CONSTRAINT INDUCTION OF PHONETIC-ACOUSTIC DECISION TREES FOR CROSSLINGUAL ACOUSTIC MODELLING

Frank Diehl, Asunción Moreno, Enric Monte

TALP Research Center
Universitat Politècnica de Catalunya (UPC)
Jordi Girona 1-3, 08034 Barcelona, Spain
{frank,asuncion,enric}@gps.tsc.upc.edu

ABSTRACT

In this work we focus on the construction of phonetic-acoustic decision trees for crosslingual use. Several modifications to the standard decision tree growing procedure are proposed aiming on integrating phonological and acoustical knowledge from source and target languages. This results in multilingual source models which already reflect characteristic of the target language though not trained on target speech data. Test results confirm the validity of this approach by improved system performances.

Index Terms— crosslingual, decision tree, acoustic modelling

1. INTRODUCTION

In recent years crosslingual acoustic modelling has attained the concern of many researchers [1], [2], [3]. This interest is driven from the circumstance that automatic speech recognition (ASR) technology highly relies on the availability of appropriate speech material. If an ASR system is ported to a new language a dedicated language specific speech database is needed to train the acoustic models. To alleviate this problem substantial efforts to create spoken language resources where undertaken by the speech community during the last decade. However, system designers are still faced with the problem of scarce training material. Reasons are manifold. A company might not be able to buy a large speech database, the available databases do not match the desired environmental conditions, or a database for the required language may not be available.

To attenuate the dependency on speech data some researchers focus on the crosslingual use of acoustic resources. Individual phonemes of different languages exhibit considerable similarities. Thus one tries to exploit these similarities by using already available acoustic models from one or more source languages to construct the acoustic model set of the target language.

As in the monolingual case it has also been shown in the crosslingual case that context dependent acoustic modelling outperforms context independent models [4], [1]. In the context dependent case the source-to-target model mapping is achieved by predicting target models from the phonetic-acoustic decision tree of the source language. One configures the question space of the decisions tree, e.g. phonetic broad classes or phonetic features, according to the target language. With these questions the tree is entered and traversed to its leaves. Finally, the source models associated to the tree's leaves are assigned to those target models which correspond to the applied questions.

Though the described approach is quite powerful it has turned out that it is not sufficient for generating high quality target models. Acoustic inter-language differences result to be too big that pure model mapping might give good results. Model adaptation applying a limited amount of target data is an indispensable measure in crosslingual acoustic modelling. In addition, beside the acoustic mismatches one is also faced by phonotactic inter-language differences. This results in a structural model mismatch due to the underlying decision tree reflects the specific phonotactic peculiarities of the source language. That is, some models of the source languages might not be used by the target language and, even worse, significantly different target models may be mapped onto one source model. To cope with this problem so-called polyphone decision tree specialization (PDTS) [5] was proposed to adapt the source language decision tree to the target language. PDTS consists in expanding the source language decision tree by restarting the tree growing process using some adaptation data of the target language. By this procedure unknown target-context information is effectively introduced in the existing tree.

In [5] and [6] PDTS has shown to perform quite well. However, the authors of [7] and [6] also report cases where its use was fairly effectless. In [7] it is argued that this might be caused by the fact that PDTS solely adapts the upper part of the tree, i.e. expands the leaves, whereas the more important splits of a decision tree appear near to the tree's root. Thus, the authors proposed some kind of bottom-up PDTS. They constructed the tree by using the adaptation material to build the bottom of the tree and the source languages data to build its upper parts, i.e. the leaves.

In both cases, PDTS and bottom-up PDTS the adjustment of the decision tree to the target language takes place in a sequential manner. The tree is grown in two steps applying first the source and afterwards the target data or vice versa. During each step neither phonological nor acoustic knowledge of the languages is mixed. The resulting tree consists therefore of two segments which are essentially monolingual.

In this work we investigate several tree growing procedures to be used in crosslingual applications. The procedures aim on integrating phonological and acoustical knowledge from the source and the target language. The integration takes place in a simultaneous parallel manner and not sequential as described in [5] and [7]. The final goal is to provide multilingual model definitions which are better suited for a crosslingual use. That is multilingual models which already reflect characteristic of the target language before applying further model refinements as e.g. PDTS.

