UNSUPERVISED SPEAKER ADAPTATION FOR TELEPHONE CALL TRANSCRIPTION

R. Wallace[*]

Speech and Audio Research Laboratory
Queensland University of Technology
2 George Street, Brisbane, Australia

K. Thambiratnam, F. Seide

Microsoft Research Asia
5F Sigma Center, 49 Zhi Chun Road
Beijing, P.R. China 100080

ABSTRACT

The use of the PC and Internet for placing telephone calls will present new opportunities to capture vast amounts of un-transcribed speech for a particular speaker. This paper investigates how to best exploit this data for speaker-dependent speech recognition. Supervised and unsupervised experiments in acoustic model and language model adaptation are presented. Using one hour of automatically transcribed speech per speaker with a word error rate of 36.0%, unsupervised adaptation resulted in an absolute gain of 6.3%, equivalent to 70% of the gain from the supervised case, with additional adaptation data likely to yield further improvements. LM adaptation experiments suggested that although there seems to be a small degree of speaker idiolect, adaptation to the speaker alone, without considering the topic of the conversation, is in itself unlikely to improve transcription accuracy.

Index Terms— Speaker adaptation, acoustic model adaptation, language model adaptation, unsupervised adaptation, speech recognition

1. INTRODUCTION

One of the effects of the current mass migration to the Internet as a means of communication is that phone calls will increasingly be made online, with smart devices. Coupled with exponential growth in data storage, this will present new opportunities to automatically capture vast amounts of speech for a particular speaker. Speech recognition will thus have an opportunity to exploit unprecedented amounts of speaker-labeled data for speaker-dependent (SD) recognition, and potentially improve recognition accuracy substantially.

Commercial dictation systems are now available that report near perfect accuracies. This is largely made possible by the use of a formal speaking style, the availability of high-quality and uncompressed audio, and leveraging large amounts of accumulated speaker-labeled training data. Recognition of phone calls could likewise utilise knowledge of the speaker identity for speaker adaptation and, if captured before transmission, could also be performed on high-quality audio. Smart devices, capable of placing telephone calls and simultaneously recording speech or performing real-time recognition, will soon become commonplace. The medium of the phone call is a natural way to collect speech data, as there is no burden placed on the user to complete any sort of initial or ongoing training, other than carrying on with their usual daily activities. Unlike dictation systems, if out-of-the-box transcription accuracy is not satisfying or useful, there is no inconvenience to users, who in any case would still be making the telephone calls. Their data could be used to continuously and transparently improve their models over time. Should this lead to significant improvements in transcription accuracy, a wide range of valuable applications become realisable such as search-able transcripts and real-time multi-modal tele-conferencing.

This paper demonstrates the gains achievable by utilising the speaker identity and up to an hour of adaptation data per speaker, by applying well established techniques in acoustic and language model adaptation to the domain of conversational telephone speech. The effect of the amount of data used, the feasibility of capturing speaker idiolect in an n-gram language model, and the results of supervised and unsupervised adaptation are also presented.

1.1. Prior work

Acoustic model adaptation and the benefits it provides for SD recognition have been clearly established for some time [1, 2]. However, much of the work has been in relation to adaptation when the identity of the speaker is unknown [3], or when the adaptation is based on only a few minutes or a few utterances of speech [4]. The application of telephone calls made from smart devices will have neither of these limitations, and as such, new techniques may be developed to exploit the available data.

Language model (LM) adaptation has also seen significant interest [5], primarily in terms of domain adaptation [6, 7] or topic adaptation [8]. LM adaptation for a particular speaker, however, is much less well known. The essential task of speaker LM adaptation is to capture idiolect, which is a linguistic term referring to the variety of language used by a particular individual. Speaker idiolect has been used to provide discriminative information for speaker identification [9, 10], whilst linguists concede that uniquely characterising a speaker's idiolect is likely to be very difficult indeed [11]. Nevertheless, it is conceivable that knowledge of the characteristic language used by the speaker would be useful for training a language model for speech recognition. It can be difficult to distinguish between the effects of speaker adaptation and topic adaptation, as speakers may tend to focus on a certain topic throughout a particular discussion or series of discussions, for example in lecture speech adaptation [8]. Where attempts have been made to separate the effects of speaker and topic adaptation, both have been shown to contribute to improvements in decoding accuracy [12, 13].

