
Proceedings Paper:
Ragni, A. and Gales, M. (2018) Automatic speech recognition system development in the "wild". In: Interspeech 2018. Interspeech 2018, 02-06 Sep 2018, Hyderabad, India. International Speech Communication Association (ISCA), pp. 2217-2221. ISSN: 1990-9772.

https://doi.org/10.21437/interspeech.2018-1085

Interspeech 2018
2-6 September 2018, Hyderabad

Automatic Speech Recognition System Development in the "Wild"

Anton Ragni and Mark J. F. Gales

Department of Engineering, University of Cambridge
Trumpington Street, Cambridge CB2 1PZ, UK
{ar527,mjfg}@eng.cam.ac.uk

Abstract

The standard framework for developing an automatic speech recognition (ASR) system is to generate training and development data for building the system, and evaluation data for the final performance analysis. All the data is assumed to come from the domain of interest. Though this framework is matched to some tasks, it is more challenging for systems that are required to operate over broad domains, or where the ability to collect the required data is limited. This paper discusses ASR work performed under the IARPA MATERIAL program, which is aimed at cross-language information retrieval, and examines this challenging scenario. In terms of available data, only limited narrow-band conversational telephone speech data was provided. However, the system is required to operate over a range of domains, including broadcast data. As no data is available for the broadcast domain, this paper proposes an approach for system development based on scraping "related" data from the web, and using ASR system confidence scores as the primary metric for developing the acoustic and language model components. As an initial evaluation of the approach, the Swahili development language is used, with the final system performance assessed on the IARPA MATERIAL Analysis Pack 1 data.

Index Terms: cross-domain development, confidence, web data, speech recognition

1. Introduction

Speech data present in the "wild" comes in various forms such as Youtube, podcasts and radio news. This provides a contrast to conversational telephone speech (CTS), broadcast news and voice search style data for which numerous systems have been developed [1, 2, 3, 4, 5, 6]. Handling these new and emerging types of data using automatic speech recognition (ASR) systems built on out-of-domain data is a challenging problem for any language. This problem stands at the core of the new IARPA initiative for machine translation for English retrieval of information in any language (MATERIAL). Participating institutions in this program are expected to develop ASR technology that would enable accurate machine translation, cross-language information retrieval and summarisation when no target domain data is given nor for training or testing. In the first year of the program the participants were given narrow-band CTS data from IARPA's previous initiative BABEL. To facilitate the development of approaches capable of handling potentially large domain mismatches, an analysis pack was released containing small quantities of transcribed wide-band news (BN) and topical (BT) broadcasts. It it important to note that such data will not be immediately available in the later stages of the program. Furthermore, handling other domains may be required during the end of period evaluation.

Lack of target domain data creates a challenge in developing statistical models. Even within a given domain over-tuning on training data may result in poor performance. Larger mismatches are expected to lead to a larger degradation. This in turn may have a negative impact on lightly and semi-supervised approaches [7] useful in such limited data scenarios. To ensure that decisions made in training will not impact generalisation on evaluation data it is common to use supervised metrics, such as word error rates (WER), frame classification accuracies, proxies to minimum Bayes' risk and conditional maximum likelihoods, on held-out development data. If no such supervised data is available the problem becomes more complicated. Essentially, an unsupervised criterion correlated with the target domain performance is required, which is the subject of this paper.

The previous work [8] in this area has looked at WER between acoustic model predictions with a weak and strong [9, 10] language model. The premise was that high agreement between the two predictions suggests the acoustic model is well matched to the target domain. This technique was employed to help track the progress of ASR system development and monitor possible drifts in the deployment domain. Though high correlation with the ASR performance on the target domain data was reported, large domain mismatches were not explored. Another work [11] looked at the harder problem of predicting WER by computing an expected loss between hypothesised and target domain text. The hypothesised text was generated by means of a phone-to-phone confusion model estimated on a mismatched domain training data. Reasonably close WER predictions were demonstrated on digit recognition and Wall Street Journal tasks. However, these results were reported only for the final model.

This paper looks at the development of ASR systems under highly mismatched training and evaluation conditions. Furthermore, the precise nature of the evaluation domains is assumed unknown at the time of development. The approach proposed in this paper consists of automatically acquiring a sample of related domain text and audio data from the web and using an unsupervised metric to track the development and tuning. The metric is an average of related domain confidence scores mapped to out-of-domain development data. It is shown to be capable of tracking the target domain performance. This paper also proposes an unsupervised method for tuning interpolated language models to the target domain by optimising the metric on the related domain data.