2. CROSSLANGUAGE TREE INDUCTION

2.1. Language-entropy constraint

In [8] the authors investigate the crosslingual system performance respective the use of monolingual and multilingual source models. It

This work was granted by the CICYT under contract TIC2006-13694-C03-01/TCM.

turned out that multilingual source models clearly outperformed the monolingual ones. However, in [9] the same authors report that the multilingual system was actually not so much multilingual. It turned out that though training the decision tree on a multilingual set of source data the generated tree resulted in as far as possible monolingual subtrees and therefore models. Thus, the improved performance of the multilingual over the monolingual models stems largely from the few models which are in fact multilingual, i.e. trained on speech material of several languages.

Bearing the small number of real multilingual models in mind one may suppose that a higher fraction of multilingual models may further improve the crosslingual performance of the source models. The idea behind that reasoning is that acoustically broader multilingual models are better suited for a crosslingual use than highly specific monolingual ones. We therefore propose to constrain the induction process of the decision tree in such a way that the resulting models are more multilingual.

In the used speech recognition system a phonetic-acoustic decision tree is grown by minimizing the overall acoustic model entropy from split to split. This is equivalent to splitting each node t by that question q^*

$$q^* = \arg\max_q \Delta H_t^{(A)}(q) \tag{1}$$

which maximizes its acoustic entropy reduction

$$\Delta H_t^{(A)}(q) = n_t H_t^{(A)} - n_{t,L} H_{t,L}^{(A)}(q) - n_{t,R} H_{t,R}^{(A)}(q) \tag{2}$$

where L and R indicate the left and right child node and n_t, $n_{t,L}$, and $n_{t,R}$ stand for the frame count of the node to split and its children. To measure the multilinguality of a node we introduce a so-called language entropy $H_t^{(L)}$. It is defined on the normalized relative frame counts $n_{t,k}/n_t$ of the speech data of each language $k \in \{1, ..., K\}$ falling in state t, i.e.

$$H_t^{(L)} = -\sum_{k=1}^{K} \frac{n_{t,k}}{n_t} log \frac{n_{t,k}}{n_t}. \tag{3}$$

As $H_t^{(L)}$ ranges between 0 and $logK$ for a pure monolingual and a maximum multilingual node, respectively, language entropy values close to $logK$ indicate a high multilinguality. Corresponding to (2) we also define the language entropy decrease $\Delta H_t^{(L)}(q)$ coming along with question q.

We are now ready to define a modified node splitting rule which aims on increasing the multilinguality of the tree. Instead of directly choosing the question q which maximizes the acoustic entropy decrease when splitting node t we take that question out of the set of the N_t best questions which minimizes the language entropy decrease. In detail, using the nomenclature $\max^{(N_t)}$ to specify the operation '*the N_t biggest values of*' we define the auxiliary question set $\mathcal{Q}$ as

$$\mathcal{Q} = \{q | \arg\max_q{}^{(N_t)} \Delta H_t^{(A)}(q)\}. \tag{4}$$

The set $\mathcal{Q}$ defines the N_t questions giving the biggest decrease in acoustic entropy. The node splitting rule can now be stated as

$$q^* = \arg\min_{q'} \Delta H_t^{(L)}(q') \qquad \forall q' \in \mathcal{Q}. \tag{5}$$

We call the parameter N_t splitting depth. It is dynamically set to a ceiling of the fraction μ of the number of possible questions N_t^{max} one can form in node t, i.e. $N_t \in \{1, ..., N_t^{max}\}$ with

$$N_t = \begin{cases} \lceil \mu N_t^{max} \rceil & \text{if} \quad \mu \in]0, 1] \\ 1 & \text{if} \quad \mu = 0 \end{cases} \tag{6}$$

Note that for $\mu = 0$ the new splitting rule goes over in the original one, i.e. without the language entropy constraint.

2.2. Source language tree with question set constraint

In Section 1 we have already mentioned that usually only a subset of the source language models are used by the target language. As an example we refer to a previous study [6] where a multilingual German-Spanish-English decision tree was used to define Slovenian and French models. It was found that in case of Slovenian only 1017 of the original 1500 models, i.e. tree leaves, were used by the target models. In case of French this number yet lowered to 696, that is, to less than half of the source models.

