jul. 1997. [Online]. Available: https://doi.org/10.1023/A:1007379606734

[9] D. Chen, B. Mak, C. Leung, and S. Sivadas, "Joint acoustic modeling of triphones and trigraphemes by multi-task learning deep neural networks for low-resource speech recognition," in *ICASSP 2014, Florence, Italy, May 4-9*, 2014, pp. 5592–5596. [Online]. Available: https://doi.org/10.1109/ICASSP.2014.6854673

[10] H. Lin, L. Deng, D. Yu, Y. Gong, A. Acero, and C. Lee, "A study on multilingual acoustic modeling for large vocabulary ASR," in *ICASSP 2009, 19-24 April 2009, Taipei, Taiwan*. IEEE, 2009, pp. 4333–4336. [Online]. Available: https://doi.org/10.1109/ICASSP.2009.4960588

[11] G. Heigold, V. Vanhoucke, A. W. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," in *ICASSP 2013, Vancouver, BC, Canada, May 26-31*, 2013, pp. 8619–8623. [Online]. Available: https://doi.org/10.1109/ICASSP.2013.6639348

[12] N. T. Vu, D. Imseng, D. Povey, P. Motlícek, T. Schultz, and H. Bourlard, "Multilingual deep neural network based acoustic modeling for rapid language adaptation," in *ICASSP 2014, Florence, Italy, May 4-9*, 2014, pp. 7639–7643. [Online]. Available: https://doi.org/10.1109/ICASSP.2014.6855086

[13] G. Zavaliagkos, M. Siu, T. Colthurst, and J. Billa, "Using untranscribed training data to improve performance," in *The 5th International Conference on Spoken Language Processing, Sydney Convention Centre, Sydney, Australia, 30th November - 4th December 1998*. ISCA, 1998. [Online]. Available: http://www.isca-speech.org/archive/icslp_1998/i98_1007.html

[14] J. Z. Ma and S. Matsoukas, "Unsupervised training on a large amount of Arabic broadcast news data," in *ICASSP 2007, Honolulu, Hawaii, USA, April 15-20*. IEEE, 2007, pp. 349–352. [Online]. Available: https://doi.org/10.1109/ICASSP.2007.366244

[15] "Machine Translation for English Retrieval of Information in Any Language (MATERIAL) program," https://www.iarpa.gov/index.php/research-programs/material.

[16] 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, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[17] H. Hadian, H. Sameti, D. Povey, and S. Khudanpur, "End-to-end speech recognition using lattice-free MMI," in *INTERSPEECH 2018, Hyderabad, India, 2-6 September 2018*, B. Yegnanarayana, Ed. ISCA, 2018, pp. 12–16. [Online]. Available: https://doi.org/10.21437/Interspeech.2018-1423

[18] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *INTERSPEECH 2015, Dresden, Germany, September 6-10*, 2015, pp. 3586–3589. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2015/i15_3586.html

[19] 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 *INTERSPEECH 2018, Hyderabad, India, 2-6 September 2018*, B. Yegnanarayana, Ed. ISCA, 2018, pp. 3743–3747. [Online]. Available: https://doi.org/10.21437/Interspeech.2018-1417

[20] B. K. Khonglah, S. R. Madikeri, S. Dey, H. Bourlard, P. Motlícek, and J. Billa, "Incremental semi-supervised learning for multi-genre speech recognition," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*. IEEE, 2020, pp. 7419–7423. [Online]. Available: https://doi.org/10.1109/ICASSP40776.2020.9054309

[21] P. Ghahremani, V. Manohar, H. Hadian, D. Povey, and S. Khudanpur, "Investigation of transfer learning for ASR using LF-MMI trained neural networks," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2017, Okinawa, Japan, December 16-20, 2017*. IEEE, 2017, pp. 279–286. [Online]. Available: https://doi.org/10.1109/ASRU.2017.8268947