IDIAP RESEARCH REPORT

ADAPTATION EXPERIMENTS ON FRENCH MEDIAPARL ASR

Gyorgy Szaszak

Idiap-RR-10-2013

MARCH 2013

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Adaptation Experiments on French MediaParl ASR

György Szaszák

March 23, 2013

Abstract

This document summarizes adaptation experiments done on French MediaParl corpus and other French corpora. Baseline adaptation techniques are briefly presented and evaluated in the MediaParl task for speaker adaptation, speaker adaptive training, database combination and environmental adaptation. Results show that by applying baseline adaptation techniques, a relative WER reduction of up to 22.8% can be reached in French transcription accuracy. For the MediaParl task, performance of systems trained on directly merged databases and of systems trained on databases combined via MAP adaptation did not differ significantly when large amount of data was available. During the experiments, French data recorded in Switzerland behaved in a similar way compared to French data recorded in France, which suggests that French spoken in Valais is close to the standard French spoken in France, and differencies in ASR accuracies between models trained on Swiss MediaParl and on French BREF are more likely caused by environmental factors or more spontaneity in speech.

1 Introduction

MediaParl is an Idiap project financed by the Parliament of the canton of Valais in Switzerland. The goal of the MediaParl project is to provide speech recognition for Swiss languages initially for parliament data available from recordings at the Valaisian Parliament. These recordings are structured into the MediaParl database, a bilingual speech database, containing recordings in both French and German. As the deputies are elected for 4 years, parliament staff are relatively permanent, allowing for using and testing of different speaker adaptation approaches. On the other hand, the available data in the whole MediaParl database is not so ample, this means that it is worth investigating combinations of other native speech databases in French or German in order to ensure sufficient data set for model training.

The goal of this report is to present and evaluate state-of-the-art adaptation techniques in the French MediaParl task, including different baseline speaker adaptation methods like Maximum Likelihood Linear Regression (MLLR) and its forms, Speaker Adaptive Training (SAT), Maximum a Posteriori adaptation (MAP) and the combined MLLR-MAP adaptation method Structural Maximum a Posteriori Linear Regression (SMAPLR), this latter used so far initially for TTS adaptation. Both speaker and environment adaptation is addressed. The baseline techniques were also investigated for database combination, where adaptation is primarily used for shifting the acoustic environment and to a lesser extent, adapting the models to Swiss accented speech, but does not aim at speaker dependent model transform.

This report does not focus specially on the bilingual aspect or non-native speaker adaptation, as this task - at least partly - has already been assessed (Imseng et al., 2012) and constitues a future challenge.

2 Speaker adaptation

The main purpose of the present report is to assess speaker adaptation. First, a brief basic overview of underlying theory is provided. More details on each algorithm can be found following the respective citation provided in the text.

2.1 MLLR Adaptation of means

Maximum Likelihood Linear Regression (MLLR) is an often used linear transformation specified by (Gales and Woodland, 1996):

$$\hat{\mu} = A\mu + b = W\xi. \tag{1}$$

This means that the transformed mean vector ($\hat{\mu}$) is obtained by applying the W transform to the extended original mean vector (ξ), obtained by extending the original n-dimensional mean vector $\mu = [\mu_1, \mu_2, ..., \mu_n]$ as:

$$\xi = [1, \mu_1, \mu_2, ..., \mu_n]^T. \tag{2}$$

This transform is linear and it can be easily decomposed as:

$$W = [b, A]. \tag{3}$$

The transform can be computed by ML estimation for the adaptation data from the speaker.

