ICASSP 2005

ACOUSTIC TRAINING FROM HETEROGENEOUS DATA SOURCES: EXPERIMENTS IN MANDARIN CONVERSATIONAL TELEPHONE SPEECH TRANSCRIPTION

Stavros Tsakalidis [1], *William Byrne* [1,2]

Center for Language and Speech Processing, The Johns Hopkins University,
3400 N. Charles St., Baltimore, MD, 21218 U.S.A [1]

Department of Engineering, Cambridge University
Trumpington Street, Cambridge, CB2 1PZ U.K. [2]

Email: `stavros@jhu.edu, wjb31@cam.ac.uk`

ABSTRACT

In this paper we investigate the use of heterogeneous data sources for acoustic training. We describe an acoustic normalization procedure for enlarging an ASR acoustic training set with out-of-domain acoustic data. A larger in-domain training set is created by effectively transforming the out-of-domain data before incorporation in training. Baseline experimental results in Mandarin conversational telephone speech transcription show that a simple attempt to add out-of-domain data degrades performance. Preliminary experiments assess the effectiveness of the proposed cross-corpus acoustic normalization. Furthermore, we investigate the behavior of speaker adaptive training in conjunction with the cross-corpus normalization procedure.

1. INTRODUCTION

The common refrain in automatic speech recognition systems (ASR) is that when it comes to acoustic training, there's no data like more data. At the same time, data added should somehow be similar to the existing data set which in itself should be closely related to the final task to which the ASR system will be applied. The refrain should in fact be modified to state that 'there's no data like more data that's similar to the data already available.' Any number of contributing factors, such as language, dialect, acoustic channel, sampling rate, domain or topic, speaking style, speaker age and education, enter into the characterization of an acoustic training set. Typically data is available from a single source, such as a single controlled data collection effort that gathers speech from a known population under somewhat controlled circumstances. This yields a relatively homogeneous collection of speech, and models trained on such a collection will perform well when incorporated into an ASR system evaluated on new speech of a similar nature. However, if a second collection of speech differs in any of these or other dimensions, for instance if the acoustic channel varies, simply adding the second collection to the first collection to create a large acoustic training set may in fact lead to degradation in recognition performance. Loosely speaking, if the model is not able to account for the added variability the acoustic model training process will be disrupted. This paper focuses on simple acoustic normalization techniques that we show make it possible to augment an acoustic training set with speech data that would otherwise lead to performance degradation.

Our task is to build a Mandarin Conversational Telephone Speech ASR system for the CallFriend (CF) [1] domain using data not only from the closely matched CF training corpus but also with data added from two out-of-domain sources. The calls for the CF test and training corpora were domestic with both parties located in the continental United States and Canada. The second corpus, the CallHome (CH) corpus is fairly similar to the CF corpus in that in both speakers were simply taking advantage of a free phone call. However the CH corpus calls originated in the US with the other speaker(s) in locations overseas. The third collection was based entirely in China. The Flight Corpus (FL) [2] consists of telephone conversations between travel agents and customers calling to ask about flights and to make reservations. Since these three databases contain conversational Mandarin collected over the telephone, it is reasonable to investigate whether the CH and FL data can be helpful in building ASR systems for the CF domain.

Many approaches have been proposed to model unwanted variations in sampled speech and language. In the front-end, speaker and channel normalization techniques modify the spectral representation of the speech waveforms in an attempt to reduce non-informative variability between speakers and channels. As an example, cepstral mean normalization [3] is used to reduce distortion introduced by the transmission channel. In the course of training of acoustic models, inter-speaker variabilities are modelled and directly incorporated into the training process by speaker adaptive training [4, 5] and stochastic matching [6] techniques. Language modeling can also benefit from techniques to incorporate out-of-domain data [7].