The rest of this paper is organised as follows. Section 2 describes the development in the "wild". The following Section 3 discusses confidence scores. Section 4 details the unsupervised

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via Air Force Research Laboratory (AFRL) contract # FA8650-17-C-9117. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, AFRL or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

10.21437/Interspeech.2018-1085

approach to language model interpolation weights estimation. Experimental results are presented in Section 5. Conclusions drawn from this work are given in Section 6.

2. Development in the "wild"

There are numerous audio domains present in the "wild". Often these are not covered by a typical laboratory or commercial system even for languages with strong presence in the speech technology sector. The situation with lower resource languages as expected is even worse. Table 1 gives an overview of some of Swahili audio data available on the web at the time of writing this paper. Swahili is a language of Niger-Congo family spoken as a *lingua franca* in much of East Africa. It is estimated to have 2 million L1 (native) and 50-100 million L2 speakers. According to Table 1 there are 82 hours of conversational telephone speech (CTS) and 14 hours of read transcribed narrow-band data available. This data was collected for IARPA's previous BABEL initiative. The remaining is an untranscribed wide-band data that comes from Youtube and news stations such as Voice of America, Swahili (VOA). All these domains are expected to be distinct from the transcribed CTS data. Furthermore, Youtube data is highly dynamic with a wide range of constantly changing domains. The standard process for building acoustic models would involve transcribing large quantities of data for each domain. However, as these domains change over time an interesting question is: is it possible to generalise from an out-of-domain data given only a sample of untranscribed data in a related domain? For example, is it possible to calibrate a neural network acoustic model on the related domain data so that it correlates with the performance on the target domain. Provided target and related domains are sufficiently close and an unsupervised metric highly correlated with the performance is available, the answer to this question will be positive.

Table 1: *Swahili web audio data summary*

Hrs	Source					
	Babel		Youtube		News	
	CTS	Read	BBC	Unk	VOA	Other
Audio	82	14	149	1445	511	104
Sub/Tra	82	14	–	470	–	55

In addition to handling the mismatch in the acoustics any such approach would also need to handle changes in the language. Table 2 shows sources and quantities of text data scraped from the web by BABEL participants [12, 13]. Excluding the Babel program data, the participants were able to scrape blogs, transcripts of TED talks as well as general web text using Bing search engine with language codes `swa` (BingA) and `swh` (BingH). Similar to audio, this data comes from a range of domains. Often it is hard to decide which source text will be useful for a given domain. Therefore, automatic procedures, such as language model interpolation, are commonly used to learn optimal language model combination weights on some target domain development text. This is hard to ensure if the target domain identity is unknown. A similar question which arises is: can these weights be determined in an unsupervised fashion that would correlate with the target domain performance?

Table 2: *Swahili web text data summary*

Words	Sources				
	Babel	TED	Blogs	BingA	BingH
Total	294k	44k	1185k	6040k	2980k
Unique	24k	9k	47k	245k	184k

3. Confidence scores

One standard approach to judge reliability of predictions in ASR are confidence scores [14]. These scores are traditionally derived from a lattice following the recognition run. In the simplest case, confidence scores are lattice arc posterior probabilities. More complex schemes include arc posterior based confusion networks [15, 16] and general feature based neural network models [17]. All these approaches enable confidence estimate to be given to each hypothesised word.

It has been observed however that confusion networks yield over-estimated confidence scores [16]. This is believed to be a consequence of not encoding all possible paths into a lattice. As a result, the lattice weight (normalisation term) applied to yield arc posterior probabilities is underestimated hence yielding higher than expected estimates. To yield more reliable estimates confidence scores can be transformed to better represent the confidence in prediction. This can be accomplished by training a mapping that maximises confidence of correctly predicted words and minimises confidence of incorrectly predicted words [16]. Note that deletion errors are not covered and have to be treated separately [18]. The typical approach uses a decision tree to learn a piecewise linear mapping.

Several studies have found that simple statistics of confidence scores, such as average, can be used to assess complexity of a task [16], compare similarly performing ASR systems [17]. This paper proposes to use these statistics to predict cross-domain generalisation. Given an initial acoustic model trained on narrow-band data, a sample of related to target domain wide-band data from the web is recognised to yield the average mapped confidence score. Following an update of the initial acoustic model, such as changes to model topology, training approach, data normalisation and speaker adaptation, the confidence score is recomputed. If the new confidence score is significantly higher the update is accepted. If not then either an adjustment that would lead to a significant increase in the confidence score is found or the update is rejected. The acoustic model performance on the narrow-band data is monitored throughout the process and changes leading to large degradations are rejected. Though simple this process requires specifying what is the significant change in the confidence score that would lead to generalisation. In addition, a special attention needs to be paid to deletion errors. Rising number of deletion errors may signal that the confidence score is over-estimated and hence may enable updates that fail to generalise. Both of these issues would need to be investigated.

