N-BEST BASED SUPERVISED AND UNSUPERVISED ADAPTATION FOR NATIVE AND NON-NATIVE SPEAKERS IN CARS

P. Nguyen[1,2], Ph. Gelin[1], J-C. Junqua[1] and J-T. Chien[3,]*

[1]Panasonic Technologies Inc., Speech Technology Laboratory, Santa Barbara, California
[2]Institut Eurécom, Sophia-Antipolis, Cedex, France
[3]Depart. of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan
(nguyenp@eurecom.fr; gelin, jcj@research.panasonic.com; jtchien@mail.ncku.edu.tw)

1. ABSTRACT

In this paper, a new set of techniques exploiting N-best hypotheses in supervised and unsupervised adaptation are presented. These techniques combine statistics extracted from the N-best hypotheses with a weight derived from a likelihood ratio confidence measure. In the case of supervised adaptation the knowledge of the correct string is used to perform N-best based corrective adaptation. Experiments run for continuous letter recognition recorded in a car environment show that weighting N-best sequences by a likelihood ratio confidence measure provides only marginal improvement as compared to 1-best unsupervised adaptation and N-best unsupervised adaptation with equal weighting. However, an N-best based supervised *corrective* adaptation method weighting correct letters positively and incorrect letters negatively, resulted in a 13% decrease of the error rate as compared with supervised adaptation. The largest improvement was obtained for non-native speakers.

2. INTRODUCTION

Adaptation techniques can be classified into two main classes: 1) supervised and 2) unsupervised. While supervised techniques are based on the knowledge of the adaptation data transcriptions, unsupervised techniques determine the transcriptions of the adaptation data automatically using the best models available and consequently provide often limited improvements as compared to supervised techniques. Given a small amount of adaptation data, one of the common challenges of supervised adaptation is to provide adapted models which accurately match a user’s speaking characteristics and are discriminative. On the other hand unsupervised adaptation has to deal with inaccuracy of the transcriptions and the selection of reliable information to perform adaptation. For both sets of techniques it is important to adjust the adaptation procedure to the amount of adaptation data available.

Among the techniques available to perform adaptation, transformation-based adaptation (e.g. Maximum Likelihood Linear Regression or MLLR [1]) and Bayesian techniques such as Maximum A Posteriori (MAP) [2] adaptation are most popular. While transformation-based adaptation provides a solution for dealing with unseen models, Bayesian adaptation uses *a priori* information from speaker independent models. Bayesian techniques are particularly useful in dealing with problems posed by sparse data. In practical applications, depending on the amount of adaptation data available, transformation-based, Bayesian techniques or a combination of both may be chosen.

An N-best paradigm has been proposed in [3] for unsupervised adaptation. This method jointly optimizes control parameter sets and recognized word sequences. It was shown to be effective for "difficult" speakers with poor recognition rates. In their work Matsui and Furui showed that utterance verification can be useful to reduce the amount of calculation. In [4] the log likelihood ratio between the first and second candidate was used to select only reliable recognition results in an unsupervised adaptation scheme. This was aimed at improving the training efficiency.

In this paper, we present a new set of techniques exploiting N-best hypotheses in supervised and unsupervised adaptation. First, we clarify the trade-offs between transformation-based and Bayesian adaptation in the context of unsupervised adaptation when recognition accuracy on the adaptation data varies. Then, we propose a weighting scheme of N-best strings for unsupervised adaptation where weights are assigned to the N-best hypotheses according to a likelihood ratio confidence measure. Finally, we present a corrective adaptation procedure weighting incorrect models by a log likelihood ratio between the current and the best hypothesis and show that corrective adaptation outperforms supervised adaptation. Results are presented for both native and non-native speakers in the context of a continuous letter recognition task in cars.

3. UNSUPERVISED ADAPTATION USING MLLR AND MAP

3.1 Introduction to MLLR and MAP

MLLR, when used for adapting the emitting distribution mean vectors of the Hidden Markov Models (HMMs) can be written as an affine transformation :

$$\hat{\mu} = W\mu + b ,$$

[*] This work was done while J.-T. Chien visited Speech Technology Laboratory, U.S.A

where $\hat{\mu}$ and μ are respectively the adapted and original mean vector; W and b are the transformation matrix and bias derived to optimize the maximum likelihood through the optimization of Baum's "auxiliary function", Q [5]:

$$Q(\mu, \hat{\mu}) = \sum_{\theta \in \text{ states}} L(O, \theta|\mu) \log(L(O, \theta|\hat{\mu})),$$

where $L(O, \theta|\mu)$ stands for the likelihood of the observation O, and the sequences of states, θ, given the specific mean vector μ. In the following experiments involving MLLR one global matrix was used.