We describe an acoustic normalization procedure for enlarging an ASR acoustic training set with out-of-domain acoustic data. The approach is a straightforward application of model-based acoustic normalization techniques to map the out-of-domain feature space onto the in-domain data. A larger in-domain training set is created by effectively transforming the out-of-domain data before incorporation in training. Performance will be measured by improvements on the in-domain test set.

2. CROSS-CORPUS NORMALIZATION

We start with a collection of C training sets ($c = 1, \ldots, C$), where $c = t$ denotes the in-domain data set. Assuming that the in-

This work was supported by the NSF (U.S.A.) under the Information Technology Research (ITR) program , NSF IIS Award No. 0122466.

domain training data are sufficiently representative of the type of speech expected to be recognized, our modelling technique transforms the out-of-domain feature space to match the space of the in-domain train population. The transformed acoustic feature vector o is found as $Ao + b$, where A is a nonsingular matrix and b is a vector. It is these transforms $[b\ A]$ that will be estimated over the out-of-domain training sets. Although this modeling approach is quite general and could be extended to a variety of normalization techniques and estimation criteria, we study only transform-based acoustic normalization in HMMs under the maximum likelihood (ML) estimation criterion.

The emission density of state s which is assumed to be Gaussian is reparametrized as

$$q(\zeta|s,c;\theta) = \frac{|A^{(c)}_{\mathcal{R}(s)}|}{\sqrt{(2\pi)^m|\Sigma_s|}} e^{-\frac{1}{2}(T^{(c)}_{\mathcal{R}(s)}\zeta-\mu_s)^T\Sigma_s^{-1}(T^{(c)}_{\mathcal{R}(s)}\zeta-\mu_s)}.$$

Note the dependence on c ; the observation distribution depends on the training set to which it is applied. Here $T_r^{(c)}$ denotes the extended source dependent transformation matrix $[b_r^{(c)}\ A_r^{(c)}]$ associated with states $S_r = \{s|\mathcal{R}(s) = r\}$ for classes $r = 1, \ldots, R$; ζ is the extended observation vector $[1\ o^T]^T$; and μ_s and Σ_s are the mean and variance for the observation distribution of state s. The Σ_s are constrained to be diagonal covariance matrices. We assume that the in-domain data does not need to be normalized at the corpus level, and this is in fact a key step in the modeling approach. To this end, we simply set $A_r^{(t)} = I$ and $b_r^{(t)} = \mathbf{0}\ \forall r$. The entire parameter set is specified as $\theta = (T^{(c)}_{\mathcal{R}(s)}, \mu_s, \Sigma_s)$.

Our goal is to estimate the transforms and the HMM parameters under the ML criterion. The estimation is based on the observed random process $(\hat{w}_1^{\hat{n}}, \hat{o}_1^{\hat{l}})$ that consists of an $\hat{n}$-length word sequence $\hat{w}_1^{\hat{n}}$ and an $\hat{l}$-length sequence of m-dimensional acoustic vectors $\hat{o}_1^{\hat{l}}$. To incorporate information about the source identity into the statistical framework, we modify the observed random process to include a sequence that labels each observation vector by the source that produced it: $(\hat{w}_1^{\hat{n}}, \hat{o}_1^{\hat{l}}, \hat{c}_1^{\hat{l}})$. The train objective therefore becomes the maximization of $p(\hat{o}_1^{\hat{l}}|\ \hat{w}_1^{\hat{n}}, \hat{c}_1^{\hat{l}}; \theta)$. This estimation is performed as a two-stage iterative procedure. At each iteration, we first maximize the ML criterion with respect to the affine transforms while keeping the Gaussian parameters fixed, and then reestimate the Gaussian parameters using the updated values of the normalizing transforms.