Section 2 describes the data and procedures used, while the following two sections describe experiments in acoustic and LM adaptation respectively.

2. EXPERIMENTAL PROCEDURE

Telephone calls that are recorded before transmission on smart devices may be captured at a high bandwidth, with a close talking mi-

[*] Work performed during an internship at Microsoft Research Asia

978-1-4244-2354-5/09/$25.00 ©2009 IEEE

crophone. This will, in itself, lead to much improved recognition accuracy compared to a low bandwidth signal captured at a telephone exchange. As such a database with a substantial quantity of speech per speaker was not available, a subset of the Switchboard-1 (SWB1) conversational telephone speech corpus was used, which provides an adequate basis for comparison of the relative gains of SD adaptation. The 62 speakers with the largest quantity of speech were selected, with 31 of these used as the development set for tuning, and 31 as the test set. Sixty minutes of adaptation data (around 13 conversation sides), and ten minutes of evaluation data were selected per speaker. Word error rates (WER) are reported as an average across the test speakers. Normalised hand transcriptions were used for supervised adaptation experiments.

The baseline acoustic model was a 72-mixture, 18000 tied state ML-trained HMM set trained on 1700h of data comprising The Fisher English Corpus Part 1 and 2 [14] (Fisher). The baseline LM was a Katz back-off 3-gram model, with 52k/1.4m/1.5m 1/2/3-grams respectively, also trained on the Fisher set. All experiments used a 22.6k vocabulary. The baseline WER on the SWB Eval2000 evaluation data set was 30.1%. The evaluation data used in experiments reported below was a more difficult, with a WER of 35.1%.

Unsupervised adaptation experiments used the 1-best word-level transcript of the adaptation data, produced by decoding with the baseline acoustic and LMs, with a WER of 36.0%. Posteriors were estimated for confidence thresholded adaptation experiments using lattices generated with a beam width of 240.

3. ACOUSTIC MODEL ADAPTATION

Maximum Likelihood Linear Regression (MLLR) [1] and Maximum A Posteriori (MAP) estimation [2] are two of the most widely used methods in acoustic model adaptation. MLLR can provide robust adaptation for limited amounts of adaptation data, by applying a linear transformation to groups of mixtures, whilst MAP estimation adjusts the individual mixtures for which observations occur, using a Bayesian framework to incorporate the background (usually speaker-independent) model as prior knowledge. In any case, for optimal decoding accuracy the complexity of the adaptation needs to be matched to the amount and quality of adaptation data available.

Both global (Global MLLR) and 256-class regression tree-based (Rtree MLLR) MLLR adaptation were investigated. Speech and non-speech phones were forcibly separated for transform estimation and only mean vectors were adjusted. MAP estimation (MAP) with a priori weight 10 was used to re-estimate mixture means, variances and weights. All three adaptation approaches were used in both isolated and chained configurations.

As shown in Table 1, Global MLLR provided WER gains of 2.7% and 2.2% absolute for supervised and unsupervised adaptation respectively. The loss from using unsupervised transcripts was only 0.5%. Rtree MLLR gave WER gains of 5.7% and 4.4% for supervised and unsupervised adaptation. The larger gains here indicate that 60min of data was sufficient to estimate up to 256 separate linear transformations, even using an automatic transcript, however the gap between supervised and unsupervised adaptation of 1.3% was larger than Global MLLR because of the increased complexity of the transformation and therefore sensitivity to transcript errors. Using MAP alone gave a 5.9% WER gain for supervised adaptation, which was greater still than Rtree MLLR. However, gains for unsupervised MAP were considerably less. MAP adaptation is more sensitive to errors than MLLR due to the lack of mixture pooling.

