Accent- and Speaker-Specific Polyphone Decision Trees for Non-Native Speech Recognition

Dominic Telaar[1], *Mark C. Fuhs*[2]

[1]Cognitive Systems Lab, Karlsruhe Institute of Technology, Germany
[2]M*Modal Inc., Pittsburgh, PA, USA

dominic.telaar@kit.edu, mark.fuhs@mmodal.com

10.21437/Interspeech.2013-733

Abstract

Acoustic models in state-of-the-art LVCSR systems are typically trained on data from thousands of speakers and then adapted to a speaker using, e.g., various combinations of CMLLR, MLLR and MAP. This adaptation step is particularly important for speakers with accents that are not well represented in the training set. The present study explores how to improve performance on South-Asian-accented speakers (SoA-accented) with the availability of thousands of US-accented, hundreds of SoA-accented, and tens of hours of speaker-specific training data. We employ a decision tree similarity measure to analyze how varying co-articulations across accents and people manifest themselves in the decision tree. Modeling these variations in addition to adapting the GMMs of an existing baseline system to a speaker improved WER for small systems (1k GMMs), but improvement for systems with larger trees (2k, 3k GMMs) was modest. Overall, GMM adaptation/retraining yields significant performance benefits, and training a SoA-accent-specific system is particularly worthwhile when lacking speaker adaptation data.

Index Terms: speech recognition, accented speech, decision trees

1. Introduction

Given the widespread use of English as a second language, many research groups have addressed the issue of increasing ASR performance for non-native speakers [1, 2, 3, 4, 5]. Typical differences between native and accented speakers are different pronunciations, grammar and speaking style. Due to a lack of accented speech resources, recognizers normally are trained on a collection of native speech and various schemes are employed to adapt the ASR system to a specific accent [4] or multiple accents at once [1]. Building on these findings, in this work we compare accent specific systems to speaker specific systems and analyse differences between speakers within an accent group due to significantly bigger accented speech resources at our disposal.

The motivation for analysing speaker-dependent systems is based on [6], where Oh found that non-native speakers of English do have different co-articulations compared to native speakers, but gradually (depending on their experience with English) tend to adopt native-like co-articulations of English. Therefore, we expect a retrained decision tree to emphasize different co-articulations for non-native speakers and in turn lead to an improved word error rate. Previous work has shown that adapting the polyphone decision tree to a new domain, e.g. new languages as in [7], helped to reduce word error rate via Polyphone Decision Tree Specialization (PDTS). Subsequent work [4, 5, 8] successfully used PDTS to cross-adapt various systems. We focus on polyphone decision trees [9], since nearly all modern speech recognition systems use these to model phonetic co-articulations.

To our knowledge we are the first to quantify the difference between polyphone decision trees and link the final word error rate of the systems to an information theoretic analysis of these trees. We investigate how best to leverage different types of training data to improve recognition accuracy for South-Asian-accented (SoA-accented) speakers on a medical dictation task. The available training data includes, in geometrically decreasing quantities, mostly US-accented, SoA-accented and speaker specific. We choose SoA-accented speakers for two reasons. First, since both India and Pakistan have been British colonies, we expect the SoA-accented speakers in our database to share some co-articulations. Hence, if accented polyphone decision trees are beneficial we expect to observe an improvement in accuracy for a SoA-accented tree compared to our baseline. Second, South-Asian accents are the most common non-US accents in our training data.

The remainder of this paper is structured as follows: Section 2 describes our training and test sets. Section 3 describes an information theoretic measure for quantifying differences between decision trees. Section 4.1 describes the experimental setup of our recognizers. Section 4.2 covers recognition performance experiments, where we examine how beneficial the US-accented data is without and especially with adaptation, as well as the benefits of SoA-accented and speaker-dependent (SD) decision trees at different tree sizes. We explore to what extent these different sets of training data lead to different decision trees in Section 4.3. Finally, we conclude our work in Section 5.

