State-Level Data Borrowing for Low-Resource Speech Recognition based on Subspace GMMs

Yanmin Qian[1], *Daniel Povey*[2], *Jia Liu*[1]

[1] Tsinghua National Laboratory for Information Science and Technology,
Department of Electronic Engineering, Tsinghua University, Beijing, China
[2]Microsoft Research, Redmond, WA, USA

qianym07@mails.tsinghua.edu.cn, dpovey@microsoft.com, liuj@mail.tsinghua.edu.cn

Abstract

Large vocabulary continuous speech recognition is always a difficult task, and it is particularly so for low-resource languages. The scenario we focus on here is having only 1 hour of acoustic training data in the “target” language. This paper presents work on a data borrowing strategy combined with the recently proposed Subspace Gaussian Mixture Model (SGMM). We developed data borrowing strategies based on two approaches: one based on minimizing K-L Divergence, and one that also takes into account state occupation counts. We demonstrate improvements versus the baseline SGMM setup, which itself is better than a conventional HMM-GMM system. The SGMMs are more robustly estimated by borrowing data from the non-target language at the acoustic-state level. Although we tested the approach for SGMMs, we expect the general idea of borrowing data from a non-target language to be applicable for conventional GMMs as well.

Index Terms: speech recognition, low-resource language, subspace gaussian mixture model

1. Introduction

Speech is the most convenient medium for human-to-human communication and should in principle also be convenient for human-to-machine interaction. The performance of speech processing systems has improved dramatically, but state-of-the-art systems require for their training a large amount of language-specific transcribed speech data. However, demand exists for speech recognition systems in languages that have only limited available training data; quickly developing ASR systems for resource-insufficient domains or languages is a research topic that has recently attracted interest [1][2].

Several strategies have been previously proposed. Developing a multilingual speech recognition system is a popular approach to deal with the low-resource problem [3][4]. In these systems a universal phone set is obtained based on the principle that the speech units with similar sounds across different languages are grouped together and represented by a single phonetic symbol. The International Phonetic Alphabet (IPA) or data-driven based phone clustering methods have been used to obtain universal phoneme units. After collecting a large set of speech data covering all speech units, a “universal” set of acoustic models can be trained, so an ASR system can be built even for languages with little or no training data. However the universal phone set is not as accurate as the language-specific one, and phone clustering induces more confusion among models, so the performance of these systems is not very promising.

Another way to deal with this problem is so-called automatic speech attribute transcription (ASAT), which has been proposed for developing multilingual or low-resourced ASR systems [5][6]. This tries to deal with the problem that it is hard with a limited set of training languages to get complete coverage of a universal phone set such as the IPA. Articulatory features are a solution to this problem, because all the phones can be modeled by a small number of articulatory attributes and most of the attributes, such as voicing, nasality, and friction, can be identified in any particular language. Most of the research in this area has up till now been focused on phone-level rather than word-level transcription.

The Subspace Gaussian mixture model (SGMM) is a recently proposed acoustic model that is especially suited for low-resource applications [7][8]. The majority of the trainable parameters of an SGMM are, in typical configurations, globally shared and not specific to any individual acoustic state; the only parameters specific to acoustic states are some relatively low-dimensional (e.g. 40-dimensional) vectors that represent fewer parameters than a typical GMM-based system. Therefore, when training SGMMs we can borrow other languages' data for model training without sharing the acoustic states, and obtain more robust estimates of the globally shared parameters [9].

This paper reports experiments with SGMMs, but addresses the question of whether it is helpful, in addition to sharing the global parameters, to merge some of the acoustic states across languages. This is an idea that would be equally applicable to conventional models such as GMMs. Our experimental setup is similar to [7]: we have limited amounts of training data in English, Spanish and German to imitate the low-resource situation. We were able to show statistically significant improvements versus the previously described SGMM system, which is substantially better than a GMM-based system.

The remainder of this paper is organized as follows: In Section 2, we describe the SGMM and then describe our new data borrowing strategy in detail. In Section 3, we describe our experimental setup and present experimental results. We summarize and give conclusions in Section 4.