The main reason for this behavior is that the source languages may comprise phonemes and phonetic characteristics which do not apply to the target language. When traversing the source language tree by questions corresponding to the target language some regions of the tree are therefore simply not visited. On the other hand, the target language will in general comprise phonetic properties which were not seen in the source languages. These properties may make the difference between phonemes of the target language yet the source tree is not able to distinguish them. As a consequence, predicting target models out of the phonetic-acoustic decision tree of a source language will often result in a one-to-many mapping, meaning that one source model will be assigned to a whole group of target models. In summary, on the one hand the model mapping process selects highly selective source models for the target models which represent on the other hand quite wide-stretched partitions of the acoustic target space.

Furthermore, even if the one-to-many mapping problem would not exist, we allude that an exact acoustic match between a phoneme of the source and the target language is nearly never given. The acoustics of languages are just too different. It might happen that the phonetic system used to define the question sets indicate such a perfect match. However, in general this is not the case. This is due to the fact that the applied question sets are typically based on IPA [10] or SAMPA [11] transcriptions. Yet SAMPA but also common IPA transcriptions are language specific. In case of SAMPA this is inherent to SAMPA itself. In case of IPA this is caused by the commonly applied *principles of phonological contrast* [10]. Hence, neither from an acoustic nor from a phonological point of view one can expect to find exact matching source models for the target language.

Altogether, one might be in doubt about assigning highly selective source models to the target language. Instead it may be advantageous to use source models which are less selective, i.e. acoustically broader. We therefore propose to broaden the source models acoustically by avoiding the fragmentation of the acoustic source space in partitions which can not be accessed by the target language. This can easily be obtained by the following question set constraint:

- Constrain the question set of the source languages to that questions which also appear in the target language.

2.3. Target language tree with question set constraint

In crosslingual speech recognition it is common practice to base the definition of the target models on the phonetic-acoustic decisions tree of a set of source languages. On the other hand, using a decision tree which is directly dedicated to the target language might give better results. In this work we have found that also with solely 426 sentences target data a quite reliable decision tree for the target language can be build. Yet, the target speech material might be too scarce to train high quality target models. To cope with this problem we propose to train the models by the speech data of the source languages. However, as the decision tree is trained by target data not all resulting models are trainable by source data (see the reasoning in the previous section). To overcome this problem we propose the following constraints for the growing process of the target language tree:

- Reduce the question set of the target language to such questions which also appear in the source language.
- Only split a node if its resulting children nodes can be trained by source language data (trainable in the sense of sufficient training material).

The resulting target language specific decision tree defines therefore a model set which is trainable by the source language data.

3. SYSTEM OVERVIEW

We use a SCHMM system calculating every 10ms twelve mel-cepstrum coefficients (MFCC) (and the delta energy) using cepstral mean subtraction. First and second order differential MFCCs plus the differential energy are employed. For each stream a codebook is constructed consisting of 256 and 32 (delta energy) Gaussian mixtures, respectively. We use 3-state state-tied left-to-right demiphones. Demiphones [12] can be thought of as triphones which are cut in the middle giving a left and a right demiphone. For the state tying a binary decision tree is used applying phonetic questions derived from the IPA-chart.

4. THE MAPPING TASK

Experiments were carried out using SpeechDat-II fixed telephone databases. The task was to map speaker independent multilingual Spanish-English-German models to Slovenian. The multilingual systems were trained with phonetically rich sentences of 3000 speaker, 1000 from each language. The target data set which is used for constructing the tree of Section 2.3 and for the acoustic adaptation of the models comprises 50 speakers. It is balanced respective sex and consists of 426 phrases so-called phonetically rich sentences.
Testing the systems was hampered by the fact that quite high error rates were observed and, as a consequence, the confidence margins for the word error rates (WER) were quite big. We could therefore not conclude that a single system was better than another. To overcome this problem we based the system evaluation on a two-way analysis of variance test (ANOVA) which tests for the hypothesis that the mean WERs (calculated over several test sets) of two systems are equal [13]. Thus, instead of a single test 8 similar but independent tests were run for each system configuration. Afterwards the mean word error rates of two system configurations were compared by ANOVA. In the following sections we therefore present mean WERs. Statistically significant (95% confidence interval) different results, with respect to some reference results, are marked boldface. The 8 test sets themselves consist of single phonetically rich words and application words and comprise $619 - 646$ sentences. The resulting grammars, just word lists, exhibited a word based perplexity of $360 - 383$.