By cascading transformations of increasing complexity, each stage was able to progressively improve the frame alignments for the subsequent stage. As shown, a cascaded Global + Rtree MLLR system resulted in an 0.5%/0.4% gain for supervised/unsupervised adaptation respectively over Rtree MLLR alone. Cascading Global + Rtree MLLR + MAP lead to an additional gain of 3.2% and 3.2% compared to MAP alone. Overall, total WER gains of 9.1% and 6.0% absolute (26% and 17% relative) were achieved for the supervised and unsupervised adaptation respectively. Unsupervised adaptation was successful in that it retained 66% of the gain of supervised adaptation, using transcripts with a WER of 36.0%.

Adaptation technique	Supervised	Unsupervised
Baseline	35.1	35.1
Global MLLR	32.4	32.9
Rtree MLLR	29.4	30.7
MAP	29.2	32.3
Global+Rtree MLLR	28.9	30.3
Global+Rtree MLLR+MAP	26.0	29.1

Table 1. Acoustic model adaptation using 60min of adaptation data per speaker (% WER). Unsupervised adaptation used automatic transcripts with 36.0% WER.

These results were compared to those achieved without knowledge of the speaker identity. It was found that utilising 210 hours of additional SWB1 data to adapt the baseline models resulted in a 31.6% WER - a 3.5% absolute gain. This gain is likely a result of more data rather than domain or channel mismatch. Most importantly, this gain was well short of the 9.1% gain from SD adaptation.

Table 2 shows the effect of the amount of adaptation data on the gains, using 10min, 30min or 60min of data per speaker. As expected, the more complex adaptation techniques accommodated more data before gains saturated. Importantly, gains did not appear to be saturated at 60min for either the supervised or unsupervised instances of Global+Rtree MLLR+MAP, thus indicating that further gains would be observed with even more data. This is beneficial for telephone call transcription since it is conceivable that most users would easily create hundreds of hours of call data, all of which could be used for acoustic model adaptation or training.

3.1. Data selection using confidence measures

In order to address the 3.1% absolute WER gap between supervised and unsupervised adaptation, the use of confidence thresholded transcripts was investigated. This was shown to be beneficial in [15]. To preserve as much adaptation data as possible, thresholding was performed on a per-word basis, rather than a per-utterance basis, as is commonly done.

First, each unsupervised transcript was confidence scored using the posterior probability from the recognition lattice. Occurrences of words with confidence scores above/below ρ were then marked as certain/uncertain. Each HMM (including its tied states) was then cloned to give an uncertain equivalent. The recognition lexicon was also cloned by inserting an uncertain version of each word with pronunciations mapped to corresponding uncertain phones. Adaptation was then performed using the confidence-marked transcript. By introducing uncertain HMMs, frames for low confidence words were excluded from adaptation parameter estimation. The uncertain models were discarded after adaptation.

Adaptation results shown in Table 2. For MLLR, thresholding was not useful, which can be explained by the error robustness provided by mixture clustering. However, there was a small gain of 0.3% absolute for MAP adaptation using a threshold of $\rho = 0.5$. Error rates at $\rho = 0.8$ were worse than $\rho = 0.5$ because of a reduction in effective adaptation data - at $\rho = 0.5$, 75% of frames were kept after thresholding, while only 50% were kept for $\rho = 0.8$.

	Supervised			Unsupervised			Conf. Thresh	
	10min	30min	60min	10min	30min	60min	60min	
							$\rho = 0.5$	$\rho = 0.8$
Baseline	35.1			35.1			-	
Global MLLR	32.7	32.4	32.4	33.1	32.9	32.9	-	
Global + Rtree MLLR	30.8	29.5	28.9	31.8	30.8	30.3	30.2	30.3
Global + Rtree MLLR + MAP	29.9	27.6	26.0	31.7	29.9	29.1	28.8	29.2

Table 2. Supervised, unsupervised and confidence thresholded acoustic model adaptation using various amounts of adaptation data per speaker (%WER). The threshold for confidence thresholding is given by ρ.