2.1. Corpus-Normalizing Transform Estimation

Maximum likelihood reestimation of the parameters is performed using the expectation-maximization (EM) [8] algorithm. This yields the following update rule to be satisfied by the parameter estimation procedures: given a parameter estimate θ, a new estimate $\bar{\theta}$ is found so as to satisfy

$$\bar{\theta} : \sum_{r,c} \sum_{s \in S_r} \sum_{\tau:\hat{c}_\tau = c} \gamma_s(\tau;\theta) \nabla_\theta \log q(T^{(c)}_{\mathcal{R}(s)}\hat{\zeta}_\tau | s, c;\ \bar{\theta}) = 0$$

where $\gamma_s(\tau;\theta) = q_{s_\tau}(s|\hat{w}_1^{\hat{n}}, \hat{o}_1^{\hat{l}}, \hat{c}_1^{\hat{l}};\theta)$ is the conditional occupancy probability of state s at time τ given the training acoustics and transcription.

A detailed derivation of the transformation parameters is contained in the work of Gales [9]. Given the updated values of the affine transforms the estimate for the mean and variance can be shown to be

$$\bar{\mu}_s = \frac{\sum_c \sum_{\tau:\hat{c}_\tau = c} \gamma_s(\tau;\tilde{\theta}) \bar{T}^{(c)}_{\mathcal{R}(s)} \hat{\zeta}_\tau}{\sum_c \sum_{\tau:\hat{c}_\tau = c} \gamma_s(\tau;\tilde{\theta})}$$

$$\bar{\Sigma}_s = \frac{\sum_c \sum_{\tau:\hat{c}_\tau = c} \gamma_s(\tau;\tilde{\theta}) \bar{T}^{(c)}_{\mathcal{R}(s)} \hat{\zeta}_\tau \hat{\zeta}_\tau^T \bar{T}^{(c)T}_{\mathcal{R}(s)}}{\sum_c \sum_{\tau:\hat{c}_\tau = c} \gamma_s(\tau;\tilde{\theta})} - \bar{\mu}_s \bar{\mu}_s^T\ .$$

Data Sources in Training				CER	
CF	CH	FL	Total Hours	SI	SI+MLLR
√			14	60.8	58.7
	√		14	62.2	59.8
		√	22	69.2	65.8
√	√		28	57.9	55.9
√		√	36	60.8	58.7
√	√	√	50	59.3	56.6

Table 1. Character Error Rate (%) of baseline systems trained from various corporus combinations as evaluated on the CF test set. Results are reported with and without unsupervised MLLR speaker adaptation.

3. EXPERIMENTAL RESULTS

The testbed used for this research was the 1 hour CallFriend development set defined by BBN [10]. As we mentioned in Section 1, the training data comes from three different Chinese corpora. These are: a 14 hour, 42-conversation CallFriend (CF) corpus; a 14 hour, 100-conversation CallHome (CH) corpus; and a 22 hour, 1790-conversation Chinese spontaneous telephone speech corpus in the flight enquiry and reservation domain (FL) [2]. Both the CF and CH collection are part of the training set defined for the EARS RT-03 evaluation.

The baseline acoustic models were built using HTK [11]. The system is a speaker independent continuous mixture density, tied state, cross-word, gender-independent, context-dependent Initial-Final (I/F), HMM system. The speech was parameterized into 39-dimensional PLP cepstral coefficients with delta and acceleration components. Cepstral mean and variance normalization was performed over each conversation side. The acoustic models used cross-word I/F with decision tree clustered states [11], where questions about phonetic context as well as word boundaries were used for clustering. Details of the ASR system design can be found in [12]. Decoding experiments were performed using the AT&T Large Vocabulary Decoder [13], using a bigram language model constructed as follows. Three bigram language models were trained over each set of transcriptions and were linearly interpolated [14] with weights chosen so as to minimize the perplexity on held-out CF transcriptions. This bigram was used for all decoding experiments.

