Robust triphone mapping for acoustic modeling

Miloš Cerňak[1,2]*, David Imseng*[1,3]*, Hervé Bourlard*[1,3]

[1] Idiap Research Institute, Martigny, Switzerland
[2] Institute of Informatics, Slovak Academy of Sciences, Bratislava, Slovakia
[3]Ecole Polytechnique Fédérale, Lausanne (EPFL), Switzerland
{milos.cernak, david.imseng, herve.bourlard}@idiap.ch

Abstract

In this paper we revisit the recently proposed triphone mapping as an alternative to decision tree state clustering. We generalize triphone mapping to Kullback-Leibler based hidden Markov models for acoustic modeling and propose a modified training procedure for the Gaussian mixture model based acoustic modeling. We compare the triphone mapping to decision tree state clustering on the Wall Street Journal task as well as in the context of an under-resourced language by using Greek data from the SpeechDat(II) corpus. Experiments reveal that triphone mapping has the best overall performance and is robust against varying the acoustic modeling technique as well as variable amounts of training data.

Index Terms: Speech recognition, acoustic modeling, triphone mapping, Kullback-Leibler divergence.

1. Introduction

State-of-the-art Automatic Speech Recognition (ASR) systems typically employ context dependent modeling in order to better take into account the canonical-to-surface form variability of pronunciation inherent to acoustic modeling. Such context dependent modeling most commonly takes the form of the triphone whose representation comprises a phone along with its preceding and following phone context. In creating triphone context models we immediately run into the problem of sparsity of the training data, since many triphone contexts will occur infrequently or not at all.

To overcome this, the decision tree state clustering (DTSC) approach [1] was introduced in which states of context dependent models are tied (thereby sharing data) according to shared properties and based on a greedy algorithm. DTSC also permits the synthesis of contexts that were unseen in the training data. However, DTSC needs a set of appropriate questions in order to develop a tree. Often this is expensive since the question are usually manually determined.

Recently, triphone mapping (TM) [2] was presented as an alternative to DTSC. TM is a data driven technique to map rare triphones to frequent ones and does not require manually determined questions. The mapping is based on context independent mono-phone models, but not limited to single Gaussian models. It was shown that TM outperforms DTSC when systems use the same number of parameters and 4k or more HMM states [2].

In this paper, we first revisit TM for Gaussian mixture based acoustic modeling and propose a modified training procedure that successively partitions the acoustic space. Furthermore, we show that it is easily generalizable to the recently proposed Kullback-Leibler divergence based hidden Markov models (KL-HMMs). A KL-HMM is an acoustic modeling technique for ASR that uses a Kullback-Leibler divergence based cost function and is very powerful if only small amounts of training data are available [3, 4]. Altogether, we show that TM can be generalized across different acoustic modeling techniques and has the best overall performance when compared to DTSC on two different databases and on different amounts of training data.

The paper is organized as follows: Section 2 reviews TM and Section 3 describes the two different acoustic modeling techniques along with the acoustic distances used for TM. Section 4 describes the experimental setup. The results are presented in Section 5 before Section 6 concludes the paper.

2. Triphone mapping

The basic concept of TM consists of creating one triphone map per phoneme that maps the set of all triphones T_a (having the same center phoneme) using only acoustic information in the form of monophone models to a subset of selected triphones T_s. The subset of selected triphones T_s is determined by applying a simple occurrence threshold λ, i.e. a triphone $t \in T_s$ if it appears at least λ times in the training data. In this study, we use the same threshold for all phonemes.

The mapping function Ω then maps the set of all triphones T_a to the subset of selected triphones T_s:

$$\begin{array}{rcl} \Omega : T_a & \rightarrow & T_s \\ t_a & \mapsto & \underset{t_s \in T_s}{\operatorname{argmin}}(TD(t_a, t_s)) \end{array} \quad (1)$$

where TD is an acoustic triphone distance, $t_a \in T_a$ and $t_s \in T_s$, i.e. each t_a is mapped to its closest t_s.