4. LANGUAGE MODEL ADAPTATION

Compared to acoustic model adaptation, the potential of LM adaptation for SD recognition is less well understood. Of previous studies [6, 7, 8, 12, 13], often the effect of adapting to the speaker's idiolect cannot be easily separated from that of adapting to the topic of the discussion, or the speaker's usual topic of discussion. One approach to deal with this ambiguity is to ignore it and deal with the effects jointly, which has the downside of limiting the applicability of the results to different corpora, as the amount of mutual information between speaker identity and topic will likely depend on the corpus, speakers and topics in question. Alternatively, the SWB1 corpus in particular is useful for examining the effect of adaptation to a speaker's idiolect alone (as used in [9]), the results of which are presented here.

The SWB1 corpus consists of calls between two previously unacquainted parties, each prompted prior to recording to discuss a pre-determined topic. This reduces the correlation between speakers and topics, as the topics are disjoint and chosen somewhat randomly. In the experiments reported here, for each speaker, there was no topic overlap between any two conversations, both in evaluation and adaptation data. This allowed any gains from LM adaptation to be attributed confidently to the learning of idiolect only.

LM adaptation was performed using the common approach of linear interpolation [8, 13, 16]. Various combinations of data sources were used to generate the LMs in experiments. The background LM was trained on the entirety of the Fisher corpus. The remainder of data was sampled from the SWB1 corpus in the form of two separate sections of 62 hrs and 210 hrs from various speakers (excluding speakers in the evaluation and development sets), as well as 60 min of adaptation data from each of the test speakers. Interpolation weights were optimised on a separate development set of 31 different development speakers. Both WER and perplexity (PPL) results are presented.

Table 3 shows that unsupervised LM adaptation using transcripts with 36.0% WER was not useful. The rest of this section thus focuses on trying to understand the failings of SD LM adaptation by examining the supervised results. Table 3 shows that interpolation of the SD models with the background LM resulted in a notable WER gain of 0.4% absolute and a perplexity gain of 6.4, over using just the the background LM. However, interpolating the background LM with an LM trained on 62 hrs of unrelated non-SD data lead to a larger WER gain of 0.7%. Unfortunately, no further gains were observed when this resulting LM was then interpolated with a SD LM. Even worse was that using additional non-speaker related data continued to improve the WER (0.7%, 1.0%, 1.1% using 62, 210 and 272 hours respectively), and still no gains were observed from further interpolating with the SD LM's. Sanity experiments such as using only 1/2-gram SD models were performed to verify that the lack of gain was not due to under-training. Also out-of-vocabulary (OOV) rates dropped from 0.6% to 0.5% in the best case, thus excluding any gains being related to new vocabulary learning.

Adaptation data				Supervised		Unsupervised	
Fisher	62h	210h	SD	WER	PPL	WER	PPL
X				35.1	97.0	35.1	97.0
X			X	34.7	90.6	35.1	94.9
X	X			34.4	89.3	35.0	94.5
X	X		X	34.4	87.9	35.0	93.7
X		X		34.1	86.5	35.0	93.1
X	X	X		34.0	85.1	35.0	92.5
X	X	X	X	34.0	84.0	35.0	91.9

Table 3. Language model adaptation using combinations of Fisher and Switchboard (SWB1) data. Speaker-dependent (SD) adaptation used a 3-gram model trained on 60min of adaptation data/speaker.

Overall, the results indicated that interpolating with large amounts of non-SD data was better than using a small amount (60 min) of tightly related speaker dependent data. Note that folding in the SD LM always improved the perplexity, and examining the resulting LMs did reveal that SD information was being learned, such as the unigrams related to a speaker's city of residence, or commonly used disfluency n-grams. Nevertheless, the trusted axiom of "more data" seemed to outweigh any SD benefits. Of course, another fair explanation is that 60 min of speech is insufficient to capture idiolect. Thus, text expansion techniques were investigated to increase the quantity of adaptation data per speaker.