3.1. Unnormalized Out-of-Domain Acoustic Data

Initial baseline experiments were performed to measure the performance of models trained using each of the three training sources. Various training sets were creating through combinations of the sources without cross-corpus normalization. Table 1 summarizes the performance of ASR systems estimated over these training

Data Sources & Normalization				CER	
CF	CH	FL	#transforms	SI	SI+MLLR
I	T		1 per corpus	57.6	55.8
I	I	T	1 per corpus	58.1	55.7
I	T	T	1 per corpus	57.8	55.5

Table 2. Character Error Rate (%) of systems by normalizing out-of-domain acoustic training data relative to in-domain data. An ‘T’ / ‘I’ indicates that a source was included in training with / without normalization, resp. Results are reported with and without unsupervised MLLR speaker adaptation.

Data Sources & Normalization				CER	
CF	CH	FL	#transforms	SAT	SAT+MLLR
I	I	I		59.4	55.6
I	T	T	1 per corpus	58.0	54.6

Table 3. Character Error Rate (%) of SAT derived systems from unnormalized and normalized out-of-domain acoustic training data relative to in-domain data. An ‘T’ / ‘I’ indicates that a source was included in speaker adaptive training with / without cross-corpus normalization, resp. Results are reported with and without unsupervised MLLR speaker adaptation.

sets. Results on the CF test set, both with and without unsupervised MLLR speaker adaptation [15] are given.

We first conducted baseline experiments to quantify the mismatch between each of the three training corpora and the test corpus in terms of recognition performance. Not surprisingly, acoustic models trained only with CF gave the best performance on the CF test set (CER 60.8%/58.7%). The CH trained acoustic models had poorer but comparable performance (CER 62.2%/59.8%). On the other hand, the FL-based acoustic models were significantly worse than either CF or CH models (CER 69.2%/65.8%).

We then investigated the combination of each of the out-of-domain corpora and the in-domain corpus in training acoustic models for the CF task. The acoustic models were obtained by pooling the in-domain training data with each out-of-domain training data and estimating the HMM parameters in the standard ML fashion, i.e. without the cross-corpus normalizing transforms. It was found that a simple merging of the CF and CH data yielded an improvement relative to using either corpus alone. However, adding the FL set to the CF data gave absolutely no improvement relative to using the CF data alone. Moreover, adding the FL set to the CF and CH sets degrades performance relative to training with CF and CH alone.

The results of this section show that the performance of acoustic models trained from a combination of in-domain and out-of-domain data depends on the similarity of each training set to the test set. Simply adding out-of-domain data can actually degrade performance.

3.2. Normalized Out-of-Domain Acoustic Data

We then conducted a series of experiments to assess the effectiveness of cross-corpus acoustic normalization as proposed in Section 2. This procedure does need a starting point from which the initial set of transforms can be estimated. All normalization experiments are seeded by the CF+CH system of Section 3.1, which was trained over the combined CF and CH training sets and was best of the unnormalized systems (CER 57.9%/55.9%). A single transform was estimated for each out-of-domain corpus in these preliminary normalization experiments. The cross-corpus normalization experiments are reported in Table 2.

We first investigated the combination of the CF and CH corpora. Applying the cross-corpus normalizing transform to the CH data gave a modest 0.3% improvement relative to the unnormalized CF+CH system when no MLLR speaker adaptation was used during decoding. However, this improvement effectively diminishes with the presence of speaker adaptation on the test side.

We then added the FL set to the CF and CH sets. We initially treated the CF and CH corpora as in-domain data sources and the FL corpus as out-of-domain source. Under this scenario, only the FL data was transformed. The normalization of the FL corpus gave an improvement (CER 58.1%/55.7%) relative to the unnormalized CF+CH+FL system. Then, we applied the cross-corpus normalization to both the CH and FL corpora. Normalizing both out-of-domain data sources yielded a slightly better result (CER 57.8%/55.5%) relative to normalizing the FL corpus alone. In conclusion, the cross-corpus normalization makes it possible to improve performance by adding a severely mismatched corpus.