2. Database

Table 1: *Training database statistics.*

Database	Speech data (hr)	#speakers
Baseline	1,405.3	1,863
US Acc.	1,099.7	1,387
Non-SoA Acc.	1,311.7	1,740
SoA Acc.	251.7	123

In the medical dictation domain, a single speaker may dictate tens of hours of medical reports which are then manually transcribed by a medical transcriptionist, or, in our case, an initial automated transcript is produced and afterwards edited by a

This work was done while Mr. Telaar was an intern at M*Modal.

transcriptionist. For the experiments, 8kHz telephony data from 1,863 speakers were selected to be part of the various training databases. Speakers accents were manually labeled by a native US English speaker based on speech samples and the speakers name. Table 1 outlines the training databases used. Baseline contains a maximum of one hour per speaker. All other databases speakers are subsets of the Baseline set of speakers. US Acc. contains all training data spoken by native US English speakers. Non-SoA Acc. contains speech from all speakers who are not categorized as being SoA-accented, including both US and other accented speech. Finally, SoA Acc. contains only SoA-accented speech, but data for each speaker is not limited to one hour per speaker as in the Baseline data set. Our test set consisted of 94k running words from 11 speakers with additional adaptation data ranging from 65 to 175 hours per speaker (see Table 3).

3. Polyphone decision tree similarity measure

In order to quantify how different co-articulations by accent groups and people are reflected in the polyphone decision trees, we employed a similarity measure based on conditional entropy. A polyphone decision tree T provides a partition of all possible polyphones into n clusters (equivalent to the n GMMs of a system), each cluster corresponds to a leaf node in the tree. Denote the probability that a randomly selected polyphone ctx falls in a particular cluster $C_i (i \in [1, n])$ by

$$P(C_i) = \sum_{ctx \in C_i} P(ctx) \tag{1}$$

where $P(ctx)$ is the probability of occurrence of the polyphone ctx estimated from the training data. The joint probability of a polyphone falling both into clusters C_i in T_1 and C_j in T_2 is

$$P(C_i, C_j) = \sum_{ctx \in C_i \cap C_j} P(ctx) \tag{2}$$

and the joint entropy of this distribution over all clusters in trees T_1 and T_2 is

$$H(T_1, T_2) = -\sum_{i,j} P(C_i, C_j) \log [P(C_i, C_j)] \tag{3}$$

Finally,

$$P(C_j|C_i) = P(C_i, C_j) / P(C_i) \tag{4}$$

is the probability of a polyphone being in C_j, given that it is in C_i. The entropy of this distribution measures how dispersed the polyphones in C_i are across the clusters of T_2. Averaging over all C_i leads to the conditional entropy of tree T_2 given T_1,

$$H(T_2|T_1) = -\sum_{i,j} P(C_i, C_j) \log [P(C_j|C_i)] \tag{5}$$

In order to compare similarities between trees with different number of models, we measure the fraction of joint entropy that is not mutual:

$$DIFF(T_1, T_2) = 1 - \frac{I(T_2; T_1)}{H(T_1, T_2)} \tag{6}$$

$$= \frac{H(T_1, T_2) - I(T_2; T_1)}{H(T_1, T_2)} \tag{7}$$

$$= \frac{H(T_1|T_2) + H(T_2|T_1)}{H(T_1, T_2)} \tag{8}$$

$I(T_2; T_1)$ corresponds to the mutual information of T_1 and T_2. If $DIFF(T_1, T_2) = 0$, the trees are identical; if $DIFF(T_1, T_2) = 1$, the clusters are not correlated across trees.

4. Experimental setup and results

4.1. System setup

For all experiments, we extracted a 32 dimensional feature vector from a standard MFCC front-end using the first 12 cepstral coefficients along with 11 delta and 9 delta-delta features. The baseline acoustic model consisted of 3k triphone Gaussian mixture models (GMMs) initialized with split-and-merge training and a total of 50k Gaussian distributions. All subsequent recognizers consisted of 50k Gaussian distributions regardless of the number of triphone models. We used a 4-gram language model with a vocabulary of 51k words. Language model weights, word and filler penalties were the same for all experiments. To better adapt to non-native accents, speaker adaptation was performed using MAP, or, for individual GMMs exceeding 1,000 adaptation samples, ML retraining.