4. Language model interpolation

One standard approach to incorporate new text data into an existing n-gram language model is language model interpolation

$$p(w_i|w_{i-1}, w_{i-2}, \ldots) = \sum_{j=1}^{N} \lambda_j p_j(w_i|w_{i-1}, w_{i-2}, \ldots) \quad (1)$$

where N is the number of language models, $\boldsymbol{\lambda} = \{\lambda_j\}$ are positive interpolation weights that add up to one, $p_j(w_i|w_{i-1}, \ldots)$ is a probability of word w_i given past word history $w_{i-1}, w_{i-2}, \ldots$ computed by the j^{th} language model.

Given some target domain text, these weight can be set using an expectation-maximisation procedure [19]. When the target domain text is not available, some target domain audio can be recognised to yield hypothesised text [20]. This is a simple unsupervised process that can be applied once the target or related domain audio is available. The disadvantage is that weights are estimated on transcriptions that contain errors.

Alternatively, weights can be estimated by directly maximising the average mapped confidence score. Given the complexity of the underlying function, a derivative free optimisation, such as Powell's method [21], can be performed. A special care is needed to ensure that weights form a valid distribution. On each iteration weights are updated by

$$\hat{\boldsymbol{\lambda}} = \boldsymbol{\lambda} + \hat{s}\mathbf{d} \tag{2}$$

where $\mathbf{d}$ is a direction vector and $\hat{s}$ is an optimal step size. For $\hat{\boldsymbol{\lambda}}$ to be a valid weight distribution both $\hat{s}$ and $\mathbf{d}$ must obey certain constraints. One suitable choice of the direction vector that preserves the positivity and sum-to-one constraint subject to certain restrictions on the optimal step size $\hat{s}$ is given by

$$d_i = \begin{cases} 1, & \text{if} \quad i = n \\ -\frac{1}{N-1}, & \text{otherwise} \end{cases} \tag{3}$$

where $n \in [1, N]$ is the current iteration number. Note that a different direction vector is used on each iteration as in Powell's method. These directions however are not mutually conjugate unlike in Powell's method which may have an impact on convergence. For sufficiently small positive $\hat{s}$ such direction vector will increase the n^{th} weight and decrease the remaining weights equally and uniformly thus preserving the positivity and sum-to-one constraint. Given the direction vector in equation (3), for $\hat{\boldsymbol{\lambda}}$ to be a valid weight distribution

$$\lambda_i + sd_i \geq 0 \tag{4}$$

$$\lambda_i + sd_i \leq 1 \tag{5}$$

must hold for all i. These constraints can be re-expressed as

$$\max_i(a_i) \leq s \leq \min_i(b_i) \tag{6}$$

where

$$a_i = \begin{cases} -\frac{\lambda_i}{d_i}, & d > 0 \\ \frac{1-\lambda_i}{d_i}, & d < 0 \end{cases}, \quad b_i = \begin{cases} \frac{1-\lambda_i}{d_i}, & d > 0 \\ -\frac{\lambda_i}{d_i}, & d < 0 \end{cases} \tag{7}$$

Given the bounds in equation (6), the optimal step size $\hat{s}$ can be found through a line minimisation

$$\hat{s} = \arg\max_s \{\mathcal{F}(\boldsymbol{\lambda} + s\mathbf{d}; \mathcal{D})\} \tag{8}$$

where $\mathcal{F}(\boldsymbol{\lambda})$ is the average mapped confidence score of an ASR system using an interpolated language model with weights $\boldsymbol{\lambda}$ to recognise domain data $\mathcal{D}$. The complete process is repeated on the next $n + 1^{st}$ iteration using $\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}$. Upon reaching the final iteration N the optimisation process can be repeated.

Unlike the unsupervised adaptation approach described at the beginning of this section, maximising the average mapped confidence score is expected to be less sensitive to recogniser errors. In addition, the weights can be estimated before any target domain data is available by using some related domain data scraped from the web. The disadvantage of this approach is a high computational cost of evaluating $\mathcal{F}(\boldsymbol{\lambda})$. This can be reduced to some extent by doing lattice rescoring, as in this work, rather than a full recognition run, bracketing the maximum tightly and using other optimisation methods [22].