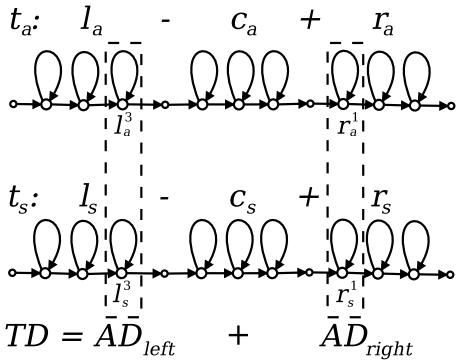

Figure 1: *Acoustic triphone distance $TD(t_a, t_s)$ calculation based on acoustic left and right contextual distances, keeping the same center phonemes $c_a = c_s$.*

We assume that each monophone is modeled with three states as shown in Figure 1. The acoustic triphone distance TD can then be expressed as:

$$\begin{aligned} TD(t_a, t_s) &= AD_{left} + AD_{right} \\ AD_{left} &= AD(l_a^3, l_s^3) \\ AD_{right} &= AD(r_a^1, r_s^1) \end{aligned} \quad (2)$$

where l and r stand for monophones of the left and right context respectively. Hence, AD_{left} is the acoustic distance between the third state of the left context monophones and AD_{right} is the acoustic distance between the first states of the right context monophones. Note that TD is not limited to three-state models. It might be generalized for monophone models with more than three states.

In this study, we measure the acoustic distance with the Kullback Leibler (KL) divergence[1] between the probability distributions F and G associated with the involved states. We use the symetric KL-divergence for the acoustic distance calculation of Eq. (2), as it is a measure between two monophones, where the two arguments are considered interchangeable. For discrete random variables, the KL-divergence is defined as:

$$D(F, G) \stackrel{def}{=} \sum_i [F(i) - G(i)] \log \frac{F(i)}{G(i)} \quad (3)$$

and for continuous random variables, the KL-divergence is defined as:

$$D(F, G) \stackrel{def}{=} \int [f(x) - g(x)] \log \frac{f(x)}{g(x)} \mathrm{dx} \quad (4)$$

where $f(x)$ and $g(x)$ are the probability density functions (pdfs) of F and G respectively.

Originally, it was proposed to partition the acoustic space directly from monophones to triphones [2]. We hypothesize that a successive partition of the acoustic space (SPAS) can be beneficial for acoustic modeling. Therefore, we propose to first model biphones before triphones are modeled. More specifically, we determine T_s and the mapping function Ω. Then

1. Based on T_s, we determined a set of selected biphones B_s by dropping the right context, i.e. given a selected triphone $l - c + r$, the corresponding selected biphones is $l - c$.
2. Instead of initializing a triphone $l - c + r$ with the center model c, we first trained a model for the biphone $l - c$ with c as seed.
3. Then, we initialized the triphone $l - c + r$ with the biphone $l - c$ and train it.

3. Acoustic modeling

We considered two different probability distributions for the acoustic modeling:

1. Mixtures of N Gaussian distributions (GMM) with probability distribution of state s, F_s:

$$F_s = \sum_{a=1}^{N} \pi_a \mathcal{N}(x; \mu_a; \sigma_a) \quad (5)$$

where $\mathcal{N}(x; \mu_a; \sigma_a)$ stands for a Gaussian distribution with mean μ_a and variance σ_a and π_a is the weight of the a^{th} Gaussian. Hence we can write the associated pdf f_s as:

$$f_s(x) = \sum_{a=1}^{N} \pi_a p_a(x) \quad (6)$$

where $p_a(x)$ is the pdf of the a^{th} Gaussian.

2. Categorical distributions with the probability distribution of state s, F_s:

$$F_s = y_s \quad (7)$$

where y_s is a categorical distribution with K dimensions and $y_s(k)$ stands for the probability of the class k (while being in states s).

Since the categorical distribution is discrete, the KL divergence can directly be computed as given in (3). The KL-divergence between two Gaussian distributions could be calculated according to (4). However, the KL-divergence between two GMMs has no closed form solution.

In this study, we compare two different acoustic distances. 1) A simplified distance metric between two Gaussians as used by Young and Woodland [6]. 2) An approximation of the relative entropy between two GMMs as proposed by Hershey and Olsen [7].

