CONFIDENCE-MEASURE-DRIVEN UNSUPERVISED INCREMENTAL ADAPTATION FOR HMM-BASED SPEECH RECOGNITION

Delphine CHARLET

France Télécom R&D
DIH/IPS, 2 av. Pierre Marzin
22307 Lannion Cedex - FRANCE
delphine.charlet@francetelecom.fr

ABSTRACT

In this work, we first review the usual ways to take into account confidence measure in unsupervised adaptation and then propose a new unsupervised incremental adaptation based on a ranking of the adaptation data according to their confidence measures. A semi-supervised adaptation process is also proposed: confidence measure is used to select the main part of the data for unsupervised adaptation and the remaining small part of the data is handled in a supervised mode. Experiments are conducted on a field database. Generic context-dependent phoneme HMMs are adapted to task- and field-specific conditions. These experiments show a significant improvement for unsupervised adaptation when confidence measure are used. In this work, we also show that the adaptation rate (that measures how important adaptation data are considered with respect to prior data) influences a lot the efficiency of the confidence measure in unsupervised adaptation.

1. INTRODUCTION

Unsupervised adaptation is required when no transcriptions of the adaptation data is available. In more and more speech recognition applications, the manpower and the delay required for manual transcriptions is prohibitive. Unfortunately, if the baseline performance of the recognizer is not high enough, many recognition errors occur, and as a consequence, the unsupervised adaptation can deteriorate some parts of the model. Therefore the idea of controlling the unsupervised adaptation process with a confidence measure on the recognized utterances is quite appealing [1] [2] [3] [4].

In this work, generic context-dependent phoneme HMMs are adapted to field- and task-specific conditions through a new unsupervised incremental adaptation approach: adaptation data are selected or weighted according to their ranked confidence measure. A semi-supervised adaptation scheme based on a splitting of the data between supervised and unsupervised adaptation is also proposed.

The paper is organized as follows: section 2 recalls incremental adaptation principles; section 3 presents the confidence measure used in this work; section 4 reviews the ways confidence measures are usually integrated in the adaptation process and proposes a new approach based on a ranking of the data according to their confidence measure. Section 5 presents and discuss the experimental evaluation of the proposed adaptation methods. Finally, section 6 proposes semi-supervised adaptation process.

2. UNSUPERVISED INCREMENTAL ADAPTATION

In previous work [5], incremental adaptation has shown to be a simple but very effective way to adapt a generic context-dependent phoneme HMM to field- and task-specific conditions. It is equivalent to MAP estimation with specific choice of priors. A comparative study made in [6] concerning the relative weighting of prior and new parameters, showed that such an adaptation gives very good results with simple prior weighting. This adaptation procedure is an iterative process, where, at each iteration, the following reestimation formulae are applied, for each Gaussian pdf:

$$\mu = \frac{\sum_{i \in Init} x_i + \sum_{j \in Adapt} x_j}{N_{Init} + N_{Adapt}} \quad (1)$$

$$\sigma^2 = \frac{\sum_{i \in Init} x_i^2 + \sum_{j \in Adapt} x_j^2}{N_{Init} + N_{Adapt}} - \mu^2 \quad (2)$$

where $\{x_i, i \in Init\}$ is the set of N_{Init} frames of the prior data associated to the pdf, and $\{x_j, j \in Adapt\}$ is the set of N_{Adapt} frames of the adaptation data aligned with the pdf.