2.1.1 Global vs. model specific adaptation

In the easiest approach, only one adaptation transform is computed on all adaptation data. This is called *global* transform. However, different models or even different mixtures of different states in a HMM/GMM system might benefit from different transformations to better fit to a given speaker. This means that multiple transforms should be created and a clustering of states or mixtures is needed to be specified in order to group parameters that share a transform. The number of the clusters depend on available adaptation data: if more data is available for a given cluster, it can be split into further sub-clusters to allow for more specialized adaptation. In practice, this is usually done using a regression class tree (Gales, 1996), which allows for a flexible data-driven clustering. The regression class tree is constructed so as to group similar components (that are close in acoustic space), this means that similar components will share a common transform. The tree can be grown until each leaf has sufficient adaptation data to estimate a transform. These transforms can be referred to as *tree-based* ones, which are expected to outperform global transforms.

2.2 Variance adaptation

Variances of the speaker independent models can also be adapted, this is usually done in a two stage approach. In the first stage, mean adaptation is carried out, secondly, in a separate step, variances are also updated using already the adapted means (as a so-called parent transform) for the computation. The covariance matrices (Σ) can be updated based on the following formula (Gales and Woodland, 1996):

$$\hat{\Sigma} = B^T H B, \tag{4}$$

where H is the transform to be estimated, and B is the inverse of the Choleski factor of Σ^{-1} , this means that

$$\Sigma^{-1} = CC^T \tag{5}$$

and

$$B = C^{-1}. \tag{6}$$

2.3 Constrained MLLR (CMLLR)

The idea in CMLLR is that the means and variances can be adapted using the same transform (Digalakis et al., 1995), (Gales, 1998). This means that this transform has fewer parameters and hence less adaptation data can be sufficient for effective estimation of the transform. The mean μ and the variance Σ are transformed as follows:

$$\hat{\mu} = A\mu + b, \tag{7}$$

$$\hat{\Sigma} = A\Sigma A^T. \tag{8}$$

Transform parameters to be computed involve A and b.

2.4 Maximum a posteriori adaptation (MAP)

The basic idea of MAP adaptation is to try to establish speaker dependent models based on prior speaker independent knowledge. In this way, adapted system performance is expected to be close to that of a speaker

dependent system, but the amount of necessary training data is much lower than in case of training a pure speaker dependent system. The adaptation formula, used also in HTK, is as follows:

$$\hat{\mu}_{jm} = \frac{\tau\mu^0_{jm} + N_{jm}\mu_{jm}}{N_{jm} + \tau} \tag{9}$$

that is, to update the mean vector of mixture component m associated with state j, original mean μ_{jm} and the mean calculated on the adaptation data μ^0_{jm} are recombined where τ is a weighting parameter for the a-priori knowledge.

Comparing MLLR-like and MAP transforms, MLLR transforms usually work better than MAP with little adaptation data, however, MLLR performance often saturates when more training data is available. On the other hand, MAP has better asymptotic properties, that is, MAP exhibits convergence to real speaker dependent systems' performance, but this convergence is slow. MLLR-like and MAP transforms can be used simultaneously (Goronzy and Kompe, 1999) to benefit from the advantages of both approaches.

2.5 Structured Maximum a Posteriori Linear Regression (SMAPLR)

Structured Maximum a Posteriori Linear Regression can be regarded as a sort combination of MLLR and MAP transforms in order to exploit the advantages of both approaches: the less extensive data requirement of MLLR and the more accurate (and asymptotically better converging) estimation capabilities of MAP (Siohan et al., 2002). The idea behind Structural MAP (SMAP) is the same used for regression tree-based MLLR: model parameters (GMMs) are clustered which allows for specific transforms for each cluster. The difference is the use of a maximum a posteriori (MAP) criterion instead of the maximum likelihood (ML) one. This involves the use of some prior distributions. These prior distributions can be estimated form the speaker independent models, but also obtained along the regression tree: once a prior is known for the root of the tree, it can be used as a constraint to calculate the priors of its child nodes. A possible algorithm for this is given by Siohan et al. (2002). The constrained form of SMAPLR (CSMAPLR) is a combination of SMAPLR and CMLLR, and performs adaptation of variances too (whilst SMAPLR is usually used for the means). CSMAPLR is described in detail by Nakano et al. (2006).

