

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/328068647>

JAMLIT: A Corpus of Jamaican Standard English for Automatic Speech Recognition of Children's Speech

Conference Paper · August 2018

DOI: 10.21437/SLTU.2018-50

CITATIONS

0

READS

86

2 authors, including:



Andre Coy

The University of the West Indies at Mona

23 PUBLICATIONS 203 CITATIONS

[SEE PROFILE](#)



JAMLIT: A Corpus of Jamaican Standard English for Automatic Speech Recognition of Children's Speech

Stefan Watson¹, André Coy²

^{1,2}University of the West Indies (Mona)

stefan.watson@live.com, andre.coy02@uwimona.edu.jm

Abstract

Children's speech is low resource because few corpora exist. Jamaican English (JE) is even lower resource, as there are no existing children or adult corpora, which hinders the automatic recognition of Jamaican children's speech; data augmentation can overcome this limitation. Typically, augmentation data comes from speakers of the same dialect; however, this is not an option for JE. This work describes JAMLIT, a collection of JE spoken by children; it explores the use of data from related dialects to augment a resource-poor dialect. Augmentation is performed using British (PF-STAR) and American (CMU Kids Speech) English corpora of children's speech. Models created by adding a fraction of the JAMLIT corpus to the PF-STAR corpus improves the recognition of JE, reducing the WER by 58.1% compared to a PF-STAR baseline. With CMU, the improvement was 59.6% over baseline. Both augmented models gave WERs within 2.1% of models trained with Jamaican only data.

Index Terms: Corpus Collection, Jamaican English, Speech Recognition, Children's Speech, Data Augmentation

1. Introduction

The automatic recognition of children's speech has always posed challenges due to factors which include, but are not limited to, data sparsity, and the differences in the acoustic properties of their speech. Automatic speech recognition (ASR) performance on children's speech can be spectacular when adequately trained on large amounts of data (2100 hours in the case of [1]). However, such large datasets are not generally available in the public domain, and readily accessible children corpora such as the CMU Kids Speech Corpus (8.3 hours) [2] and the PF-STAR corpus (14 hours of English) [3] are tiny in comparison, thus marking children's speech as low resource. However, there is currently no corpus of Jamaican children's speech, which further qualifies it as low resource.

Jamaica has one officially recognized language - Jamaican Standard English (JSE), which is the language of instruction and formal communication. However, most Jamaicans are mono-literate bilinguals and speak both JSE and Jamaican Creole (JC), but write and read only JSE. Though there is a call for JC to be recognized as a separate language (see [4 and 5] for example) the relationship between JSE and JC is considered as a standard-creole continuum of language, with JSE being the acrolect and JC the basilect. JSE was initially derived from Standard British English, but it has been influenced by American English due to the proximity of the two countries as well as some level of cultural assimilation. JSE has also been impacted by JC [6]. Thus, spoken JSE, in its current form is not simply a direct derivative of Standard British English, but has evolved into a distinct variant of English with its own phonology [7].

The nature of JSE, as described, underscores the need for a corpus of JSE to be used in developing any application that employs speech and language technology (SLT), as the use of existing corpora would likely lead to less than optimal performance of the various tools, given the phonological differences. This paper describes the collection of such a corpus.

2. Recognition of children's speech

Despite the limitations caused by data sparsity, the best recognition performance on children's speech is generally achieved using models estimated from children's data [1, 8, 9, 10]. To address data sparsity, several data augmentation approaches have been proposed; these approaches can be broadly separated into in-domain data augmentation [11, 12] and out-of-domain augmentation [13, 14, 15, 16]. Data augmentation has been applied to a variety of problems including, cross-language ASR [17, 18], ASR for low resource languages [19, 20], disordered speech [21] and ASR using children's speech [22].

It has been shown that out-of-domain (adult speech) augmentation using unmodified data does not lead to improvement in the recognition of children's speech [13, 14, 16, 22]. However, when these out-of-domain data are perturbed - using techniques such as vocal tract length perturbation (VTLN) [11] and stochastic feature mapping (SFM) [23] - improved performance can be achieved. While out-of-domain augmentation has produced good results for languages with available corpora, for low resource languages it has produced mixed results [20, 24]. Similarly, the results obtained using in-domain data augmentation are mixed. For low-resource languages there is some improvement in word error rates [19, 12]; while, for limited data children's corpora, the word error rates actually increase [22].