2. Data borrowing strategy

2.1. Subspace Gaussian Mixture Model

The most basic form of the Subspace Gaussian Mixture Model (SGMM) can be expressed in the following three equations:

$$p(x \mid j)=\sum_{i=1}^{I} \omega_{ji} N(x; \mu_{ji}, \sum{}_{i}) \tag{1}$$

$$\mu_{ji} = M_i v_j \tag{2}$$

$$\omega_{ji} = (\exp w_i^T v_j) / (\sum\nolimits_{i'=1}^{I} \exp w_{i'}^T v_j) \tag{3}$$

where $p(x \mid j)$ is the distribution of features in HMM state j . The model is a mixture of Gaussians, but unlike the conventional GMM, the number of mixture components I is the same for all states and is typically quite large, e.g. several hundred. The covariance $\sum_i$ for each Gaussian in the mixture is globally shared across states (we use full covariances). The most important difference is that the mean μ_{ji} and mixture weights ω_{ji} are not direct parameters of the model, and instead they are expanded from a state-specific vector v_j, via globally shared parameters M_i and w_i, as illustrated in Equations (2) and (3).

This model has a more complicated structure than a GMM; however a well-tuned SGMM typically has fewer parameters than the well-tuned GMM system [7]. Moreover, the majority of the parameter count in a SGMM system consists of shared parameters M_i, $\sum_i$ and w_i, which for well-tuned systems trained on small amounts of data can be 8~10 times larger than the state-specific parameters v_j. This leads to a natural method of training SGMMs in a multilingual way: the state-specific SGMM parameters are trained as separate language-specific states, and the common SGMM parameters are, however, shared across languages. This can be thought of as a single system covering multiple languages, in which the phones from distinct languages are given distinct names.

We mention at this point that the SGMMs we use are a slight extension of the simplified version described above: we introduce sub-states, where each state j has M_j sub-states each with its own mixture weight c_{jm} and vector v_{jm}. The extended equations with sub-states are given in [7].

It was previously shown [9] that training the globally shared parameters across languages can lead to substantial improvements if the amount of training data in the target language is limited. Our work here builds on that previous work, and attempts to address the question of whether in addition to sharing the global parameters, it might be advantageous to also share some of the speech states. This is a similar idea to sharing phones across languages (which we also explore in our experiments), but is more fine-grained. We emphasize that although we do the experiments in the SGMM framework, the idea of sharing states across languages is equally applicable for normal GMM models, although the details would be different.

In the multilingual SGMM framework introduced in [9], the target-language and non-target language models are trained at the same time (the statistics for updating the shared parameters are shared across languages). At a point towards the start of training, we decide for each target-language HMM-state whether or not it should be shared, and if so select some non-target-language HMM-state to share it with. We explore two techniques: one based on an approximated K-L divergence (Section 2.2), and one that uses K-L divergence but also takes into account occupation counts (Section 2.3).

2.2. Minimum K-L divergence principle

The first of our two methods uses an approximated K-L divergence between SGMM states. The basic method is: for each state in the target language, find the "closest" state in some non-target language, and if this falls below some threshold, share with that state; otherwise leave the state in question unshared. The distributions in SGMM states are just a special case of Gaussian Mixture Models (GMMs), and exactly computing the K-L divergence between GMMs is quite hard (e.g. see [10]). However, because there is a correspondence between the states of the models (i.e. the index *i* is shared), by making the assumption that the Gaussians are "far apart" and there is insignificant overlap in distribution between differently numbered Gaussians, we can obtain a convenient closed-form expression for the K-L divergence. We additionally make the approximation that the Gaussian priors are the same across states, i.e. we use a global rather than state-specific Gaussian prior. This is quite crude, but our main aim was to obtain a distance measure that makes sense; we do not believe that the algorithm should be particularly sensitive to the exact distance measure used as long as it is reasonable.

The distance measure between SGMM states j and k is defined as follows:

$$Dis(j,k)=\sum_{i=1}^{I}p(i)(v_j-v_k)^T M_i^T \textstyle\sum_i^{-1} M_i(v_j-v_k) \quad (4)$$

where $p(i)$ is the prior on the Gaussian index i, M_i and $\sum_i$ are the shared projection matrix and Gaussian covariance, v_j and v_k are the state-specific parameters of state j and k. The prior of Gaussian i can be defined as:

$$p(i)=\sum\nolimits_{t=1}^{T}\sum\nolimits_{j=1}^{J}\gamma_i^j(t) \quad (5)$$

where $\gamma_i^j(t)$ is the occupation probabilities per-Gaussian and per-state as defined in the standard forward-backward or Viterbi algorithm [7]. Note that it is possible to simplify (4) into an inner product of the difference between the two vectors, with a particular matrix, so this distance measure is the same as the Euclidean distance in an appropriately pre-scaled space. The threshold ***e*** controls the amount of shared states, i.e. borrowed data. We evaluate this criterion and tie states before introducing sub-states, in order to avoid complications arising from sub-states.