3 Speaker Adaptive Training (SAT)

In Speaker Adaptive Training, speaker specific transformations are used during the training process itself in order to create models of individual speaker characteristics. The transformations typically used are MLLR (or rather CMLLR). These speaker specific transforms are used for "de-individualizing" the utterances during the model (re-)training. In this sense, SAT can be interpreted as a special form of normalization. The underlying assumption is that human speech is a product of two components (Anastasakos et al., 1997): the first component is the raw speech representing purely phonetic variations, while the second one is speaker specific and represents speaker variation (caused by physiological differences, age, gender or even the different acoustic environments and so on). Interpreting the second component as a filter, the final speech product is the convolution of the two. Thus, by modelling speaker variations, it is possible to normalize them and to create a model set independent of speaker characteristics (which is then also further adaptable to each speaker separately). In this sense, speaker independent models are independent not in the sense of being all-speakers models trained on a large number of speakers, but are ideally free of any speaker specific influence.

This also means that in SAT, speaker transforms can operate on the feature level, so instead of creating an adapted (specialized) model for each speaker, features can be transformed so that they become speaker independent (as much as possible):

$$\hat{o} = Ao - b. \tag{10}$$

If sufficient training data is available from the speakers, SAT can be done using speaker transforms obtained by regression tree-based clustering. Otherwise, global adaptation data is used.

4 Adaptation experiments for French MediaParl

For adaptation experiments, we use the French version of the MediaParl ASR, trained on French utterances of the MediaParl database. The MediaParl database is presented in detail by Imseng et al. (2012). The baseline setup

of the ASR used for the present experiments is based on PLP features and cepstral mean normalization (CMN), acoustic models are tied-state HMM triphone models with 8 Gaussian mixtures. Training was performed with the HTS variant (version 2.2) of the HTK toolkit on training data labelled as 'Grandconseil French train sentences without noise' at the Idiap repository, containing a total of 6256 utterances from 106 speakers. Note that this setup is slightly different from the one used by Imseng et al. (2012): the split of the database for training, development and testing sets is different, the phoneme set and dictionaries are identical, whilst language models are the same too with the exception that no EuroParl data is used in present report to train language models. (Differences in the training settings are explained by the fact that for present experiments not all bilingual speakers were reserved for testing, but many of them are rather involved here in the training set.)

The original test set is further split up into two subsets, one of them holds 10 utterances from each speaker for final testing (referred to further as adaptation-test set), while the other subset holds the remaining files for computing MLLR transforms (referred to as adaptation set). If a speaker within the test set has fewer then 20 utterances, each of them is entirely removed from the test set. The adaptation set contained 421 utterances, while the adaptation-test set contained 90 utterances (2661 words) from 9 speakers (speakers 004, 013, 062, 075, 096, 102, 126, 135, 162).

4.1 MLLR in French MediaParl

MLLR adaptation is performed in a supervised, static manner in two steps: first a global transform is computed (base transform), then a regression tree is generated, based on which tree-based transforms are also computed (tree based transform). The regression tree used in the experiment had maximum 32 leaves (for speakers with insufficient adaptation data this can be less as trees are pooled to ensure enough adaptation data for each terminal node).

Results are presented in Table 1. Results (word accuracy and WER improvement compared to baseline) for global (base) and tree-based transforms are presented separately. As SMAPLR and CSMAPLR are structured MAP transforms, they can be regarded as tree-based ones for comparison.

Adaptation type	Adaptation kind	Adapted parameters	Acc [%]	WER reduction [%]
Baseline	-	-	72.83	0.00
MLLR	base	MEAN	74.78	8.86
MLLR	base	MEAN+VAR	74.26	8.33
CMLLR	base	MEAN&VAR	74.78	11.36
MLLR	tree	MEAN	75.08	13.10
MLLR	tree	MEAN+VAR	75.42	15.08
CMLLR	tree	MEAN&VAR	75.57	15.96
SMAPLR	tree	MEAN	75.05	12.93
CSMAPLR	tree	MEAN&VAR	75.01	12.70

Table 1: MLLR based adaptation results for French MediaParl

Results show that any type of the examined adaptation modes yielded improvement in recognition performance. The highest improvement is seen with tree-based constrained MLLR (CMLLR). CSMAPLR did not perform better than CMLLR.