As a secondary goal, the paper examines the augmentation of a limited-resource, English language, children's corpora using other dialects of English. For English, the dominant dialects are American and British; children's speech corpora exist for both dominant varieties, but none are available for dialects derived from them. The current study uses a small, unpublished, corpus of Jamaican English and established British and American corpora to answer two questions: firstly, does ASR performance improve when Jamaican data is augmented using speech from dialects related to Jamaican English; and secondly, which, if any, of the related dialects leads to the greater improvement?

The following sections describe the corpus in detail and explore the effect that augmenting training data from other dialects, with limited amounts of JSE data, will have on recognition accuracy. The experiments and recognition results are presented and finally, a discussion and conclusion are presented.

3. Corpus Collection Methodology

Students were recruited from public (government run) primary schools based on the overall performance of the schools in the

Grade Four Literacy Test (GFLT). The GFLT is a standardised test used by the Jamaican Ministry of Education as a proxy for literacy in primary schools. Ten schools were chosen from Kingston and St Andrew in the Southeast of the island and students recruited from grades three, four and five (ages 8 - 11).

The recordings were done in a sound-proof recording studio of a radio station on the campus of the University of the West Indies in Kingston. Students were transported from their schools to the recording studio. Speech was captured using a single, studio grade *Shure Dynamic Mini* microphone, recording equipment and *Adobe Studio 6* software. The microphone and participant were positioned at the same horizontal distance from each other (approximately eight inches) for each recording, while being adjusted for height. Each recording was done with a sampling rate of 48 kHz and 16- bit encoding with the same audio equipment and studio scheduled for the same time each day.

Each recording session lasted approximately 15 minutes, inclusive of a short practice session. The students sat inside the sound room and were instructed to read prompts displayed on a computer monitor. The studio engineer directed the students and changed the prompts. The computer itself was located in the studio room and not in the sound booth so as to minimize external noise. While a recording session was in progress, those children not recording were engaged with fun activities. However, after 90 minutes they were obviously ready to move on to something else. It was decided that no more than six students would be taken to the studios at any given time.

3.1. Challenges

The first challenge was how to design a prompting system for the recording session. The system designed had to accomplish the task of ease of use and interaction. Therefore, a two-button interface was designed within a full screen browser interface as the prompting system for the reader. The user was familiarised with the system by allowing them to read numbers from 1 to 10, then colours and finally simple sentences. All the sentences and words were in large, bold text in a legible font so that the students could easily see and read from the screen. The text was also displayed in the center of the screen so that users never had to move the position of their head while reading. The two buttons on the interface allowed the user to move forward to another sentence and back to a previous sentence in order to facilitate the desire to skip sentences or to go back and make corrections. In initial tests prior to actual recording, it was found that some students became distracted by having to navigate through the prompts, while others had little experience using the computer and so were not confident enough to use the system efficiently. It was thus decided that the studio engineer should manipulate the prompt system.

The reader could read a sentence as many times as was required in order to ensure correct pronunciations of all the words in the sentence. The participants were only required to read the sentences at normal reading levels.

4. Corpus

JAMLIT, a phonetically balanced set of 42 simple sentences was developed, with words taken from the word lists contained in the curriculum document for each grade. The final sentence list includes sentences such as:

The thief jumped through the window and broke his arm to avoid being caught.

My teacher always tells us, "Stop the noise and sit while in class."

"What I enjoy most about fish, is that it tastes awesome."

The number of recordings stand at 219 with 120 boys and 99 girls with provisions for more recordings to be done. Currently the data stands at 15.6 GB with over 9400 recordings totaling just less than 55 hours. The data are split into training (80%) and test (20%).

The sentences were transcribed by a single transcriber using the Transcriber software tool. The BEEP pronunciation dictionary was employed for phonetic transcription of the data.

4.1. Age-based Differences

Differences in age between the participants was expected to affect their reading performances. The grade range chosen (3-5) was selected solely on the premise of the Grade Four Literacy Test which is the standard used to judge the literacy of children in Jamaica. The effect that the age gap between the participants had on their currently literacy levels was apparent. The younger participants tended to struggle with words that were 1) unknown to them or 2) known but mispronounced. Whether or not these instances were due to the impact of the JC spoken informally is unclear. It is to be noted that the sentences recorded were a combination of words from the syllabus of all three grades. Therefore, the participants at a lower grade would not encounter all the words that those at higher grade levels are expected to know.