1. Young and Woodland used the square root of the KL-divergence as interstate distance [6]. They also tried a related but much simpler distance metric which gave similar performance. That metric is still used by the HTK[2] toolkit. We implemented the same simplified distance metric between two states s_1 and s_2:

$$AD_{HTK}(s_1, s_2) = \sqrt{\frac{1}{V_x} \sum_{k=1}^{V_x} \frac{(\mu_{s_1,k} - \mu_{s_2,k})^2}{\sigma_{s_1,k} \sigma_{s_2,k}}} \quad (8)$$

where V_x is the dimensionality of feature vector x, and μ_s and σ_s are means and variances from the Gaussian associated to state s, respectively.

2. As already mentioned, the KL-divergence has no closed form solution for GMMs. The KL-divergence as defined by Kullback and Leibler is the sum of two relative entropies.

$$D(F, G) = D(F \parallel G) + D(G \parallel F)$$

The second acoustic distance that we use was proposed by Hershey and Olsen [7]. They used the Monte Carlo simulation to approximate the relative entropy between two GMMs associated to states s_1 and s_2 respectively, with pdfs $f_{s_1}(x)$ and $f_{s_2}(x)$. They drew a sample x_i from the pdf f_{s_1} such that:

$$E_{f_{s_1}} [\log f_{s_1}(x_i)/f_{s_2}(x_i)] = D(F \parallel G)$$

Hence:

$$AD_{KL}(s_1, s_2) = \frac{1}{n} \sum_{i=1}^{n} \log f_{s_1}(x_i)/f_{s_2}(x_i) \quad (9)$$

and using n i.i.d samples $\{x_i\}_{i=1}^n$

$$AD_{KL}(s_1, s_2) \to D(F \parallel G) \quad (10)$$

as $n \to \infty$. To draw a sample x_i from a GMM with pdf f, we first draw a discrete sample a_i according to the weights π_a (see Equation 6). Then, we draw a continuous sample from the corresponding pdf p_a.

[1] Kullback and Leibler originally named divergence what nowadays is often referred to as symmetric version of the Kullback-Leibler divergence [5].

[2] `http://htk.eng.cam.ac.uk/`

4. Experimental setup

For the experiments, we used GMMs as well as categorical distributions. The GMMs were used in the standard HMM/GMM framework as described in Section 4.1. The HMM/GMM experiments investigated DTSC and TM on Wall Street Journal (WSJ) data. Since we already showed that KL-HMMs are very powerful if only a small amount of training data is available [3, 4], we explored TM for KL-HMMs as described in Section 4.2 on limited amounts of SpeechDat(II) Greek data.

4.1. HMM/GMM system

We developed HMM/GMM systems using WSJ0 and WSJ1 continuous speech recognition corpuses [8]. All systems used three-state, cross-word triphone models, trained from 39 dimensional MFCCs (12 cepstral plus energy coef.) including delta and delta-delta features, with cepstral mean normalization. Training was performed with the HTS [9] variant of the HTK toolkit on the *si_tr_s_284* set of 37,514 utterances.

The pronunciation dictionary was based on the CMU pronunciation dictionary. We used the standard bigram and trigram backed-off language models *tcb20onp.z* from WSJ1 database, pruned to 20k target words defined by *wlist20o.nvp* from WSJ0 database. The standard test set *st_et_20* consisted of 303 utterances.

4.1.1. DTSC

As a baseline, we tied triphone models with DTSC based on the minimum description length (MDL) criterion [10]. The MDL criterion allows an unsupervised determination of the number of states. In this study, we obtained 12,685 states and modeled each state with a GMM consisting of 16 Gaussians.

4.1.2. TM

For the TM training we set the threshold $\lambda = 119$ to obtain 4226 selected triphones and a total of 12,678 states. Again, we modeled each state with a mixture of 16 Gaussians. The TM training used the same number of re-estimation iterations as DTSC training.

4.1.3. Successive Partitioning of Acoustic Space (SPAS)

As described in Section 2, we also implemented a slightly modified training procedure that successively partitions the acoustic space. Note that the SPAS does not consider all biphones that are present in the training data, but only the set of selected biphones B_s that constitute the set of selected triphones T_s. We observed that the number of biphones in B_s is roughly one half of all biphones (i.e. the successive partitioning increased model complexity from 41 monophones, to 765 selected biphones and then 4226 selected triphones).

Since SPAS training first re-estimates biphone and then triphone models, we adapted the number of re-estimation iterations to ensure that the total number of iterations is the same for DTSC, TM and SPAS.