5. Experiments

Experiments were conducted using IARPA BABEL Swahili release pack B data for training and development, IARPA MATERIAL analysis pack 1 data for evaluation. In addition, web text and audio data described in Section 2 were used for language model training and confidence score estimation. The development and evaluation audio data is summarised in Table 3. The

Table 3: *Test audio data summary*

Id	Set	Band	Type	Dur (hrs)
BNB	Babel	narrow	CTS	10.2
MNB	Material	narrow	CTS	1.8
MWB		wide	BN,BT	5.6
VWB	VOA	wide	BM	9.7

Material data contains a small sample of narrow-band, CTS, data possibly recorded under different to Babel channel conditions. The remaining Material data comes from news (BN) and topical (BT) broadcasts. A sample of untranscribed VOA data containing mixed broadcast (BM) web audio was selected to approximate the wide-band Material data.

The acoustic model was trained in 2 stages. The first stage utilised only "clean" portion of Babel training CTS data that excludes utterances with partial and mispronounced words to bootstrap a maximum likelihood Gaussian mixture model (GMM) ASR system. The GMM ASR system was used to process the remaining training data to yield "clean" transcripts. Overall this increased the amount of training data from 43 to 57 hours of speech. The second stage utilised the expanded training data to build first a speaker adaptive discriminatively trained GMM and then an interleaved Time-Delay Neural Network and Long Short-Term Memory (TDNN-LSTM) model trained using Lattice-Free Maximum Mutual Information (LF-MMI) criterion [23]. Both stages whenever possible avoided custom configurations to avoid over-tuning to the CTS data. The stage one system used an HTK [24] configuration that had been previously employed for all Babel tasks [25, 26], multi-genre English broadcast transcription [27] and many others. The stage two system used a Kaldi [28] configuration that had been previously employed for multi-genre English broadcast [27] and spoken English language assessment. The only modification to the Kaldi configuration done in this paper was adding pitch and probability of voicing [29] to filter-bank input features.

A total of 5 language models were built on the training and 4 web text sources respectively. Table 4 shows out of vocabulary (OOV) rates on different development and evaluation sets when only training text is used. Unsurprisingly, the OOV rate

Table 4: *Test set out-of-vocabulary rates*

LM	OOV (%)		
	BNB	MNB	MWB
`train`	9.4	7.0	14.9
`+web`	5.9	3.7	3.6

on the Material wide-band (MWB) data text is the highest due to language domain differences, CTS versus BN/BT. The OOV rate on the narrow-band Babel (BNB) and Material (MNB) text is on a larger side as well due to the limited size of the training text. The addition of the web text in Table 4 provides a large decrease in the OOV rate on the MWB and less so on the BNB

and MNB data. Phonetic pronunciations for unseen web text words were derived using a grapheme-to-phoneme model [30].

Given the initial CTS acoustic model (AM), it is interesting to see if additional improvements on the BNB data could generalise to the MNB and MWB data. To investigate this, the average mapped confidence score was monitored both on the BNB and VWB data as the changes were made. Note that confidence scores on the MNB and VWB data were mapped using the same piecewise linear mapping applied to the BNB data. The changes included (*ii*) adding more layers, (*iii*) speaking rate perturbation [31], (*iv*) ensemble [27] and (*v*) unsupervised language model adaptation (Section 4). Figure 1 plots the average mapped confidence score against WER on BNB, MNB and MWB data as the changes were introduced. In the latter two