4.2. Gender-based Differences

The most surprising observation was the disparity between the genders and reading levels. Female participants generally outperformed the males at the same age with a few exceptions where the males were on par with the females. In some cases, female participants, who were at a lower grade than their male counterparts, performed on par and even outperformed them. This is significant as the participants selected were considered the best readers among their peers. Most of the schools chosen were co-education institutions, except for one all-female school. The students were chosen from the same classrooms which would mean that each student should be exposed to the same teaching material, methods and curriculum. However, the disparity between the genders was apparent in the number of errors made, even when the difference in average reading time is taken into account.

5. Experimental Framework

5.1 Acoustic Data

Three children corpora, the CMU Kids Speech corpus [2], the PF-STAR corpus [3] and the JAMLIT corpus, introduced in this paper, were used to represent American, British and Jamaican English. The PF-STAR corpus comprises 158 speakers ranging from 4-15 years of age, with approximately 14 hours of data sampled at 22.05 kHz. The corpus is divided into 7.4 hours of training data from 86 speakers, 5.8 hours of test data (60 speakers) and 54 minutes of evaluation data from 12 speakers. The CMU corpus contains 5180 utterances, totaling 8.3 hours, produced by 24 males and 52 females with ages ranging from 6-11, sampled at 16 kHz.

5.2 Language Models

Owing to the limited vocabulary of the three speech corpora being used, a language model (LM) was developed from other

sources. As the vocabularies of the corpora are very different from each other and from any other corpus, it is difficult to determine the utility of a particular language model for this task, thus three different language models are used and the recognition performance compared. Three LMs, the TEDLIUM, WSJ and BBC LMs, were selected for use in this study.

The TEDLIUM model has a 157.6K word vocabulary, with pronunciations derived from the CMU dictionary and the Festival Speech Synthesis System. It is an interpolated, 4-gram back-off model with Kneser-Ney discounting [25].

The second LM is the WSJ LM is a 20K word tri-gram LM constructed from the words in the WSJ corpus [26]. A pronunciation dictionary of 33K words was included in the database. The third LM used was the BBC LM, which was developed for the MGB challenge [27] and contains 640M words taken from subtitles recorded on the BBC over 34 years. The lexicon was taken from the 2015 MGB Challenge, which has a vocabulary of 238.6K words. For this sets of experiments the model used was a Kneser-Ney smoothed 3-gram LM using the 150K most common words, as described in [28].

5.3. Acoustic Modelling

All the models in this experiment were developed using the Kaldi Toolkit [29]. As there are no Kaldi recipes for any of the corpora employed, existing recipes are adapted to generate and test Gaussian-based (GMM-HMM) and Deep Neural Network (DNN) acoustic models; each data set was prepared using similar steps to those in the TEDLIUM and WSJ recipes.

The data were sampled at 16 kHz, and initial GMM-HMM models were trained from each dataset. From the data, 39-dimension Mel-frequency cepstral coefficients were extracted. Monophone and triphone models were created, after which, a number of adaptation techniques (linear discriminative analysis (LDA) with maximum likelihood linear transformation, speaker adaptive training (SAT) through feature-space maximum likelihood linear regression (fMLLR)) were applied to the triphone models. Deep Neural Network (DNN) modelling with sequential Training for the DNN and Minimum Phone Error (MPE) was performed using the *nnet1* script in Kaldi. Decoding is done after each stage, to view the improvements in the Word Error Rate (WER) gained by using each technique.

5.3 Experiments

The first set of experiments tested the performance of matched training and test data. The results of these experiments served as a baseline for determining the improvement that is achievable using the proposed data augmentation method. The effect of different language models was explored in a second set of experiments. This was meant to highlight the importance of determining the best language model for a low-resource dialect.

Finally, models were created using CMU and PF-STAR data and augmented with different amounts of unmodified data from the JAMLIT training set. Sets of 1000 utterances (1 hr. 45 mins), 1500 utterances (2 hrs. 30 mins) and 2000 utterances (3 hrs. 20 mins) were randomly chosen and added to both datasets and combined models are created. These models were used to recognise speech from the JAMLIT test set.

6. Results

Figure 1 shows the results of matched recognition, obtained for all stages of the model building process, using each of the language models. As expected the DNN models produced the best result for all combinations of acoustic and language model, with WERs of 27.3%, 49.3% and 22.9% for PF-STAR, CMU and

JAMLIT, respectively. The steepest reductions in word error rate (WER) occurred between the mono-phone and tri-phone stages, with more gradual reductions in WER for the other models. This pattern holds across the three corpora and for most acoustic and language model combinations (there was an increase of 0.1% and 2.3%, respectively, in WER between the SAT and DNN models for the CMU dataset using the WSJ and TEDLIUM language models).