5. TESTS AND DISCUSSION

The tests are divided in three sections, one for each method proposed in Section 2. In each case the test consists in constructing a corresponding decision tree followed by training the models with the complete training material of the three source languages. Next, target models are predicted from the decision tree and the source models are mapped, i.e. copied, to their counterparts of the target language. These so-called 'PRED' models are adapted to the target language in two steps. First, they are just retrained by one iteration Baum-Welch training using the target data. These models are called 'TRAI'. Second, a dedicated acoustic model adaptation scheme namely maximum a posteriori convex regression (MAPCR) [6] is applied. The resulting models are named 'ADAP'. Finally the results are compared to a corresponding crosslingual Slovenian model set which was constructed in the traditional way, i.e. by using a source language decision tree exclusively induced with data from the source languages. The corresponding results will be named 'Reference'.

5.1. Language-entropy constraint

In this paragraph we investigate the influence of the language-entropy constraint, see Section 2.1, on the crosslingual system performance. We start by analyzing the influence of the control parameter μ on the multilinguality of the decision tree. In Figure 1 we depict the decrease in the accumulated language entropy for growing tree sizes. The unconstrained case $\mu = 0$ is compared with the constraint cases $\mu = 0.02$, $\mu = 0.03$, and $\mu = 0.04$. Obviously, the modified node splitting rule according to equation (4) and equation (5) makes the model sets more multilingual. As bigger μ as bigger become the accumulated language entropies of the trees.

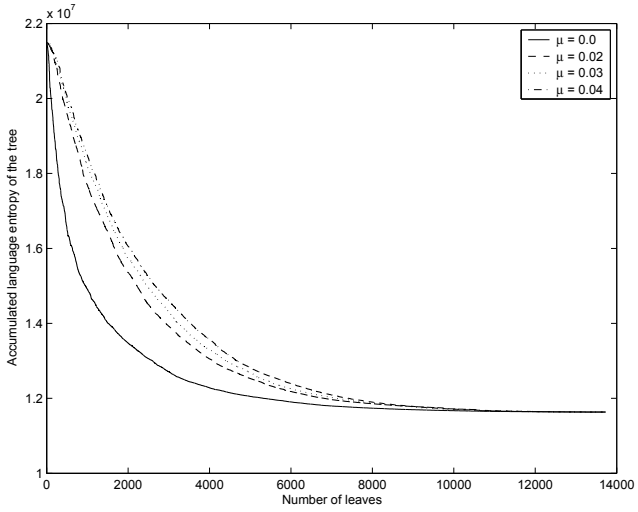

Fig. 1. *Accumulated language entropy versus number of leaves.*

Next we focus on the system performances which were achieved when applying the corresponding PRED-models, see Table 1. Statistically significant different results respective the unconstrained case, i.e. $\mu = 0$, are marked boldface. Table 1 confirms the assumption that an improved crosslingual modelling is possible when applying source models which are more multilingual. Between $\mu = 0.02$ and $\mu = 0.05$ significantly lower mean WER than in the unconstrained case were obtained. Best performance was achieved by $\mu = 0.03$ lowering the mean WER by 2.78% respective the reference, i.e. the $\mu = 0$ case.

Table 1. *Mean WERs for PRED-models*

μ	0.00	0.01	0.02	0.03	0.05	0.07
PRED	47.33	48.08	**46.67**	**44.55**	**45.75**	**48.87**

We went on by retraining and adapting the best PRED models, $\mu = 0.03$, and the reference set, $\mu = 0$, using the adaptation data. As can be seen from Table 2 the modified tree growing procedure also pays off after acoustic model adaptation. Though the reductions in mean WERs are not as high as in the PRED case one still obtains reductions of 0.9% in terms of mean WER.

Table 2. *Mean WERs for TRAI- and ADAP-models*

	TRAI	ADAP
Reference, $\mu = 0$	20.77	18.47
Constraint tree, $\mu = 0.03$	**19.87**	**17.58**

5.2. Source language tree with question set constraint

In this paragraph we analyze the influence of reducing the question set of the source languages to such questions which are also valid for the target language, see Section 2.2. A decision tree was built and the PRED, TRAI, and ADAP model sets were created. In Table 3 the corresponding test results are compared with the reference results.