In order to emphasize the influence of field data, a weighting factor on the prior distribution can be used. This weighting factor can be seen as a threshold on the prior weight

This work is partially supported by the SMADA project. The SMADA project is partially funded by the European Commission, under the Action Line Human Language Technology in the 5th framework IST Programme

$N_{MaxInit}$. The weight actually assigned to the prior parameters is then N_{prior}:

$$N_{prior} = \min(N_{Init}, N_{MaxInit}) \quad (3)$$

Thus, the weight assigned to the prior parameter can not exceed $N_{MaxInit}$. This weight threshold is a control parameter of the adaptation process. The lower $N_{MaxInit}$ is, the more the adaptation data are taken into account with respect to prior data. Then, the parameters of the pdfs are reestimated as if the prior parameters had been estimated with N_{prior} frames.

$$\mu = \frac{\frac{N_{prior}}{N_{Init}} \sum_{i \in Init} x_i + \sum_{j \in Adapt} x_j}{N_{prior} + N_{Adapt}} \quad (4)$$

3. CONFIDENCE MEASURE

The confidence measure used in this work is the difference between the log-likelihood of the first and second candidates in an N-best decoding approach, normalized by the length of the utterance. This measure is a classical confidence measure when N-best decoding is available, that has proven to be quite effective [7] [3]. For a given utterance X of T frames for which the first hypothesis is word W_i^1 and the second hypothesis is word W_j^2, the confidence measure is:

$$cm = \Delta_{llk} = \frac{log(P(X/W_i^1)) - log(P(X/W_j^2))}{T} \quad (5)$$

4. CONFIDENCE-MEASURE-DRIVEN UNSUPERVISED INCREMENTAL ADAPTATION

To cope with recognition errors that might occur in unsupervised adaptation and degrade model parameters accuracy, confidence measure can be used to control the way adaptation utterances are taken into account. The following formula presents the framework that is generally used for the integration of confidence measure, in the case of the pdf mean:

$$\mu = \frac{\frac{N_{prior}}{N_{Init}} \sum_{i \in Init} x_i + \sum_{j \in Adapt} w(x_j) x_j}{N_{prior} + \sum_{j \in Adapt} w(x_j)} \quad (6)$$

The function $w(x_j) = f(cm(x_j))$ is the weighting function on the accumulated counts for the adaptation data, this weight is a function of the confidence measure for the data. To design the function f, two basic ways appear. First, it can be an actual weighting function whose only requirement is to be monotonic and to increase with confidence measure values. When the confidence measure is mapped to a probability, the simplest way is to consider this probability as the weighting factor of the given utterance, as it is done in [2].

The second way to take into account confidence measure is a selection of the adaptation utterances whose confidence measures are above a defined threshold θ. This is done for instance in [1] [3] [4]. It can be seen as a particularly simple weighting function with:

$$\begin{aligned} w(x_j) &= 1 \text{ if } cm(x_j) > \theta \\ w(x_j) &= 0 \text{ if } cm(x_j) \leq \theta \end{aligned} \quad (7)$$

In this case of data selection, the upper limit of the integration process efficiency can be estimated: [1] and [4] have evaluated their unsupervised adaptation scheme in a so-called "cheated" framework, where they used a perfect confidence measure (only utterances that were correctly recognized were used in adaptation), so as to evaluate the best performance they can expect for confidence-measure-driven unsupervised adaptation.

In this work, we evaluate and compare these 2 kinds of integration of the confidence measure in an unsupervised adaptation process. In both approaches, the weighting function w is based on a ranking of the adaptation utterances according to their confidence measure. At each iteration, the confidence measure is computed for each utterance. Then, the utterances are ranked according to their confidence measure and they are either selected or weighted according to their rank.

In the framework of selecting adaptation utterances, the threshold is defined as follows:

- Threshold $T_{N\%}$ set to select the $N\%$ utterances best ranked according to their confidence measure.

Thus, only utterances having confidence measure above the threshold are used. An interesting point is to consider various thresholds during the adaptation. At the beginning of the process, the model may not be very accurate and may lead to a certain amount of recognition errors. Thus, it is better not to use a lot of utterances. However, as the adaptation goes on, the system is supposed to get better and better, so it is supposed to make less and less errors, hence a threshold that enables more utterances to be selected can be used. In our selection scheme, we use a succession of 3 blocks of iterations using respectively $T_{50\%}$, $T_{75\%}$ and $T_{90\%}$.