The following algorithm summarizes the state-tying procedure; the threshold ***e*** controls the amount of state tying that takes place. The overall training schedule is as described in [9]; we applied this algorithm on the second "epoch" of training as defined in [9] (an epoch corresponds to eight passes over the data).

Algorithm 2.1 Data Borrowing with Minimum K-L Divergence Principle on SGMM

```
Follow the SGMM multilingual training schedule, and
finish the second epoch of the normal SGMM training
  for each state j in the target language do
    for each state k in the non-target languages do
      calculate the KL divergence $Dis(j,k)$
    end for
    select the relative minimum KL divergence
    if the minimum KL divergence is smaller than threshold e
    then share the target state j with the non-target state k
    else leave the state j unchanged
    end if
  end for
```

2.3. State occupation principle

We also tried a second approach that makes use of the state occupation counts. The basic intuition is that if there is a large amount of data available to train a particular target-language state, there is no need to share it with a non-target language because the only point of this procedure is to overcome data sparsity. The way we apply this intuition is to select a count cutoff ε, and to treat target-language states with counts above and below this value differently, in that we apply two different distance thresholds: a large threshold ***e2*** for states with "small" counts and a smaller threshold ***e1*** for states with "large counts". Let the state occupation counts be γ^{j}.

The modified algorithm is as follows:

Algorithm 2.2 Refined Data Borrowing with State Occupation Principle on SGMM

```
Follow the SGMM multilingual training schedule, and finish the second epoch of the normal SGMM training
  for each state j in the target language do
    for each state k in the non-target languages do
      calculate the KL divergence $Dis(j,k)$
    end for
    select the relative minimum KL divergence
    if the minimum KL divergence is smaller than KL threshold e1
    then
      share the target state j with the non-target state k
    else if the minimum KL divergence is smaller than KL threshold e2 (e2 > e1), and the state occupation of state j is lower than occupation threshold $\varepsilon$
    then
      share the target state j with the non-target state k
    else
      leave the state j unchanged
    end if
  end for
```

3. Experiments and Results

3.1. Experimental data and Baseline system

Our experiments are on the Callhome English, German and Spanish databases [11], and are based on those in [9]. The conversational nature of speech in Callhome database along with high out-of-vocabulary rates, use of foreign words and telephone channel distortions make the task of speech recognition on this database challenging.

The database contains 80 spontaneous telephone conversations in each of English, German and Spanish, with about 15 hours of speech per language to be used as training data. To imitate the low-resource application, we select the English as the target language and use 1 hour of randomly chosen speech from the English corpus as the target-language training data. Besides this, we use the entire 15 hours of German and 16 hours of Spanish training data. The 20 conversations of the English evaluation set, roughly containing 1.8 hours of speech, form our test set.

The features are 39-dimensional and based on PLP features with energy, first and second order deltas, plus per-speaker mean and variance normalization. We use a 42-phone set for English, 46 for German and 28 for Spanish. We use the 1 hour English data to train a baseline HMM-GMM system with only 550 states and 4 Gaussians per state, and we pool all the three languages' data to train a multilingual HMM-SGMM system which has 1500, 1771 and 1623 tied states for English, German and Spanish respectively. The number of tied states was tuned separately for the GMM baseline and the multilingual setup. The number of Gaussian components I is 400, and dimension of the state-specific vector v_j is 40. We used the SRILM tools [12] to build a language model which is a trigram with a word-list of 62K words obtained by interpolating individual models trained from English Callhome corpus, the Switchboard corpus [13] and the Gigaword corpus [14]. We use the HDecode and Kaldi decoders to decode the GMM or SGMM model respectively, and score the results with the NIST scoring scripts.

The first two lines of Table 1 summarize the HMM-GMM baseline and HMM-SGMM baseline results for our experiments. It is clear that the multilingual SGMM approach gives substantial improvement (more than 10% absolute). This is the approach previously reported in [9].