4.2. KL-HMM system

A KL-HMM is an HMM that uses a categorical distribution as its output distribution. The name is taken from the Kullback-Leibler divergence distance measure that is employed. More specifically, each state of the HMM is modeled with a categorical distribution and phoneme posterior probabilities given the acoustics serve as features. The categorical distributions can be trained with a Viterbi segmentation optimization algorithm.

The idea is to estimate posterior probabilities with a Multilayer Perceptron (MLP) that can be trained on large amounts of out of language data. The KL-HMM parameters can then be estimated with only low amounts of within language data (target language).

As we did earlier [3], we used data from the SpeechDat(II) databases (*corpus S*). We used 63 hours of data in five European languages, namely British English, Italian, Spanish, Swiss French and Swiss German to train the MLP.

Greek was the target language. To simulate limited resources, the amount of training data varied from 5 hours to 5 minutes. For evaluation, we used a test set with 10k different words. Since we had no access to an appropriate language model, we simply built a language model with all the sentences from the test set. The language model had a perplexity of 44. In this sense, results should be considered as optimistic.

4.2.1. KL-HMM BO

The standard KL-HMM system was based on triphones. Without state tying, we limited ourselves to word-internal triphones only (as opposed to cross-word triphones for all the other systems). During decoding, we backed off (BO) to the context independent model of the center phoneme if a triphone was not seen during training. Each triphone was modeled with three states.

4.2.2. KL-HMM DTSC

The second KL-HMM system used an adapted version of a decision tree [11] and was therefore based on cross-word triphones. For the adapted version of the decision tree clustering, it was not obvious how to use the MDL criterion for the automatic determination of the number of states. Therefore, the optimal number of states was determined on a development set. The size of the development set varied depending on the amount of available data.

4.2.3. KL-HMM TM

The third KL-HMM system used the TM approach with the KL-divergence as given in (3) and was therefore also based on cross-word triphones. We adjusted the threshold λ to obtain similar number of states as for system KL-HMM DTSC. Note that we did not investigate SPAS training for KL-HMM systems yet.

5. Results

In this section, we first present the results of the HMM/GMM systems on WSJ data and then the results of the KL-HMM systems on SpechDat(II) Greek data. For the significance test, we used the bootstrap estimation method [12] and a confidence interval of 90%.

5.1. HMM/GMM system

We compared the DTSC to the TM systems using both acoustic distances AD_{HTK} and AD_{KL}, given in (8) and (9) respectively (1 milion samples were used in (9)). We hypothesize that AD_{KL} improves over AD_{HTK} and expect that TM performs better than DTSC and SPAS better than TM. Table 1 shows the results. Since DTSC uses a phonetic decision tree for state clustering instead of an acoustic distance as in TM-based system, the acoustic distance field of DTSC entry is empty.

Table 1: *Sentence (SRA) and word (WRA) recognition accuracies for WSJ task. The systems DTSC, TM and SPAS are described in Section 4.1.*

Method	Acoustic distance	SRA [%]	WRA [%]
DTSC	–	25.5	*90.3*
TM	AD_{HTK}	27.0	*89.9*
TM	AD_{KL}	*28.2*	*90.4*
SPAS	AD_{HTK}	*27.3*	*90.1*
SPAS	AD_{KL}	**28.8**	**90.5**

For WRA, all the systems perform similar (bold numbers are the best numbers in a column and italic numbers are not significantly different comparing to the best numbers). However, the test set contained 4.3% out-of-vocabulary (OOV) words. Since all systems perform around 90% WRA, the OOV rate might be responsible for the marginal improvement in WRA. Therefore, we run a McNemar test with a 90% confidence interval on SRA results. Indeed, this test revealed that both TM and SPAS (with AD_{KL}) significantly outperforms DTSC. While both systems produced almost identical number of substitutions and deletions, the number of insertions considerably decreased in SPAS recognition results. SPAS outperforms TM in sentence recognition accuracy, even thought the improvement was not significant.

Comparing different acoustic distances, there seems to be a tendency that AD_{KL} performs better than AD_{HTK}. However, the effect is not very pronounced. Altogether, SPAS with the KL-divergence based acoustic distance given in (9) performs best and significantly outperforms DTSC in SRA.