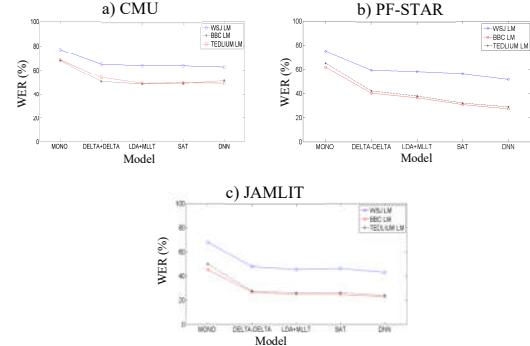


Figure 1: Recognition performance for all three datasets using matched acoustic models and language models.

6.1. The Effect of Different Language Models

When using the WSJ LM, the WERs were between 9 and 29% higher than those for the TEDLIUM and BBC LMs. This is consistent with the fact that the WSJ LM is the weakest of the three LMs employed. WERs achieved on matched recognition using the BBC LM are on average 1.6% lower than those achieved using the TEDLIUM models.

6.2 Baseline Recognition Results Using Matched Data

The result of decoding the PF-STAR test set with the DNN acoustic model and the TEDLIUM LM was 29.1%; this compares favourably with the WER of 29% achieved in [22] for a similar experiment using DNNs with restricted Boltzmann machine pretraining and the BBC LM. Further improvement was seen when the BBC LM was used; the WER of 27.3% is similar to the best result of 27.2% achieved in [22] when PF-STAR was augmented WSJCAM0 data perturbed by SFM.

Recognition of the JAMLIT corpus yields error rates of 43.1%, 24% and 22.9%, for the WSJ, TEDLIUM and BBC LMs, respectively. The results for the CMU models are the worst of the three. The WER of 49% was achieved with the BBC LM. This is not necessarily indicative of the quality of the dataset, as it is not meant to be a standalone corpus but rather as data for adaptation of existing speech models [30].

Table 1. WER (%) for DNN training of CMU, PF-STAR and JAMLIT models on PF-STAR, CMU and JAMLIT test data. The BBC LM is used in all tests.

DATA	MODEL		
	CMU	PF-STAR	JAMLIT
CMU	49.3	75.9	85.5
PF-STAR	74.0	27.3	83.0
JAMLIT	84.1	85.7	22.9

6.3 Recognition with Augmented Data

The previous section shows that the best recognition performance was achieved with the BBC LM, thus the rest of the experiments were performed using that model. Table 1 shows the

result of decoding the three English datasets with each acoustic model and the BBC LM.

These results illustrate the need for data from the target dialect, as models that perform well on matched data do not perform as well on cross-dialect data. The table also shows that PF-STAR and CMU models perform equally poorly on the JAMLIT data, while the JAMLIT models perform slightly better when tested with the PF-STAR data.

The results from the augmented CMU and PF-STAR models (see Table 2) show dramatic improvements on the baseline models from the initial monophone stage for both corpora. For CMU models tested on JAMLIT data, the WERs fell from 84.1% to 27.3%, 27.4% and 25.6% with the addition of increasing amounts of JAMLIT data. Adding JAMLIT data to the PF-STAR data yielded WERs of 27.2%, 26.2% and 24.9%, this compares to a baseline of 85.7%.

The augmented models give results which approach the standalone JAMLIT models (within 2% for PF-STAR) albeit with less than 50% of the total JAMLIT training data. For comparison, models were created using the 2000 utterances (3 hrs. 20 min) of the JAMLIT adaptation data. The WER achieved was 24.7%, almost identical to the augmented PF-STAR models, and 1% different from the augmented CMU models.

Table 2. *WER (%) for DNN training of CMU and PF-STAR models augmented with varied amounts of JAMLIT data. The BBC LM is used in all tests.*

# of JAMLIT Utterances	MODEL	
	CMU	PF-STAR
1000	27.3	27.2
1500	27.4	26.2
2000	25.6	24.9

7. Discussion

This paper introduced the JAMLIT corpus and further, sought to provide insight into the possibilities of acoustic modeling for dialects that are more resource poor than related dialects. The focus was on children's speech, which is traditionally resource poor. Jamaican English (JE) was taken as the target language, because of its relationship to American English (AE) and British English (BE), and because of the fact that no corpus exists for Jamaican children's speech, while there are established corpora for the other dominant dialects. The aim of the paper was to answer two questions: firstly, does ASR performance improve when Jamaican data is used to augmented speech from dialects related to JE; and secondly, which, if any, of the related dialects leads to the greater improvement?