4.2 Speaker adaptive training for MediaParl

In speaker adaptive training, models are retrained using speaker specific transforms. The transforms used were CMLLR transforms, generated on tied-state, 8 Gaussians triphone models. The same models were then retrained by using the obtained transforms in 5 subsequent re-estimation cycles. For evaluating SAT performance, CMLLR transforms have to be generated also for the test speakers.

SAT results are shown in Table 2. for French MediaParl task.

SAT based on	SAT transform type	Adapted parameters	Acc [%]	WER reduction [%]
Baseline	-	-	72.83	0.00
CMLLR	base	MEAN&VAR	75.76	17.06
CMLLR	tree	MEAN&VAR	76.74	22.77

Table 2: SAT results for French MediaParl

As expected, tree-based speaker adaptive training outperforms the global one, and SAT even outperformed single CMLLR transforms applied during the decoding for non retrained models.

4.3 Combining SAT and MLLR

Combining SAT and MLLR involves MLLR adaptation for models obtained during SAT. Results are presented in Table 3. for French MediaParl. Single tree based SAT without additional MLLR outperforms the combined test results. When using variance adaptation beside SAT+MLLR, performance was significantly worse (65.28%) than the baseline one. This means that a pure tree based SAT was found more efficient. Using CMLLR for SAT models did not improve results, indeed, the SAT itself is already CMLLR-based.

SAT based on	SAT transform type	Adaptation type	Transform type	Acc [%]	WER reduction [%]
Baseline	-	-	-	72.83	0.00
CMLLR	base	MEAN MLLR	base	75.61	16.19
CMLLR	base	MEAN MLLR	tree	76.51	21.43
CMLLR	tree	MEAN MLLR	base	76.44	21.03
CMLLR	tree	MEAN MLLR	tree	74.45	9.44

Table 3: SAT+MLLR results for French MediaParl

In order to further analyze SAT and SAT+MLLR capabilities, experiments were extended for French ASR models trained on the BREF database, a considerably larger corpus which contains more utterances from each speaker (Lamel et al., 1991), allowing for an exhaustive analysis of the performance when data sparsity is unlikely to cause problems. The baseline setup of French BREF acoustic models was as follows: PLP features were used with cepstral mean normalization, models were tied-state GMMs of up to 16 mixture components. Testing set involved all utterances from speakers (JDM, JZF, K2F, K3F, K4F, K5M, K6M, K7M), split again into adaptation set (4564 utterances) and adaptation-test set (480 utterances, 60 from each speaker). All other speakers were assigned to training test. Tri-gram language model and dictionary were generated over the training set transcripts. Results are presented in Table 4.

SAT based on	SAT transform type	Adaptation type	Transform type	Acc [%]	WER reduction [%]
Baseline	-	-	-	95.85	0.00
No SAT	-	MEAN MLLR	tree	95.90	1.20
CMLLR	tree	No MLLR	-	96.28	10.36
CSMAPLR	tree	No MLLR	-	96.12	6.51
CMLLR	tree	MEAN MLLR	tree	96.30	10.84

Table 4: SAT, MLLR and SAT+MLLR results for French BREF

Although baseline performance was already very high in the implemented BREF recognition task, MLLR, SAT, and combined SAT+MLLR all yielded some additional improvements. Comparing results to MediaParl ones in Table 3., where combined SAT+MLLR were less effective than SAT alone, a possible explanation of the difference can be raised. Namely, it is possible that if more adaptation data is available for the MediaParl experiment, combined SAT+MLLR could bring some modest performance improvement compared to SAT alone. For BREF, the adaptation set contained 4564 utterances (sentences) from 8 speakers, while in MediaParl only 421 sentence utterances of comparable length were available. This also explains why regression trees for several speakers could not use final leaves to compute the transformations but had to be pooled in MediaParl, whereas in BREF, all 32 final leaves had sufficient adaptation data associated for all of the speakers.