5.2. KL-HMM system

For KL-HMM, we compared systems KL-HMM BO, DTSC and TM. We hypothesize, that for very low amounts of data (5 minutes), KL-HMM DTSC and TM both outperform KL-HMM BO, because of data sparsity. For larger amounts of data however, we expect that system KL-HMM TM and BO perform equally well because there is enough data. However, KL-HMM DTSC might perform worse because there is a mismatch between the cost function used for decoding and DTSC. For decoding, we used the Kullback-Leibler divergence, but the DTSC algorithm adapted to KL-HMM used the relative entropy, because there is no closed form solution for the Kullback-Leibler divergence [11].

The results in Table 2 are consistent with both hypotheses. Bold numbers are the best numbers (given an amount of training data) and italic numbers are not significantly different.

6. Conclusions

We successfully generalized TM to KL-HMM based acoustic modeling and improved TM for GMM based acoustic modeling by introducing successive partitioning of the acoustic space during the training procedure.

Experiments on WSJ and Greek SpeechDat(II) data revealed that TM is robust and has the best overall performance. On the WSJ task, TM significantly outperforms DTSC. On Greek SpechDat(II) data, TM significantly outperforms DTSC if five hours of training data are available and performs similar to DTSC if only five minutes of data are available.

Table 2: *Word recognition accuracies (WRA) on Greek SpeechDat(II) for variable amounts of training data. The systems KL-HMM BO, DTSC and TM are described in Section 4.2.*

Amount of training data	System	WRA [%]
5 min	KL-HMM BO	76.6
	KL-HMM DTSC	**81.5**
	KL-HMM TM	*80.6*
75 min	KL-HMM BO	*83.2*
	KL-HMM DTSC	*83.6*
	KL-HMM TM	**83.8**
300 min	KL-HMM BO	*84.1*
	KL-HMM DTSC	83.0
	KL-HMM TM	**84.2**

7. Acknowledgements

M. Cerňak would like to thank Peder A. Olsen from IBM T. J. Watson Research Center for providing implementation details of Monte Carlo simulation. His research was partially supported by the VEGA project number 2/0202/11.

This research was supported by the Swiss NSF through the project Interactive Cognitive Systems (ICS) under contract number 200021_132619/1.

8. References

[1] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proceedings of the workshop on Human Language Technology*, ser. HLT '94. Stroudsburg, PA, USA: ACL, 1994, pp. 307–312.

[2] S. Darjaa, M. Cerňak, M. Trnka, M. Rusko, and R. Sabo, "Effective triphone mapping for acoustic modeling in speech recognition." in *Proc. of Interspeech*, 2011, pp. 1717–1720.

[3] D. Imseng, H. Bourlard, and P. N. Garner, "Using KL-divergence and multilingual information to improve ASR for under-resourced languages," in *Proc. of ICASSP*, 2012, pp. 4869–4872.

[4] D. Imseng, R. Rasipuram, and M. Magimai-Doss, "Fast and flexible Kullback-Leibler divergence based acoustic modeling for non-native speech recognition," in *Proc. of ASRU*, 2011, pp. 348–353.

[5] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[6] S. Young and P. Woodland, "State Clustering in Hidden Markov Model-based Continuous Speech Recognition," *Computer Speech and Language*, vol. 8, no. 4, pp. 369 – 383, 1994.

[7] J. R. Hershey and P. A. Olsen, "Approximating the Kullback Leibler divergence between Gaussian mixture models," in *Proc. of ICASSP*, vol. 4, 2007, pp. 317–320.

[8] D. B. Paul and J. M. Baker, "The design for the wall street journal-based CSR corpus," in *Proceedings of the workshop on Speech and Natural Language*, ser. HLT '91. Stroudsburg, PA, USA: Association for Computational Linguistics, 1992, pp. 357–362.

[9] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, "The HMM-based Speech Synthesis System Version 2.0," in *Proc. of ISCA SSW6*, 2007, pp. 131–136.

[10] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL principle for speech recognition," in *Proc. of Eurospeech*, 1997, pp. I –99–102.

[11] D. Imseng and J. Dines, "Decision tree clustering for KL-HMM," Idiap Research Institute, Tech. Rep. Idiap-Com-01-2012, 2012.

[12] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in ASR performance evaluation," in *Proc. of ICASSP*, 2004, pp. I–409–412.