The study shows that models created using children's data from one dialect of English do not ideally recognise speech from a related dialect. This has implications for applications, such as intelligent tutors, which will likely be accessed online, or from media sent from other countries. These applications will likely be used 'straight out of the box' and would adopt models from related dialects without modification.

The answer to the first question marks a key contribution made by this work. It has been shown that small amounts of unmodified data from the target dialect can be used to augment data from a related dialect to produce a model that performs almost as well as a full model from the target dialect. It also shows that although children's speech from dialects of English are best recognized with training and test data from the same demographic, data augmentation using cross-dialect data from children can lead to comparable results.

The second question is answered by observing that the BE models outperform the AE models, whether using raw data, or data augmented with Jamaican English. This is not entirely surprising, given the historical origins of English in Jamaica. What is somewhat surprising is the significant improvement in WER when the AE models were augmented with small amounts of JE, the WER plummeted by 60%. This decrease was significant enough for the WER to be comparable with the WER achieved with models developed using pure Jamaican data. This is even more surprising, given that the resulting WER achieved by the augmented models on Jamaican data (25.6%) was lower than that achieved with the unmodified models decoding in-domain (AE) data (49.3%). This is not a case of additional data improving performance, as JE models created with little over three hours data are compared to those created with almost three times as much; the difference in the WERs is only 1.8%. It is clear, that simply adding more in-domain target data to target models does not yield significant improvements (as shown when SFM was used to augment PF-STAR [22]). However, it has been shown that adding in-domain target data to in-domain data from related dialects can lead to dramatic improvement in the recognition performance of the models created from these dialects.

The most important implication of this work is that acoustic models, for children, can be developed for limited-resource dialects of English using data from related dialects that are more data rich, possibly obviating the collection of extensive corpora for the limited-resource dialect.

8. Conclusions

This study shows that models created using unmodified children's data from one dialect of English do not ideally recognise speech from a related dialect. However, adding relatively small amounts of data from the target dialect can, in some cases, improve performance dramatically. This study illustrated this principle for the case of Jamaican children's speech and its related dialects - American and British English. Though this study explored this approach for English, there is no indication that similar results cannot be achieved with other languages.

9. Acknowledgements

This work has been partially supported by a Research and Publications grant from the University of the West Indies (Mona), an LDC Fall 2015 data scholarship and the CloudCAST International Network (IN-2014-003) funded by the Leverhulme Trust

10. References

- [1] H. Liao, G. Pundak, O. Siohan, M. K. Carroll, N. Coccaro, Q.-M. Jiang, T. N. Sainath, A. Senior, F. Beaufays, and M. Bacchiani, "Large vocabulary automatic speech recognition for children," In Interspeech, 2015, pp. 1611–1615, 2015.
- [2] M. Eskenazi, J. Mostow, and D. Graff, "The CMU Kids Corpus," Linguistic Data Consortium, 1997.
- [3] A. Batliner, M. Blomberg, S. D'Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, S. Steidl, and M. Wong, "The PF Star children's speech corpus," in Ninth European Conference on Speech Communication and Technology, 2005, pp. 2761–2764, 2005.
- [4] H. Devonish and K. Carpenter, Towards Full Bilingual Education: The Jamaican Bilingual Primary Education Project, Social and Economic Studies 56, 1-2, pp. 277-303, 2007.
- [5] F. Marijke. "Introducing Jamaican Creole into the Jamaican Educational Curriculum." The English Languages: History, Diaspora, Culture vol. 1, no. 1 (2010)