4.4 Using MAP for speaker adaptation

As in MediaParl the data available from each speaker is usually insufficient to perform MAP adaptation, experiments were conducted on French BREF with settings described in the previous subsection (adaptation set contained 4564 utterances from 8 speakers, testing set 480 utterances from the same 8 speakers (60 utterances from each of them). 16 GMM triphone models were adapted separately for each speaker in the following way: first, missing triphone were synthesized from tied 16 GMM model set, then MAP adaptation was run in 3 subsequent iterations with $\tau = 10$ using the speakers' adaptation data. Then a full coverage triphone set was synthesized. The baseline performance of 95.85% dropped to 93.74% after MAP adaptation.

4.5 Using MAP to combine training data from different sources

Adaptation can be used to perform environment adaptation too. MAP adaptation was tested for this purpose in adapting French BREF and GlobalPhone acoustic models with French MediaParl data (MediaParl train set) and testing on MediaParl data (MediaParl test set). French BREF was also adapted with GlobalPhone to see whether there is difference between French-French and Swiss-French adaptation performance.

Additionally, models trained on mixed BREF+MediaParl data were also evaluated to compare them to baselines and to results obtained with MAP adaptation based database combination.

Experimental setups were as follows:

- for BREF: the ASR setup presented in subsection 4.3.
- for GlobalPhone: Training set (speakers 001-079) was used as training set to train 16 GMM tied-state triphone models with standard PLP features + CMN when creating baseline to be adapted with MediaParl. This same set was used as adaptation set when BREF models were adapted with GlobalPhone data. Testing set (speakers 080-089) was used for testing in all scenarios without further split up.
- for MediaParl: Training set ('Grandconseil French train sentences without noise') was used as adaptation set, testing set (speakers 004, 013, 062, 075, 096, 102, 126, 135, 162) was used for testing.
- for BREF+MediaParl: the ASR setup presented in subsection 4.3, but with merged training and testing sets for BREF and MediaParl.

Acoustic Models	MAP on	tested on	Acc [%]
MediaParl	-	MediaParl	73.84
BREF	-	MediaParl	57.29
BREF	MediaParl	MediaParl	72.92
BREF+MediaParl	-	MediaParl	70.54
BREF+MediaParl	MediaParl	MediaParl	70.92
GlobalPhone	-	MediaParl	55.63
GlobalPhone	MediaParl	MediaParl	73.47
GlobalPhone	-	GlobalPhone	71.13
BREF	GlobalPhone	GlobalPhone	71.82

Table 5: MAP results for French ASR (τ=10)

MAP results (Table 5.) show that by using MAP, approximately the same accuracy can be reached as by training directly on the database representing the target domain. No significant difference is seen in this regard between French-French and Swiss-French adaptation data. The adapted models are expected to be more robust on the target domain than models trained directly to the target domain. So to conclude, if sufficient training data is available, MAP adapted models cannot be expected to yield superior performance compared to models trained directly on the target domain, but MAP adapted models can be more robust (this hypothesis needs further experimental confirmation). MAP based database combination performed slightly better then directly merging the databases.

The effect of changing the τ constant was also evaluated using BREF models and GlobalPhone adaptation data. No significant difference was noted depending on the value of τ (see Table 6.) .

τ	Acc [%]
5	71.81
10	71.82
15	71.78
20	71.78

Table 6: MAP results with GlobalPhone adaptation data for BREF models depending on τ

4.6 Combining MAP and MLLR

Combining MLLR and MAP adaptation is reported to further improve performance in speech and speaker recognition by several studies (Goronzy and Kompe, 1999), (Wang et al., 2009). However, as MAP did not yield and improvement in speaker adaptation and as in environment adaptation MLLR-like approaches are not the ideal choice, combining MAP and MLLR fell out of scope for this report. Indeed, combining MAP and MLLR in environment adaptation considerably degraded the performance.