4.2. Modeling south-asian-accented speech

To evaluate the best possible way of modeling SoA-accented speech, we started with our baseline system with 3k GMMs trained on the Baseline database (see Section 2). For comparison, we trained systems with 3k GMMs based on the Baseline and SoA Acc. databases, with weight increasingly shifted towards SoA-accented data until only accented data was used. To separately explore the impact of accented training data on the polyphone decision tree and the GMMs, one set of systems used the baseline tree, while, for another set of systems, the tree was retrained with the same training data as the GMMs. Results of adapted and unadapted recognizers can be seen in Figure 1.

The figure shows that training on SoA-accented data helps to improve the speaker-independent (SI) systems: the purely SoA-accented system has a 27.6% relative reduction in WER compared to the baseline system. After adaptation, however, performance of all systems was about the same, indicating that speaker-adapted performance was not improved by using appropriately accented training data for the SI system. A similar performance of the adapted GMMs was expected since most models (¿90% on average) were ML retrained due to their high adaptation data count.

To measure the impact of speaker-dependent (SD) systems, we trained baseline, SoA-accented and for each speaker in our test set systems with 1k, 2k, and 3k GMMs and 50k Gaussian distributions each. Table 2 lists the word error rates of all these systems. In order to achieve 3k GMMs during clustering with the SD systems, we had to add SoA-accented data to the tree training process due to the limited amount of training data for some speakers. The weight of the SoA-accented data was chosen such that the speaker with the least amount of data achieves 3k GMMs. The resulting ratio between this speaker and SoA-accented data (7:3) was used for all speakers to generate com-

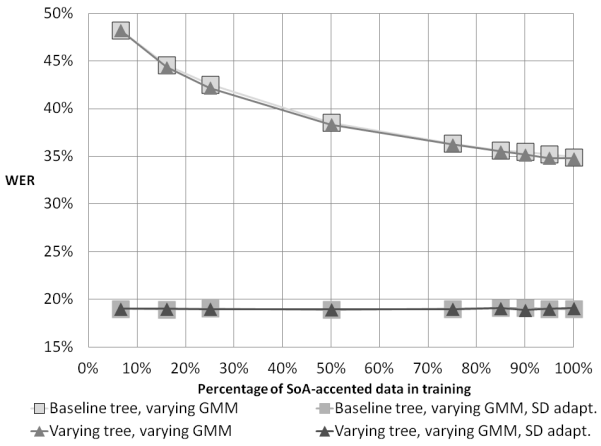

Figure 1: *WER of systems with 3k models with data weighted towards SoA or general speakers for GMMs only or polyphone decision tree and GMMs.*

parable results across speakers. For smaller systems (1k, 2k GMMs), only speaker data was used. While the SD systems outperformed the adapted baseline (15% relative) and adapted SoA-accented systems (11% relative) with 1k models, their advantage diminished with increasing number of models (refer to Table 2).

Table 2: *WER of adapted baseline, adapted SoA-accented and SD systems with varying number of GMMs.*

	1k GMMs	2k GMMs	3k GMMs
Baseline	23.91%	19.53%	19.01%
SoA Systems	22.80%	19.54%	19.09%
SD Systems	**20.21%**	**19.32%**	**18.77%**

Table 3 lists the word error rates per speaker of the 3k GMM systems. Given the large amount of speaker data, performance varies little across adapted baseline, adapted SoA-accented, and SD systems. Instead, the size of the tree plays a much stronger role in final system performance than tailoring the tree to each speaker (refer to Table 2).