- [6] S. Jantos. Agreement Variation in educated Jamaican English: A Corpus Investigation of ICE-Jamaica. Ph.D. dissertation. University of Freiburg, Germany, 2009.
- [7] H. Devonish and O. Harry. “Jamaican Creole and Jamaican English: phonology.” In: Schneider et al. (eds.): *A Handbook of Varieties of English*. Vol. 1. Phonology. Berlin: Mouton de Gruyter, 2003.
- [8] Q. Li and M. J. Russell, “Why is automatic recognition of children’s speech difficult?” in *INTERSPEECH*, 2001.
- [9] M. Gerosa, D. Giuliani, and F. Brugnara, “Acoustic variability and automatic recognition of childrens speech,” *Speech Communication*, vol. 49, no. 10, pp. 847–860, 2007.
- [10] R. Serizel and D. Giuliani, “Deep-neural network approaches for speech recognition with heterogeneous groups of speakers including children,” *Natural Language Engineering*, vol. 23, no. 3, pp. 325–350, 2017.
- [11] N. Jaitly and G. E. Hinton, “Vocal tract length perturbation (VTLP) improves speech recognition,” in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, 2013, pp. 625–660, 2013.
- [12] X. Cui, V. Goel, and B. Kingsbury, “Data augmentation for deep neural network acoustic modeling,” *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 9, pp. 1469–1477, 2015.
- [13] J. G. Wilpon and C. N. Jacobsen, “A study of speech recognition for children and the elderly,” in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. In IEEE International Conference on*, vol. 1. IEEE, 1996, pp. 349–352, 1996.
- [14] A. Potamianos and S. Narayanan, “Robust recognition of children’s speech,” *IEEE Transactions on speech and audio processing*, vol. 11, no. 6, pp. 603–616, 2003.
- [15] D. Elenius and M. Blomberg, “Adaptation and normalization experiments in speech recognition for 4 to 8-year-old children.” In *Interspeech 2005*, pp. 2749–2752, 2005.
- [16] R. Serizel and D. Giuliani, “Deep neural network adaptation for children’s and adults’ speech recognition,” in *Italian Computational Linguistics Conference (CLiC-IT)*, 2014.
- [17] P. Swietojanski, A. Ghoshal, and S. Renals, “Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR,” in *Spoken Language Technology Workshop (SLT)*, 2012, IEEE, 2012, pp. 246–251.
- [18] A. Stolcke, F. Grezl, M.-Y. Hwang, X. Lei, N. Morgan, and D. Vergyri, “Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons,” in *ICASSP*, vol. 1, 2006, pp. 321–324, 2006.
- [19] Z. Tuske, P. Golik, D. Nolden, R. Schluter, and H. Ney, “Data augmentation, feature combination, and multilingual neural networks to improve ASR and KWS performance for low-resource languages,” in *Interspeech*, 2014, pp. 1420–1424, 2014.
- [20] A. Ragni, K. M. Knill, S. P. Rath, and M. J. Gales, “Data augmentation for low resource languages,” in *Interspeech*, 2014, pp. 810–814, 2014.
- [21] H. Christensen, M. Aniol, P. Bell, P. D. Green, T. Hain, S. King, and P. Swietojanski, “Combining in-domain and out-of-domain speech data for automatic recognition of disordered speech.” In *Interspeech*, 2013, pp. 3642–3645, 2013.
- [22] J. Fainberg, P. Bell, M. Lincoln, and S. Renals, “Improving children’s speech recognition through out-of-domain data augmentation.” in *Interspeech*, 2016, pp. 2749–2752.
- [23] H. Zen, Y. Nankaku, and K. Tokuda, “Continuous stochastic feature mapping based on trajectory HMMs,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 417–430, 2011.
- [24] K. M. Knill, M. J. Gales, S. P. Rath, P. C. Woodland, C. Zhang, and S.-X. Zhang, “Investigation of multilingual deep neural networks for spoken term detection,” in *Automatic Speech Recognition and Understanding (ASRU)*, 2013 IEEE Workshop on, 2013, pp. 138–143, 2013.
- [25] A. Rousseau, P. Deleglise, and Y. Esteve, “Ted-lium: an Automatic Speech Recognition Dedicated Corpus.” in *LREC*, 2012, pp. 125–129, 2012.
- [26] D. B. Paul and J. M. Baker, “The design for the Wall Street Journal-based CSR Corpus,” in *Proceedings of the workshop on Speech and Natural Language*, 1992, pp. 357–362, 1992.
- [27] P. Bell, M. J. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Wester et al., “The MGB challenge: Evaluating multi-genre broadcast media recognition,” In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, 2015, pp. 687–693.
- [28] S. R. Gangireddy, P. Swietojanski, P. Bell, and S. Renals, “Unsupervised adaptation of recurrent neural network language models.” in *Interspeech*, 2016, pp. 2333–2337, 2016.
- [29] 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*, 2011.
- [30] G. Aist, P. Chan, X. Huang, L. Jiang, R. Kennedy, D. Latimer, J. Mostow, and C. Yeung, “How effective is unsupervised data collection for children’s speech recognition?” in *Fifth International Conference on Spoken Language Processing*, 1998.