3.2. Data borrowing at the SGMM state-level

We first construct the initial multilingual SGMM model, and finish the first two epochs of the normal SGMM training. Then we apply the Minimum KL Divergence Principle of Algorithm 2.1 to calculate the distances between the states of target language and the states of borrowed languages, German and Spanish. We vary the threshold to control the quantity of ultimately shared states. The number of states coming from German versus Spanish as we vary the threshold e is illustrated in Fig. 1; German contributes more states than Spanish, as one would expect from its closer linguistic relationship to English. The difference between German and Spanish in this regard suggests that for this technique to work it may be important to select non-target languages that are linguistically close to the target language.

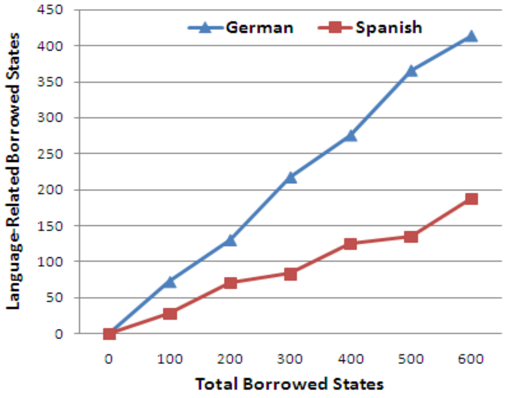

Figure 1: *Distribution by language of borrowed states using Minimum KL Divergence Principle.*

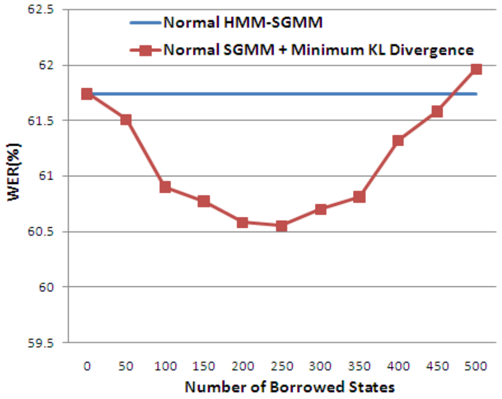

Figure 2: *WER as number of borrowed states is increased, using Minimum KL Divergence Principle.*

When we decide the final shared states between the target and non-target languages, we edit the HMM-state level training transcription obtained by Finite State Acceptors [15] and pool the real target data and borrowed data to train these shared states. Then we finish the later epochs of SGMM training including updating individual parameters and splitting sub-states as normal. We varied the number of shared states to investigate the performance of the SGMM system using Minimum KL Divergence Principle.

Fig.2 shows how the WER changes as we vary the threshold ***e*** (we plot WER against the number of borrowed states). The WER initially decreases, but then increases again if we combine states too aggressively. We get the best performance when about 20% of target-language states are shared.

Lines 3 and 4 of Table 1 compare two different methods of sharing states: the Minimum KL Divergence Principle (Algorithm 2.1) and the State Occupation Principle (Algorithm 2.2). We get about 1.2% absolute WER improvement from the Minimum KL Divergence principle, and 1.7% with the State Occupation Principle, which validates our intuition that it makes more sense to share target-language states with small data counts. In each case we tuned the number of shared states to minimize WER, which resulted in about 250 and 200 states shared respectively. We applied the matched-pairs significance test described in [16], and in either case the improvement versus the SGMM baseline was statistically significant[1] at the chosen confidence level of 99.5%.

Table 1. *Performance comparison of different systems using only 1 hour of target language data*

System description	WER
1. Conventional HMM-GMM	72.57%
2. SGMM	61.74%
3. SGMM + data borrowing Algorithm 2.1	60.55%
4. SGMM + data borrowing Algorithm 2.2	60.02%
5.SGMM + sharing states within target language	61.88%
6.SGMM + data borrowing on the phone level from non-target languages	62.59%

In order to verify that the effect we were seeing was a genuinely “cross-language” effect and not simply a gain from post-clustering the states obtained by HTK’s state clustering procedure, we did the experiment in line 5 where we applied Algorithm 2.1 to share states, but only within the target language and not across languages. This degraded performance, which confirms that the improvements we were seeing were not explainable in this way.

We also attempted to use a more conventional method based on tying phones across languages; this experiment is in line 6. We used the State Time Alignment (STA) algorithm described in [17], but replacing the Bhattacharya distance with the K-L divergence for consistency with the current experiments; we tuned it to share 20% of the target-language phones in order to be comparable to our state-tying experiments. This results in a degradation. The conclusion we draw from this is that data sharing across languages can be helpful, but phones are a too-coarse level at which to tie, and it is better to tie the context-dependent states.