Table 3: *WER of adapted baseline, adapted SoA, and SD systems, with 3k GMMs each.*

Spk	Adapt(hr)	Test(hr)	Baseline	SoA	SD
1	105.1	2.28	5.9%	5.8%	**5.7%**
2	88.6	0.89	**17.3%**	17.5%	17.4%
3	87.1	0.80	23.1%	**22.9%**	23.1%
4	89.8	1.11	11.9%	**11.8%**	12.0%
5	76.5	1.13	22.1%	21.8%	**21.2%**
6	92.7	0.98	16.8%	**16.2%**	16.8%
7	80.6	1.02	**22.5%**	**22.5%**	**22.5%**
8	174.5	1.55	28.0%	27.9%	**27.5%**
9	87.6	0.73	**27.4%**	27.8%	27.6%
10	65.1	0.45	18.9%	19.2%	**18.2%**
11	88.8	0.91	24.5%	26.3%	**23.6%**
All	1036.4	11.85	19.0%	19.1%	**18.8%**

4.3. Decision tree analysis

To further analyse the results of the previous section we use the similarity measure introduced in Section 3. Before interpretation of results is possible, we had to gauge the range of results we can expect for similar and dissimilar decision trees. Our approach is as follows: to estimate the score for similar trees, we divided the US-accented database with 1,387 speakers into six equal sets, each with roughly 230 speakers and 180 hours of speech. Decision trees were build with 1k, 2k, and 3k models. To estimate the score for very different trees, we computed the similarity between the previously trained US-accented systems and a French broadcast news recognizer. Since only a subset of phones were shared between US English and French, phones were mapped based on their IPA classification. Two phones were mapped to each other if they differed in only one attribute (e.g. central versus front position of the tongues tip for vowels).

Similarity between English and French trees was then only computed on the mapped phones (see Table [tab:Cond.-entropies-for]). In addition to these extremes, we also included comparisons comprising our systems of interest: the SoA-accented system trained on 250 hours of SoA-accented speech and speaker-dependent systems trained on the adaptation data of the test speakers. In order to verify that we could estimate the speaker-dependent trees reliably, we randomly split each speakers data into three equally sized parts and trained three trees on two parts each.

The average conditional entropy between speaker-dependent trees of the same speaker is 0.14 (between-speaker std. dev. 0.02; within-speaker std. dev. 0.004; one speaker was excluded as an outlier, who inexplicably had entropy 3x the average). For comparison, we split the SoA-accented data into five parts with varying numbers of speakers and 50 hours of speech each. The average conditional entropy between the SoA-accented systems is 0.26, which is significantly higher compared to the speaker-dependent trees.

This suggests that training of the speaker-dependent trees is not unstable due to data sparsity. Results of our similarity measure are shown in Table 4. They show that the variation between US-accented and SoA-accented speakers is greater than between different (random) groups of US-accented speakers. Single speakers have the largest dissimilarity to both the US-accented systems and, surprisingly, to the SoA-accented system as well.

Table 4: *Conditional entropy of various trees trained on US-accented English (US), SoA-accented English (SoA), French (FR), and single test speakers (SD).*

T_1(GMMs)	T_2(GMMs)	$H(T_1\|T_2)$	$DIFF(T_1, T_2)$
US(1k)	US(1k)	0.07	0.021
US(2k)	US(2k)	0.10	0.027
US(3k)	US(3k)	0.15	0.038
US(2k)	SoA(2k)	0.36	0.085
SD(2k)	SoA(2k)	0.56	0.123
SD(2k)	US(2k)	0.61	0.131
US(2k)	FR(2.25k)	0.51	0.190

In an additional experiment we used our similarity measure to compare the speaker-dependent systems to the various systems in Figure 1 where training data for decision tree and GMMs was either weighted towards SoA-accented speakers or general speakers. The results are shown in Figure 2. De-