3.3. Speaker Adaptive Training on Normalized Out-of-Domain Acoustic Data

A commonly used approach for improving ASR performance is speaker adaptive training (SAT) [4] in which speaker dependent transforms are used to reduce speaker-specific variability in the speech signal. Our training set is a collection of heterogeneous corpora, and we investigate whether cross-corpus normalization procedures can be used jointly with speaker adaptive training to improve recognition performance.

Table 3 compares the performance of SAT acoustic models trained over unnormalized acoustic data to SAT acoustic models trained over an in-domain training set created by transforming the out-of-domain corpora prior to speaker adaptive training. Throughout these SAT experiments we used a fixed set of two regression classes for the speaker depended transforms- one class for speech states and one class for silence states. The first SAT system was seeded by the unnormalized CF+CH+FL system of Section 3.1 (CER 59.3%/56.6%) and subsequently trained over the unnormalized CF, CH and CH training sets. The second SAT system was seeded by the models of Section 3.2 (CER 57.8%/55.5%) which were trained over the in-domain CF data and the normalized out-of-domain CH and CF training sets. SAT training was performed as usual, but over the cross-domain normalized data.

In the following we focus on the recognition performance incorporating unsupervised speaker adaptation over the test set. Applying SAT in the standard fashion, i.e. without cross-corpus normalization, yields 1.0% absolute gain over the unnormalized CF+CH+FL system (CER 55.6% vs. 56.6% - see Table 1). This is comparable to the gains from cross-corpus normalization alone: in Section 3.2 we found that applying the cross-corpus normalizing transforms to both the out-of-domain corpora gave 1.1% absolute gain over the same unnormalized CF+CH+FL system (CER 55.5% vs. 56.6%). Results in Table 3 show that SAT can be further improved by 1.0% (CER 54.6% vs. 55.6%) if we first compensate for the cross-corpus differences across the training sets. When we consider the combined gains from SAT and cross-corpus normalization against the ML baseline system, (CER 54.6% vs. 56.6%)

the total gain is 2.0%, an indication that the cross-corpus normalization and SAT procedures yield additive improvement, and are thus capturing complementary influences, as desired.

4. DISCUSSION

In this paper we investigated the use of heterogeneous data sources for acoustic training. Baseline experimental results showed that simply adding out-of-domain data actually degraded performance. The proposed acoustic normalization procedure made it possible to enlarge the acoustic training set with out-of-domain acoustic data that would otherwise lead to performance degradation. A larger in-domain training set was created by transforming the out-of-domain feature space to match the in-domain training population.

We have also found that cross-domain normalization can also improve Speaker Adaptive Training. Experimental results show that performing SAT over cross-corpus normalized data effectively doubles the gains obtained from SAT alone on this corpus. Interestingly, the gains from SAT and cross-corpus normalization are almost exactly additive, which is strong evidence that they are capturing different phenomena. In this we emphasize that we are not in fact proposing new modeling algorithms. What we have shown is that careful initialization and application of existing transform-based modeling techniques can be used to capture different effects in heterogeneous data.

In our preliminary experiments we studied the use of single cross-corpus transforms. We are planning to extend this idea by estimating multiple transforms for each out-of-domain corpus based on broad speech classes. Furthermore since a training corpus usually consists of a coarse group of speakers, we indent to extend the acoustic normalization technique to homogeneous clusters of speakers, or even to each speaker separately.

This framework can be applied in certain multilingual ASR systems [16, 17, 18], where the task is to train models for a 'target' language using data from a number of 'source' languages. Currently, most of these training approaches are based on multilingual speech data pooling followed by acoustic model adaptation to fit the characteristics of a target language [17]. However, as in the experiments described earlier in Section 3.1, it often happens that simply combining data from multiple languages actually hurts ASR performance in a single language. There are many interesting modeling issues involved in sharing speech across languages [19, 20], such as the varying effect of phonetic context, the presence or absence of phones, and other issues such as the role of prosodic features such as pitch and pause duration. However, none of those detailed issues can be studied unless it is possible to work with multiple data sources without degrading the baseline performance. This work is meant to be a basis to enable further studies of the more subtle issues in combining multiple data sources.