4. Conclusions

In this paper, we present our work on a data borrowing strategy for low-resource speech application. We performed experiments based on the recently proposed Subspace Gaussian Mixture Model (SGMM). The SGMM has an inherent mechanism of tying data across languages because it has a large number of globally shared (not language-specific) parameters, but we wanted to investigate whether, in addition to this mechanism, we could apply a method based on cross-language state tying to further improve results. We looked at the scenario where we have 1 hour of in-language data together with a larger amount of out-of-language data. We were able to get improvements from state tying of about 1.7% absolute. While this is smaller than the original improvements of the SGMM over the HMM-GMM baseline, it is still substantial. We showed that it is possible to improve results by tying states across languages, and our results seem to indicate that it is important to select linguistically close languages to do this tying, and that it is important to tie parameters at the context-dependent state level rather than at the phone level.

5. Acknowledgements

This work was supported by the National High Technology Research and Development Program of China (Project 2008AA040201), the Project 90920302 and 61005019 of NSFC (National Natural Science Foundation of China), and the Project 60931160443 of NSFC and RGC.

6. References

[1] P. Fung and T. Schultz, "Multilingual Spoken Language Processing", IEEE Signal Processing Magazine, vol. 25, pp:89-97, 2008.

[2] Xiaodong Cui, Jian Xue, et al., "Acoustic Modeling with Bootstrap and Restructuring for Low-Resourced Languages", in Proc. Of INTERSPEECH, pp:2974-2977, 2010.

[3] B. D. Walker, B. C. Lackey, J. S. Muller, and P. J. Schone, "Language-Reconfigurable Universal Phone Recognition", in Proc. Of EUROSPEECH, 2003.

[4] Hui Lin, Li Deng, et al., "A Study on Multilingual Acoustic Modeling for Large Vocabulary ASR", in Proc. Of ICASSP, pp:4333-4336, 2009.

[5] Sabato Macro Siniscalchi, Torbjorn Svendsen, and Chin-Hui Lee, "Toward bottom-up continuous phone recognition", in Proc. Of ASRU, 2007.

[6] Sabato Macro Siniscalchi, Torbjorn Svendsen, and Chin-Hui Lee, "Toward A Detector-Based Universal Phone Recognizer", in Proc. Of ICASSP, pp:4261-4264, 2008.

[7] D. Povey, Lukas Burget, et al., "The Subspace Gaussian Mixture Model-A Structured Model for Speech Recognition", Computer Speech and Language, vol. 25, Issue 2, pp:404-439, 2011.

[8] D. Povey, Lukas Burget, et al., "Subspace Gaussian Mixture Models for Speech Recognition", in Proc. Of ICASSP, pp:4330-4333, 2010.

[9] Lukas Burget, Petr Schwartz, et al., "Multilingual Acoustic Modeling for Speech Recognition based on Subspace Gaussian Mixture Models", in Proc. Of ICASSP, pp:4334-4337, 2010.

[10] J-Y. Chen, J.R. Hershey, P.A. Olsen and E. Yashchin, “Accelerated Monte Carlo for Kullback-Leibler divergence between Gaussian mixture models”, in Proc. Of ICASSP, pp:4553-4556, 2008.

[11] A. Canavan, D. Graff, and G. Zipperlen, “CALLHOME American English/German/Spanish Speech”, Linguistic Data Consortium, 1997.

[12] A. Stolcke, "SRILM – An Extensible Language Modeling Toolkit", in Proc. Of ICSLP,pp:901-904, 2002.

[13] J.J. Godfrey el at., “Switchboard: Telephone speech corpus for research and development”, in Proc. Of ICASSP, 1992.

[14] D. Graff, “English Gigaword”, Linguistic Data Consortium, 2003.

[15] Mehryar Mohri, Fernando Pereira, and Michael Riley, "Weight finite-state transducers in speech recognition", Computer Speech and Language, vol. 20, Issue 1, pp:69-88, 2002.

[16] L. Gillick and S.J. Cox, "Some Statistical Issues in the Comparison of Speech Recognition Algorithms", in Proc. Of ICASSP,pp:532-535, 1989.

[17] Yanmin Qian and Jia Liu, “Phone Modeling and Combining Discriminative Training for Mandarin-English Bilingual Speech Recognition”, in Proc. Of ICASSP, pp:4918-4921, 2010.

[1] This significance test is not completely valid since we tuned the number of states borrowed on the test set, but it at least shows that the improvements are of a magnitude that is potentially significant