Table 3. *Mean WERs for PRED-, TRAI-, and ADAP-models*

	PRED	TRAI	ADAP
Reference	47.33	20.77	18.47
Constraint tree	**45.23**	**20.02**	**17.71**

As in the case of introducing a language-entropy constraint also in this case significant improvements of the system performances were found ranging from $0.75 - 2.10\%$ mean WER.

5.3. Target language tree with question set constraint

In the following we analyze the tree growing method proposed in Section 2.3. A decision tree was build based on the Slovenian adaptation data. During the tree growing process the target question set was reduced to the questions which are also valid for the source languages. Multilingual models were trained on this model definition resulting actually in the PRED models which in turn built the base to generate the TRAI and ADAP models. Table 3 summarizes the corresponding test results and compares them with the reference results.

Table 4. *Mean WERs for PRED-, TRAI-, and ADAP-models*

	PRED	TRAI	ADAP
Reference	47.33	20.77	18.47
Constraint tree	**50.00**	**18.30**	17.98

Inspecting Table 4 one finds that the new unadapted, solely predicted models perform significantly worse than the reference. However, retraining them with the adaptation data results in a mean WER for the TRAI models which outperforms the reference TRAI and all other TRAI results clearly, see Table 1 and 3. This behavior might be explained by the fact that the underlying model definition, i.e. the decision tree, is based on Slovenian data. Training such models with source language data results in relatively bad PRED models due to the language mismatch between the tree and the data. However, when retraining the models with the adaptation material the data matches better the model definition, which in fact is based on the adaptation data, resulting in better TRAI models.
Finally we allude that also in the ADAP case the new models outperform the reference results yet the results are not significantly better than the reference outcomes.

6. SUMMARY

In this work we have focused on the construction of phonetic-acoustic decision trees for improved crosslingual acoustic modelling. It was proposed to integrate phonological and acoustical knowledge of the target language in the tree construction process of the source languages. Three different methods were pointed out to cope with this aim. Test results underlined the usefulness of the suggested methods. After the final model adaptation step all crosslingual model sets which were constructed according to the new procedures outperformed crosslingual model set which were generated in the traditional way.

7. REFERENCES

[1] T. Schultz and A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, no. 1-2, pp. 31–51, Aug. 2001.

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

[3] C. Liu and L. Melnar, "Training acoustic models with speech data from different languages," *ISCA Tutorial and Research Workshop (ITRW) on Multilingual Speech and Language Processing*, Apr. 2005.

[4] P. Beyerlein, U. Meinhard, and P. Wilcox, "Modelling and decoding of crossword context dependent phones in the Philips large vocabulary continuous speech recognition system," *European Conference on Speech Communication and Technology*, pp. 1163–1166, Sep. 1997.

[5] T. Schultz and A. Waibel, "Polyphone decision tree specialization for language adaptation," *International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 1707–1710, Jun. 2000.

[6] F. Diehl, A. Moreno, and E. Monte, "Crosslingual adaptation of semi-continuous HMMs using acoustic sub-simplex projection," *ISCA Tutorial and Research Workshop (ITRW) on Multilingual Speech and Language Processing*, Apr. 2006.

[7] T. Martin and S. Sridharan, "Cross-language acoustic model refinement for the indonesian language," *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 865–868, Mar. 2005.

[8] T. Schultz and A. Waibel, "Language adaptive LVCSR through polyphone decision tree specialization," *Multi-Lingual Interoperability in Speech Technlogy*, pp. 97–102, Sep. 1999.

[9] T. Schultz and A. Waibel, "Language independent and language adaptive large vocabulary speech recognition," *International Conference on Speech and Language Processing*, pp. 1819–1822, 1998.

[10] The International Phonetic Association, *Handbook of the International Phonetic Association*, Cambridge University Press, 1999.

[11] J. Wells, W. Barry, M. Grice, A. Fourcin, and D. Gibbon, "Standard computer-compatible transcription," *Doc. No. SAM-UCL-037*, 1992.

[12] J. B. Mariño, P. Pachès-Leal, and A. Nogueiras, "The demiphone versus the triphone in a decision-tree state-tying framework," *International Conference on Spoken Language Processing*, vol. 1, no. 5, Nov. 1998.

[13] T. W. Anderson, *An introduction to multivariate statistical analysis*, John Wiley & Sons, Inc., 1984.