5 Conclusions

In this report speaker and domain adaptation was analyzed for French ASR tasks with a special attention paid to the MediaParl task. Several baseline adaptation techniques were evaluated. The most improvement in speech recognition accuracy was linked to speaker adaptive training (SAT) using tree-based CMLLR speaker transforms.

To a lesser extent, mean and variance MLLR, CMLLR, SMAPLR and CSMAPLR methods also contributed to some performance improvement. MAP adaptation did not improve accuracy when using for speaker adaptation, but proved to be useful for database combination or environment adaptation. Although in the latter case performance did not differ significantly from the baseline system trained directly on target domain data, models can be expected to be more robust.

The results showing that French data recorded in Switzerland (MediaParl) and French data recorded in France (GlobalPhone) behave very similarly when using it for environment adaptation with MAP on a model set trained on French data recorded in France (BREF) strengthen the suspicion that standard French spoken in France and French spoken in Switzerland do not differ significantly from ASR modelling point of view (i.e. the French spoken in Switzerland does not seem to be an individual dialect of French worth special ASR modelling or dialect adaptation). The high differences between BREF vs. MediaParl and GlobalPhone based ASR baseline performances are therefore more likely caused by environmental conditions and more spontaneity in speech, both factors recognized as important cues determining ASR performance.

6 Acknowledgements

The author expresses his gratitude to the Parliament Service of the State of Valais, Switzerland for their financial support and for providing access to the audio-video recordings.

References

T. Anastasakos, J. McDonough, and J. Makhoul. Speaker adaptive training: a maximum likelihood approach to speaker normalization. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 1043–1046, 1997.

V. Digalakis, D. Rtischev, and L. Neumeyer. Speaker adaptation using constrained estimation of gaussian mixtures. *IEEE Transactions on Speech and Audio Processing*, 3:357–366, 1995.

M. J. F. Gales. The generation and use of regression class trees for MLLR adaptation. Technical Report CUED/F-INFENG/TR263, Cambridge University, 1996.

M. J. F. Gales. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, 12:75–98, 1998.

M. J. F. Gales and P. C. Woodland. Mean and variance adaptation within the MLLR framework. *Computer Speech and Language*, 10:249–264, 1996.

S. Goronzy and R. Kompe. A combined MAP + MLLR approach for speaker adaptation. In *Proceedings of the the Sony Research Forum '99*, volume 1, pages 9–14, 1999.

D. Imseng, H. Bourlard, H. Caesar, P. N. Garner, G. Lecorvé, and A. Nanchen. MediaParl: Bilingual mixed language accented speech database. In *Proceedings of the 2012 IEEE Workshop on Spoken Language Technology*, pages 263–268, 2012.

Lori F. Lamel, Jean-Luc Gauvain, Maxine Eskenazi, and Maxine Eskenazi. BREF, a large vocabulary spoken corpus for french. In *Proceedings of Eurospeech-91*, pages 505–508, 1991.

Yuji Nakano, Makoto Tachibana, Junichi Yamagishi, and Takao Kobayashi. Constrained structural maximum a posteriori linear regression for average-voice-based speech synthesis. In *Proceedings of Interspeech 2006*, pages 2286–2289, 2006.

Olivier Siohan, Tor Andr Myrvoll, and Chin-Hui Lee. Structural maximum a posteriori linear regression for fast HMM adaptation. *Computer Speech and Language*, 16:5–24, 2002.

H. Wang, X. Zhang, X. Xiao, J. Zhang, and Y. Yonghong. Combining MAP and MLLR approaches for SVM based speaker recognition with a multi-class MLLR technique. In *Proceedings of the Second International Symposium on Information Science and Engineering (ISISE)*, pages 447–450, 2009.