4.1. TFIDF-based text expansion

The TFIDF (term frequency inverse document frequency) is a popular means for text expansion measure for LM adaptation [8, 16]. It is commonly used to measure document similarity, and is computed using the cosine similarity between word count vectors of documents weighted by word importance. For LM adaptation, TFIDF is used to augment adaptation data with similar documents from a background corpus. In previous studies that did not isolate topic and idiolectic effects, this approach has been moderately successful. The experiments reported here aim to validate whether such gains are indeed from learning idiolect, or just a by-product of topic learning.

To ensure consistency with learning idiolect, a number of important modifications were made. Since the goal was to adapt n-grams, this necessarily assumes that idiolect is embodied in n-gram statistics. Therefore, document vectors for TFIDF were computed using n-gram counts, rather than single-word counts. Secondly, it was assumed that idiolectical patterns would occur on an utterance-level - entire conversations in the background data would not necessarily be representative of another speaker's idiolect. Therefore similarity was computed on a per-utterance rather than per-document basis, with a sliding window over adjacent utterances for smoothing.

Unfortunately, even with these modifications, there were no observable gains from using TFIDF expansion, as shown in Table 4. This lack of gain is best explained by the lack of topic overlap between adaptation and evaluation data, which was by experiment design. Since TFIDF is typically used as a topical similarity measure, it is not unreasonable that no gains are observed when topic overlap was removed.

Hours selected	Index order	Pool width	WER	PPL
3	1, 2, 3	1	34.0%	84.1
3	1, 2, 3	3	34.0%	84.3
20	1, 2, 3	3	34.1%	84.7
3	1, 2	3	34.0%	84.3
3	1	3	34.0%	84.6
3	1, 2	5	34.0%	84.4

Table 4. Supervised language model adaptation using TFIDF. "Hours selected" is the total number of hours selected using TFIDF similarity to grow adaptation data; "Index order" is the n-gram order of the elements of the vectors used in TFIDF scoring; "Pool width" is the number of adjacent utterances used to in the sliding window for smoothing TFIDF statistics.

LM	Self	Other	Δ n-grams		
	PPL	PPL	1-gram	2-gram	3-gram
SD	97.1	107.1	-	-	-
Fisher+SD	90.6	93.7	30	650	38
Fisher+62	-	-	1.4k	39.6k	5.2k

Table 5. Average PPL using speaker's LM, best PPL of other speaker's LM, and increase in 1/2/3-gram counts for various interpolation approaches.

4.2. Discussion

Overall, the LM adaptation experiments demonstrate that idiolect learning via LM adaptation does not significantly affect WER. It seems though that topical overlap is critical since prior work in LM adaptation has been successful. It can be concluded then that for tasks with large amounts of speaker data, it is not necessary to craft a speaker-dependent LM. Possibly it may be sufficient to use an ensemble of well-trained topic-dependent LMs.

Although SD LM adaptation did not improve the WER, there were always gains in perplexity, as shown in Table 3. To investigate whether this change was meaningful, PPL was evaluated for each of the 31 speakers' evaluation data using each of the 31 SD LM's. Care was taken to maintain a fixed vocabulary so that PPLs were comparable. In all cases, as shown in Table 5, the speaker's own LM achieved the lowest perplexity on the speaker's own evaluation data, demonstrating that some degree of idiolect was indeed being learnt.

The change in n-gram counts gives further insight into how idiolect was being learnt. LM adaptation can affect n-gram probabilities by either adjusting an existing n-gram probability (adjustment) or promoting a lower order n-gram to a higher order (promotion). The gain in probability is generally more from promotion, since back-off weights significantly penalise low order n-grams. Table 5 shows the number of new (or promoted) n-grams for various interpolation approaches. Clearly, SD adaptation directly from the background model introduced far fewer new n-grams than adapting to 62 hrs of speaker-independent data. This should be noted alongside the larger gains in WER and PPL for speaker-independent adaptation (a gain in WER of 0.7% compared to 0.4%). If promotion is assumed to give larger probabilities than adjustment, then it is clear why the SD models do not improve WER; SD interpolation is mostly an adjustment adaptation, that will indeed manifest PPL improvements but not necessarily WER improvements.