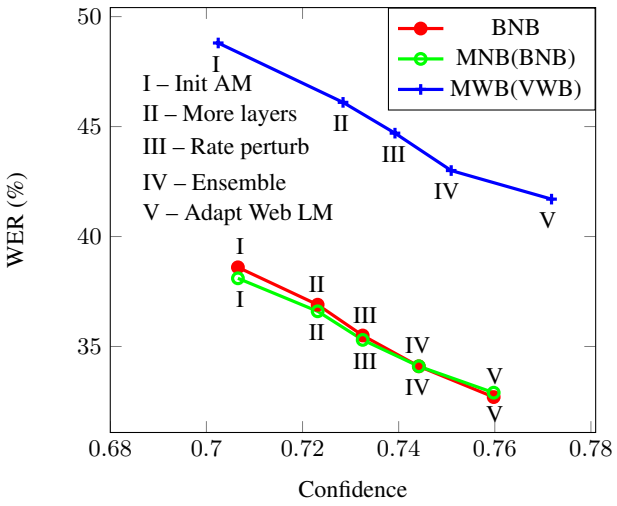

Figure 1: *Cross-domain ASR development*

cases confidence scores were computed on the BNB and VWB data and WERs were evaluated once the MNB and MWB data became available. Figure 1 shows that the narrow-band, BNB and MNB, data is well matched. Though wide-band data exhibits much higher error rate each of the changes increased the average mapped confidence score and decreased WER on the MWB data. The deletion rate was closely monitored throughout the process as changes of more than 1% absolute were found to lead to a significant degradation on the BNB data.

Figure 1 suggests that changes in the order of 0.01 in the average mapped confidence score on the BNB and VWB data lead to generalisation on the MNB and MWB data. When changes are smaller reliability of prediction based on the confidence score alone may drop. The first block in Table 5 shows the impact of varying LSTM cell size from the default value 512 for the initial acoustic model in Figure 1. Decreasing the cell size improves confidence scores and WERs on the BNB and VWB/MWB data. Though increasing the cell size leads to a significant confidence drop the WER performance is not seriously affected. The second block shows the impact of varying the size of utterance chunks, as measured in frames, processed in training. Here, the confidence scores on the BNB and VWB data move opposite ways as the number of frames is reduced. However, if the BNB error rate, including deletion error, is used as an additional factor in prediction the situation can be resolved.

Figure 1 shows that unsupervised adaptation of web language models (LM) improves performance on the BNB data. When these interpolation weights are used on the MNB data

Table 5: *Confidence and error rate based AM tuning*

Param	Value	BNB Conf	BNB Del	BNB WER	VWB Conf	MWB WER
Cell size	768	0.696	8.4	59.0	0.681	49.2
	512	0.707	9.3	58.8	0.703	49.2
	256	0.713	10.6	58.0	0.719	47.0
Chunk width	150	0.707	9.3	58.8	0.703	49.2
	100	0.712	9.9	59.2	0.699	49.6
	50	0.716	12.1	60.7	0.692	52.1

similar performance gains can be observed as shown by the line marked with hollow circles. The same trend can be seen on the MWB data when the weights are estimated on automatically generated VWB transcriptions as shown by the line marked with vertical dashes. Table 6 shows interpolation weights estimated on the hypothesised transcriptions and ensemble WER performance on the MNB and MWB data. The narrow-band data

Table 6: *Interpolation weights and word error rate performance*

Test Set	LM	Source LM weights Babel	TED	Blog	BingA	BingH	WER (%)
MNB	`train`	1.0	0.0	0.0	0.0	0.0	34.1
	`+hyp`	0.68	0.29	0.00	0.02	0.00	32.9
MWB	`uni`	0.2	0.2	0.2	0.2	0.2	43.0
	`+hyp`	0.12	0.19	0.14	0.37	0.19	42.8
	`+conf`	0.04	0.01	0.02	0.55	0.38	42.7

mostly makes use of Babel training and TED talks data. The wide-band data makes most use of Bing data which illustrates the mismatch between CTS and broadcast domains. Similar weight and WER patterns were observed if the actual test set text data was used to estimate weights. This suggests that unsupervised adaptation can also be applied from a related domain.

As described in Section 4 interpolation weights can be optimised alternatively by maximising the average mapped confidence score on the VWB data. Table 6 gives an example of fine-tuning the weights estimated on hypothesised transcriptions. The resulting distribution of weights though significantly different from the one estimated on hypothesised transcriptions yields a small 0.003 gain in the average mapped confidence score and similarly small improvement in WER. This is consistent with the previous finding that gains in the order of 0.01 are required to yield significant improvements in WER.

6. Conclusions

Developing automatic speech recognition (ASR) systems required to operate over broad domains is challenging. This paper looked at a particular scenario where the development training data comes from conversational telephone speech domain whilst target domains include news and topical broadcast. As the target domain data may not always be available during the initial development, this paper proposed to automatically scrape audio in a range of domains from the web and use ASR system confidence score as the primary metric for the development of acoustic and language model components. An automatic procedure based on maximising ASR system confidence score was proposed for interpolating web language models. Experimental results on IARPA MATERIAL Swahili Analysis Pack 1 data show promise of this approach for cross-domain development.

7. References

[1] G. Saon, T. Sercu, S. Rennie, and J. H.-K. Kuo, "The IBM 2016 English conversational telephone speech recognition system," in *Interspeech*, 2016.