On the other hand, the MAP approach maximizes the *a posteriori* probability:

$$\mu_{MAP} = \arg\max_{\mu} \; f(\mu|O),$$

which for adaptation of the means reduces to:

$$\mu_{MAP} = \frac{\tau\mu_0 + \sum_t \gamma(t) o_t}{\tau + \sum_t \gamma(t)},$$

where τ is a measure of confidence on the prior ($\tau = 15$ in our experiments) and γ is the observed posterior probability of the observation.

The two techniques can be serialized, e.g. [6], i.e. one can apply MAP after MLLR. Doing so, it is possible to take advantage of the different properties of the two techniques.

3.2 Database

All the experiments conducted in this paper were done on continuous spelled names. The training data used to build the HMMs consists of telephone speech (1222 calls) extracted from the spelled name part of the OGI database [7]. The test data was recorded in two medium size cars with a Knowles 3310 close talking microphone. 10 speakers (6 native speakers of American English, comprised of 4 males and 2 females and 4 non-native speakers comprised of one female Japanese and 3 males, one French, one Italian and one Chinese) uttered 45 spelled street names at 60 mph leading to an overall test set of 3951 letters with an average of 8.8 letters per street name.

As adaptation data we used one repetition of the alphabet produced by each speaker in a continuous mode and in five sentences ("abcdef", "ghijkl", "mnopqr", "stuvw", and "xyz") when the car was parked. The data recorded in the two cars was sampled at 8kHz. We used Perceptually-based Linear Prediction (PLP) cepstral parameters (18 coefficients comprised of 8 static cepstral coefficients + energy and their delta) whose trajectories were bandpass filtered.

3.3 Experiments and Results

Unsupervised adaptation is difficult because the transcriptions used in the adaptation process can be unreliable. To assess how this factor affects both MLLR and MAP, we ran experiments where recognition accuracy on the adaptation data varied. Our speaker-independent HMMs led to an average of 60% recognition accuracy on the adaptation data. To vary the recognition accuracy on the adaptation data we randomly selected one out of the five adaptation sentences and corrected its transcription, leading to 70% unit accuracy. Iterating the process led to 78%, 85%, 94% and 100% recognition accuracy on the adaptation data.

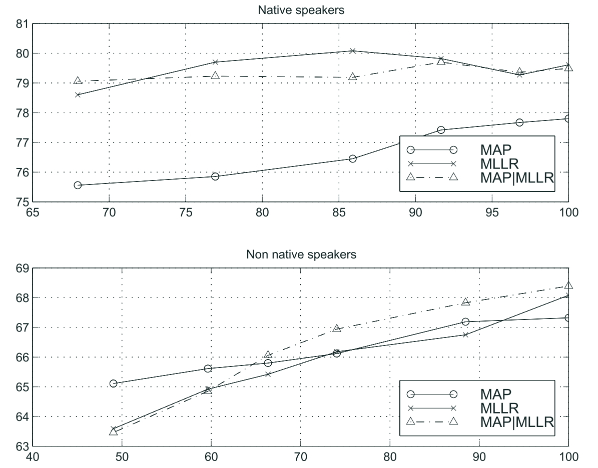

Figure 1: Sensitivity of MAP and MLLR techniques to recognition accuracy on the adaptation data for both native and non-native speakers

Figure 1. shows how MLLR, MAP and MLLR followed by MAP perform when the recognition accuracy on the adaptation data varies in the case of both native and non-native speakers. It can be seen that in the case of native speakers MLLR is less sensitive to the recognition accuracy on the adaptation data than MAP. For non-native speakers this tendency is not present. This may be due to errors which occur sometimes between two letters not belonging to the same confusable set. In this case the estimation of the MLLR matrix may not be reliable. As EM-based algorithms [8], MAP and MLLR both gather statistics for each mean. While MLLR further combines them to estimate another set of parameters shared by several means, MAP uses statistics directly to update the means. Therefore, in MLLR, the reliability of statistics is averaged into the estimation of the transformation matrix, whereas in MAP, the granularity is maximal and each parameter is associated to its seen statistics. It is thus clear that MLLR is best suited for unsupervised adaptation because the parameters are estimated globally combining all statistics, reliable and not reliable altogether. MAP, in contrast, updates model parameters on a per mean basis and hence appears as a good candidate for discriminative schemes.