5. ACKNOWLEDGMENT

Thanks to Prof. F. Zheng of Tsinghua Univ. for the use of the Chinese spontaneous telephone speech corpus in the flight enquiry and reservation domain.

6. REFERENCES

[1] RT-03 Spring Evaluation, 2003, [Online]. Available: http://www.nist.gov/speech/tests/rt/rt2003/spring/.

[2] T. F. Zheng, P. Yan, H. Sun, M. Xu, and W. Wu, "Collection of a chinese spontaneous telephone speech corpus and proposal of robust rules for robust natural language parsing," in *SNLP-O-COCOSDA*, May 2002, pp. 60–67.

[3] A. Acero, "Acoustical and enviromental robustness in automatic speech recognition," Kluwer Academic Publishers, Boston, MA, 1993.

[4] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *ICSLP*, Oct. 1996, pp. 1137–1140.

[5] H. Jiang and L. Deng, "A robust compensation strategy for extraneous acoustic variations in spontaneous speech recognition," *IEEE Trans. Spch. & Aud. Proc.*, vol. 10, no. 1, pp. 9–17, Jan. 2002.

[6] Ananth Sankar and Chin-Hui Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. Spch. & Aud. Proc.*, vol. 4, no. 3, pp. 190–202, May 1996.

[7] R. Iyer, M. Ostendorf, and H. Gish, "Using out-of-domain data to improve in-domain language models," *IEEE Sig. Proc. Let.*, vol. 4, no. 8, pp. 221–223, Aug. 1997.

[8] A. P. Dempster, A. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data," *J. Roy. Stat. Soc., Ser. B*, vol. 39, no. 1, pp. 1–38, 1977.

[9] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Comp. Spch. & Lang.*, vol. 12, no. 2, pp. 75–98, Apr. 1998.

[10] RT-03 Spring Workshop, May 2003, [Online]. Available: http://www.nist.gov/speech/tests/rt/rt2003/spring/presentations/cts-combined-sm-ok-v14.pdf.

[11] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book, Version 3.0*, July 2000.

[12] F. Zheng, Z. Song, P. Fung, and W. Byrne, "Mandarin pronunciation modeling based on the CASS corpus," *J. Comp. Sci. Tech. (Science Press, Beijing, China)*, vol. 17, no. 3, May 2002.

[13] M. Mohri and M. Riley, "Integrated context-dependent networks in very large vocabulary speech recognition," in *Eurospeech*, Sept. 1999, pp. 811–814.

[14] A. Stolcke, "SRILM - An Extensible Language Modeling Toolkit," in *ICSLP*, Sept. 2002, pp. 901–904.

[15] C. J. Leggetter and P. C. Woodland, "Speaker adaptation of continuous density HMMs using multivariate linear regression," *ICSLP*, pp. 451–454, Sept. 1994.

[16] W. Byrne et. al., "Towards language independent acoustic modeling," in *ICASSP*. IEEE, June 2000, pp. 1029–1032.

[17] C. Nieuwoudt and E. C. Botha, "Cross-language use of acoustic information for automatic speech recognition," *Spch. Comm.*, vol. 38, no. 1, pp. 101–113, Sept. 2002.

[18] Z. Wang, T. Schultz, and A. Waibel, "Comparison of acoustic model adaptation techniques on non-native speech," in *ICASSP*. IEEE, Apr. 2003, pp. 540–543.

[19] Special Session, "Multilinguality in Speech Processing," in *ICASSP*. IEEE, May 2004.

[20] D. Vergyri, S. Tsakalidis, and W. Byrne, "Minimum Risk Acoustic Clustering for Multilingual Acoustic Model Combination," in *ICSLP*, Oct. 2000, pp. 873–876.