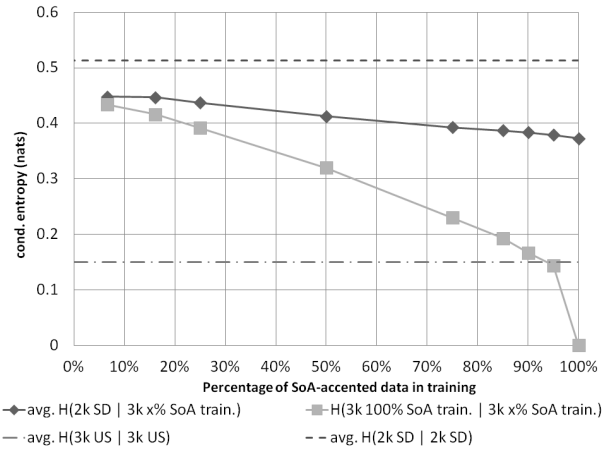

Figure 2: *Conditional entropy of speaker-dependent systems (SD) and SoA only system, given systems with data weighted towards SoA or general speakers for polyphone decision tree and GMMs (SoA).*

picted are the similarities between a system solely trained on SoA-accented data and the weighted systems. As expected, as the weighting of SoA-accented data increases, the conditional entropy drops monotonically to 0. However, the conditional entropy of the speaker-dependent systems given the various weighted systems does not reduce entropy in the same fashion. This supports the hypothesis that there are co-articulations particular to individual speakers that are not systematic even within a group of similarly accented speakers, limiting the potential benefit of accent-specific trees relative to SD trees. This is consistent with the performance observed on small systems in Table 2.

5. Conclusion

While the speaker-independent SoA-accented system clearly outperformed systems built on a mixture of SoA-accented and US-accented data, the performance after adaptation was nearly the same for all systems. Indeed, individual speakers are sufficiently different that adapting to individual speakers vastly improves performance over adapting to a general accent group. We introduced a similarity measure for polyphone decision trees in Section 3 and found a similar pattern: polyphone decision trees trained on SoA-accented speech differed from trees trained on US-accented speech, and both were dissimilar to the speaker dependent trees. However, despite large differences between US-accented trees, SoA-accented trees and SD trees nearly 2/3 of the entropy between English and French trees the impact of the polyphone decision tree on WER decreased quickly as the number of models grew. Given the amount of speaker specific adaptation data, GMM adaptation/retraining provided significant gains. However, using the data to build accent-specific or speaker-specific trees was only beneficial for smaller trees (1k GMMs), whereas improvement was minimal for larger systems (2k, 3k GMMs).

6. Acknowledgements

Our gratitude to Monika Woszczyna, Thomas Schaaf and Shahid Durrani for insightful discussions.

7. References

[1] Bouselmi, G. and Fohr, D. and Illina, I.,Multi-Accent and Accent-Independent Non-Native Speech Recognition, InterSpeech Proc., 2703–2706, 2008.

[2] Deng, Y. and Li, X. and Kwan, C. and Raj, B. and Stern, R., Continuous Feature Adaptation for Non-Native Speech Recognition, International Journal of Signal Processing, 3(4):230–237, 2006.

[3] Oh, Y.R. and Kim, H.K.,On the use of Feature-Space MLLR Adaptation for Non-Native Speech Recognition, ICASSP Proc., 4314–4317, 2010.

[4] Wang, Z. and Schultz, T. and Waibel, A.,Comparison of Acoustic Model Adaptation Techniques on Non-Native Speech, ICASSP Proc., 1:540–543, 2003.

[5] Nallasamy, U. and Metze, F. and Schultz, T., Enhanced Polyphone Decision Tree Adaptation for Accented Speech Recognition, InterSpeech Proc., 2012.

[6] Oh, E.,Coarticulation in non-native speakers of English and French: An acoustic study, Journal of phonetics, 36(2):361–384, 2008.

[7] Schultz, T. and Waibel, A., "Polyphone Decision Tree Specialization for Language Adaptation, ICASSP Proc., 1707–1710, 2000.

[8] Stuker, S., Modified Polyphone Decision Tree Specialization for porting multilingual Grapheme based ASR systems to new languages, ICASSP Proc., 4249–4252, 2008.

[9] Beulen, K. and Bransch, E. and Ney, H.,State-tying for context dependent phoneme models, EUROSPEECH, 3:11791182, 1997.