4. UNSUPERVISED ADAPTATION USING N-BEST DECODING AND LIKELIHOOD RATIO WEIGHTING

The application of the EM-algorithm for supervised adaptation is well-known. For unsupervised adaptation, the mathematical solution is clear: we need to take the expectation over all possible word sequences in the grammar. This being in practice intractable, we only need to take into account the N-best hypotheses, based on the assumption that they accumulate most of the probability mass. We subsequently describe three types of approximations of the estimation. These approximations make use of the max operator instead of the expectation. The simplest unsupervised algorithm consists of using decoded labels as true labels; we call it 1-best unsupervised adaptation:

$$\hat{\lambda} = \arg\max_{\lambda} \max_{w, \theta} f(O, \theta, w|\lambda),$$

where $f(.)$ is the likelihood function, θ is the states sequence and w is the word sequence.

Another possibility is to maximize the likelihood of all N-Best strings as follows:

$$\hat{\lambda} = \arg\max_{\lambda} \sum_{w} \max_{\theta} f(O, \theta, w|\lambda).$$

We will refer to this technique as equal-weighting. Yet the optimal solution is

$$\hat{\lambda} = \arg\max_{\lambda} \sum_{w} \varphi_w \max_{\theta} f(O, \theta, w|\lambda),$$

where φ_w is the weight applied to the word sequence w.

Each N-best string w should be weighted according to its probability. Since we use non-normalized measures of likelihood, a straightforward approximation of the probability is given by:

$$\varphi_n = \exp([L_n - L_1]\eta),$$

where L_n is the log likelihood of the n-th best hypothesis, φ_n is by definition inferior or equal to 1 and η is a heuristic parameter that represents prior confidence on the decoded labels. When $\eta \to \infty$, the best hypothesis is expected to be correct and a 1-best adaptation is performed. If $\eta = 0$, then an equal weighting is applied to N-best hypotheses. Figure 2. shows sample weights for different values of the η parameter versus the rank of the word sequence n.

4.1 Experiments and Results

We ran experiments with a value of N equal to 5 and 3 iterations of the adaptation procedure. Table 1. shows how weighting N-best hypotheses by an exponential weight compares to the 1-best case and the N-best case with equal weighting. Table 1. reveals that exponential weighting provides only a marginal improvement. However, more experimentation is necessary to find the best weighting scheme. One possible explanation for the results obtained could be that there are too many similar strings on the N-best hypotheses and an N-best paradigm has a hard time to provide a significant difference.

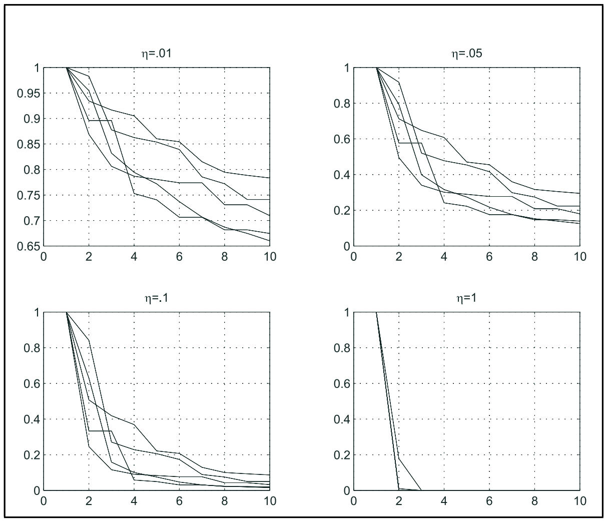

Figure 2: Exponential weighting versus the rank of the word sequence n

MLLR->MAP adaptation	η	Native	Non Nat.	Total
1-best	1	78.9%	64.7%	73.2%
N-best & exp weights	0.1	79.0%	64.2%	73.1%
//	0.05	79.4%	64.8%	73.5%
//	0.01	79.7%	65.2%	73.9%
N-best & Eq. weights	0	79.6%	65.0%	73.8%

Table 1: Unsupervised adaptation using N-best decoding (in unit accuracy)