5. CONCLUSIONS

By applying well established techniques in acoustic and LM adaptation to the domain of speaker adaptation for conversational telephone speech, some promising findings have been demonstrated which will prove useful for improved recognition of telephone calls on smart devices. Using one hour of automatically transcribed speech per speaker at an error rate of 36.0%, unsupervised acoustic model adaptation reduced the word error rate from 35.1% to 28.8%, an 18% relative gain, equivalent to 70% of the gain from supervised adaptation, with strong indications that the accuracy could be improved further still with additional data. LM adaptation experiments indicated that although there seems to be a small degree of speaker idiolect, adaptation to the speaker alone, without considering the topic of the conversation, is in itself unlikely to improve transcription accuracy.

As more extensive data becomes available, interesting questions may be raised concerning the boundaries between idiolect, topic and domain, and how incomplete knowledge of such things can be utilised in an unsupervised way to drive word error rates downward in the long-term.

6. REFERENCES

[1] C.J. Leggetter and P.C. Woodland, "Flexible speaker adaptation using maximum likelihood linear regression," in *Proc. Eurospeech'95*, 1995.

[2] J.-L. Gauvain and Chin-Hui Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 2, pp. 291–298, Apr 1994.

[3] M.J.F Gales et al., "Progress in the CU-HTK broadcast news transcription system," *IEEE Transactions Audio, Speech and Language Processing*, 2006.

[4] Chao Huang, Tao Chen, and E. Chang, "Speaker selection training for large vocabulary continuous speech recognition," in *ICASSP '02*, 2002, vol. 1, pp. 609–612.

[5] Jerome R. Bellegarda, "An overview of statistical language model adaptation," in *Adaptation-2001*, 2001.

[6] M. Bacchiani and B. Roark, "Unsupervised language model adaptation," in *ICASSP '03*, 2003.

[7] Diego Giuliani and Marcello Federico, "Unsupervised language and acoustic model adaptation for cross domain portability," in *Adaptation-2001*, 2001.

[8] D. Willettt, T. Niesler, E. McDermott, Y. Minami, and S. Katagiri, "Pervasive unsupervised adaptation for lecture speech transcription," in *ICASSP '03*, 2003, vol. 1, pp. 292–295.

[9] George Doddington, "Speaker recognition based on idiolectal differences between speakers," in *EUROSPEECH '01*, 2001.

[10] F. Weber, L. Manganaro, B. Peskin, and E. Shriberg, "Using prosodic and lexical information for speaker identification," in *ICASSP '02*, 2002, vol. 1, pp. 141–144.

[11] Malcolm Coulthard, "Author identification, idiolect and linguistic uniqueness," *Applied Linguistics*, 2004.

[12] Yuya Akita and Tatsuya Kawahara, "Language model adaptation based on PLSA of topics and speakers," in *INTERSPEECH 2004 - ICSLP*, 2004.

[13] G. Tur and A. Stolcke, "Unsupervised language model adaptation for meeting recognition," in *ICASSP '07*, 2007.

[14] Christopher Cieri, et al., "Fisher English Training Speech Part 1 & 2 Transcripts," Linguistic Data Consortium, 2005.

[15] Michael Pitz, Frank Wessel, and Hermann Ney, "Improved mllr speaker adaptation using confidence measures for conversational speech recognition," in *ICSLP 2000*, 2000.

[16] H. Nanjo and T. Kawahara, "Language model and speaking rate adaptation for spontaneous presentation speech recognition," *Speech and Audio Processing, IEEE Transactions on*, 2004.