[2] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "The Microsoft 2016 conversational speech recognition system," in *ICASSP*, 2017.

[3] M. J. F. Gales, P. C. Kim, D. Y. Woodland, D. Mrva, R. Sinha, and S. E. Tranter, "Progress in the CU-HTK broadcast news transcription system," *IEEE TSAP*, vol. 14, no. 5, pp. 1513–1525, 2006.

[4] M. Afify, L. Nguyen, B. Xiang, S. Abdou, and J. Makhoul, "Recent progress in Arabic broadcast news transcription at BBN," in *Interspeech*, 2005.

[5] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Large vocabulary continuous speech recognition with context-dependent DBN-HMMs," in *ICASSP*, 2011.

[6] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *ICASSP*, 2015.

[7] L. Lamel and J.-L. Gauvain, "Lightly supervised and unsupervised acoustic model training," *Computer speech and language*, vol. 16, pp. 115–129, 2002.

[8] B. Strope, D. Beeferman, A. Gruenstein, and X. Lei, "Unsupervised testing strategies for ASR," in *Interspeech*, 2011.

[9] R. Schlüter, B. Müller, F. Wessel, and H. Ney, "Interdependence of language models and discriminative training," in *ASRU*, 1999.

[10] P. C. Woodland and D. Povey, "Large scale discriminative training of hidden Markov models in speech recognition," *Computer Speech and Language*, vol. 16, no. 1, pp. 25–48, 2002.

[11] Y. Deng, M. Mahajan, and A. Acero, "Estimating speech recognition error rate without acoustic test data," in *Eurospeech*, 2003.

[12] G. Mendels, E. Cooper, V. Soto, J. Hirschberg, M. Gales, K. Knill, A. Ragni, and H. Wang, "Improving speech recognition and keyword search for low resource languages using web data," in *Interspeech*, 2015.

[13] L. Zhang, D. Karakos, W. Hartmann, R. Hsiao, R. Schwartz, and S. Tsakalidis, "Enhancing low resource keyword spotting with automatically retrieved web documents," in *Interspeech*, 2015.

[14] H. Jiang, "Confidence measures for speech recognition," *Speech Communication*, vol. 45, pp. 455–470, 2005.

[15] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus among words: lattice-based word error minimisation," *Computer Speech and Language*, vol. 14, no. 4, pp. 373–400, 2000.

[16] G. Evermann and P. C. Woodland, "Posterior probability decoding, confidence estimation and system combination," in *NIST STW*, 2000.

[17] M. Weintraub, F. Beaufays, Z. Rivlin, Y. Konig, and A. Stolcke, "Neural-network based measures of confidence for word recognition," in *ICASSP*, 1997.

[18] M. S. Seigel and P. C. Woodland, "Detecting deletions in ASR output," in *ICASSP*, 2014.

[19] R. DeMori and M. Federico, "Language model adaptation," in *Computational Models of Speech Pattern Processing*, K. Pointing, Ed. Springer, 1999, vol. 169, pp. 280–303.

[20] M. Bacchiani and B. Roark, "Unsupervised language model adaptation," in *ICASSP*, 2003.

[21] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, 1992.

[22] J. Nocedal and S. J. Wright, *Numerical Optimization*. Springer, 1999.

[23] V. Peddinti, Y. Wang, D. Povey, and S. Khudanpur, "Low latency acoustic modeling using temporal convolution and LSTMs," *Signal Processing Letters*, vol. 25, no. 3, pp. 373–377, 2017.

[24] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, A. Ragni, V. Valtchev, P. C. Woodland, and C. Zhang, *The HTK Book (for HTK Version 3.5)*. `http://htk.eng.cam.ac.uk`: University of Cambridge, 2015.

[25] M. Gales, K. Knill, A. Ragni, and S. Rath, "Speech recognition and keyword spotting for low-resource languages: BABEL project research at CUED," in *SLTU*, 2014.

[26] M. J. F. Gales, K. M. Knill, and A. Ragni, "Low-resource speech recognition and keyword-spotting," in *SPECOM*, 2017, pp. 3–19.

[27] Y. Wang, X. Chen, M. J. F. Gales, A. Ragni, and J. H. M. Wong, "Phonetic and graphemic systems for multi-genre broadcast transcription," in *ICASSP*, 2018.

[28] 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 *ASRU*, 2011.

[29] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *ICASSP*, 2014.

[30] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, 2008.

[31] K. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Interspeech*, 2015.