5. SUPERVISED ADAPTATION USING CORRECTIVE N-BEST DECODING

When dealing with supervised adaptation, as the correct label sequence is known, the N-best information can be used in a discriminative way. During preliminary experiments we found that using the correct segmentation is crucial to adaptation and that the segmentation can easily vary between the N-best solutions. This led us to investigate corrective adaptation based on the segmentation produced by a forced alignment of the correct label sequence. In other words, for each segment produced by the correct segmentation, an N-best pass is done to collect the N most probable labels. These N-best labels are then used to adapt the model, either with a positive or a negative weight according to the following rule:

$$\varphi_n = \begin{cases} \kappa & \text{, if correct label} \\ -\rho e^{([L_n - L_1]\eta)} & \text{, otherwise,} \end{cases}$$

where κ represents the weight given to the supervised forced alignment. It is independent of n because we want to recover the correct label the same way whatever its rank is. L_n is the likelihood of the n^th best answer. ρ and η control the amount of backoff mis-recognized letters should receive. Ensuring that $\eta > 0$ and $\kappa > (N-1)\rho$ guarantees that for a given segment, the sum of all weights will be positive for MLLR and MAP, assuming the correct label is in the N-best. Typical values for these parameters are: $\kappa = 2$, $\eta = 0.01$ and $\rho = 0.3$.

As an iterative procedure over the adaptation data can be used to further improve the models, the global training protocol can then be summarized as follows:

```
For each iteration :
    For each recorded sentence of the adaptation set,
        Make a forced alignment according to the expected labeling of the sentence.
        For each aligned segment of the sentences, use an N-best decoder to get the N-best transcriptions and their corresponding likelihood ($L_1, L_2, \ldots, L_N$).
        Accumulate the adaptation of all the N-best transcriptions, according to their weights $\varphi_n$, $n = 1, \ldots, N$.
    Apply the adaptation.
```

5.1 Experiments and Results.

Table 2. shows a comparison between letter recognition accuracy obtained with speaker-independent models after unsupervised 1-best adaptation, supervised 1-best adaptation and corrective N-best adaptation. Corrective N-best adaptation decreases the error rate by 13% as compared to supervised adaptation. The improvement is larger for difficult speakers such as non-native speakers than for native American speakers.

	Native	**Non Nat.**	**Total**
Spkr. Ind.	75.6%	64.3%	71.1%
Unsupervised ($\eta = 0.01$)	79.7%	65.2%	73.9%
Supervised 1-best	79.5%	69.0%	75.3%
Corrective 5-best	81.7%	73.7%	78.5%

Table 2: Supervised adaptation using corrective N-best decoding (in unit accuracy)

In comparison with other discriminative methods, this corrective adaptation has several advantages. It operates on a small amount of data, is computationally inexpensive and is easy to implement. Moreover it carries out discrimination specific to speaker and in practice, convergence is not an issue. Note that anti-observations (those associated with a negative weight) can be regarded as additional observation that contribute to obtaining more reliable statistics.

6. CONCLUSIONS

In this paper, we presented new techniques for supervised and unsupervised adaptation. These techniques combine statistics extracted from the N-best hypotheses with a weight derived from a likelihood ratio confidence measure. While our experiments revealed that weighting N-best hypotheses in unsupervised adaption did not bring much improvement in the context of continuous letter recognition, corrective supervised adapation decreases the error rate by more than 13%. These techniques need to be further explored to derive optimal weighting schemes. They also need to be applied to other tasks where there is less confusability between words of the lexicon.

7. REFERENCES

1. C. Leggetter and P. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density HMMs”, *Comp. Speech Lang.*, 1995, vol 9, pp. 171-185.

2. J.-L. Gauvain and C.-H. Lee, “Maximum a posteriori estimation for multivariate gaussian mixture observation of Markov chains”, *IEEE Transactions on Speech and Audio Processing*, 1994, no 2, pp. 291-298.

3. T. Matsui and S. Furui, "N-best-based instantaneous speaker-adaptation method for speech recognition", *Proc. of ICSLP-96*, 1996, vol. III, pp. 973-976.

4. S. Homma, K. Aikawa and S. Sagayama, "Improved Estimation of Supervision in Unsupervised Speaker Adaptation", *Proc. of ICASSP-97*, vol. II, pp. 1023-1026.

5. L.E. Baum, “An inequality and associated maximization technique in statstical estimation for probabilistic functions of Markov processes”, *Inequalities 3*, 1972, pp. 1-8.

6. E. Thelen, X. Aubert, “Speaker adaptation in the Philipps system for large vocabulary continuous speech recognition”, *Proc. of ICASSP-97*, 1997, vol. 2, pp. 1035-1038.

7. R. Cole, K. Roginski and M. Fanty, “ A telephone speech database of spelled and spoken names”, *Proc. of ICSLP-92*, pp. 891-893.

8. N.M. Laird, A.P. Dempster and D.B. Rubin, “Maximum-Likelihood from Incomplete Data via the EM algorithm”, *Journal of the Royal Statistical Society B*, 1977, pp. 1-38.