In the framework of a weighting of adaptation utterances, the proposed weighting function for utterance X is the following one:

- $w(X) = i/10$
  where $\{i \in 1, 2, .., 10\}$ is found as:
  $T_{10(i-1)\%} \leq cm(X) < T_{10i\%}$

This weighting function leads to a weight of 1 for the 10% best utterances and a weight of 0.1 for the 10% worst utterances.

5. EXPERIMENTS

5.1. Experimental setting

Experiments have been conducted on a field database collected from a telephone voice portal of 256 entries. This database can be divided into 4 sub-corpora:

- corpus A: 1112 utterances of entry names
- corpus B: 1046 utterances of noise tokens
- corpus C: 235 utterances of out-of-vocabulary tokens (OOV tokens)
- corpus D: 128 utterances of out-of-vocabulary requests (OOV requests)

The difference between corpus C and corpus D is that people in corpus D pronounce names that they believe to be present in the application, whereas in corpus C, they don't speak directly to the server (mainly hesitations, comments or private conversations). False alarm rate are much higher for corpus D than for corpus C. This database is further divided into 2 equally balanced corpora. One is used for adaptation, the other for test.

The recognizer is speaker independent HMM-based and observation functions are continuous multi-gaussian densities (8 gaussians per density). The acoustic modelling of words is based on contextual phones, silence and garbage models. Adaptation is made on contextual phones and garbage models.

5.2. Results and discussion

Figure 1 presents the ROC curves for the various corpora on test data, for different adaptation processes refered as:

- initial: the initial model (generic phone-library)
- supervised: after supervised adaptation
- unsupervised: after unsupervised adaptation
- selecting data: after unsupervised adaptation with selection of utterances according to their confidence measure
- weighting data: after unsupervised adaptation with weighting of utterances according to their confidence measure

The results presented in figure 1 are obtained with a threshold on the prior parameter (see section 2): $N_{MaxInit} = 100$. These results are encouraging as they show that the integration of confidence measure leads to a significant improvement with respect to unsupervised adaptation. When confidence measure are used, recognition performances stand between those obtained with supervised adaptation and those obtained with unsupervised adaptation. Moreover, selecting data appears to be more effective than the particular weighting function we experimented.

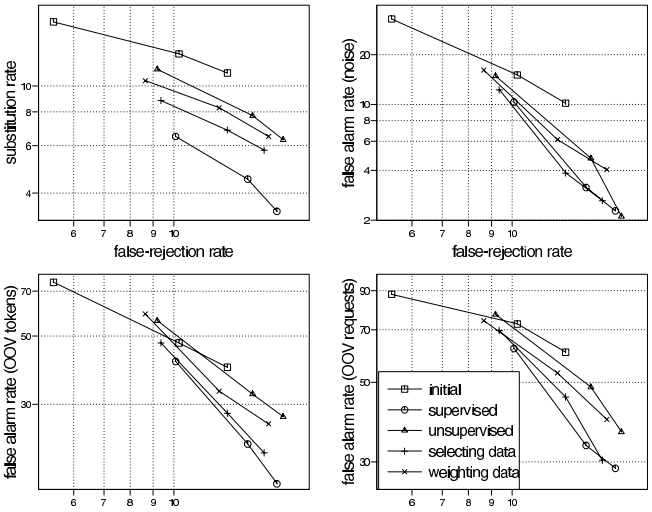

Fig. 1. Recognition performances for different adaptation processes, with $N_{MaxInit} = 100$

A very interesting point is the behaviour of such adaptation processes when the prior threshold of the incremental adaptation is modified. Figure 2 shows the performances achieved when there is no threshold on prior data. In this case, unsupervised adaptation and confidence-measure-driven supervised adaptation lead to similar performances. The following interpretation is done: when there is a threshold on prior parameters, the importance of adaptation data is increased and the sensitivity to recognition errors in adaptation data is very important. When there is no threshold on prior parameters, adaptation data are less important and sensitivity to recognition errors is not high. Hence, the influence of the confidence measure on adaptation performances depends a lot on the adaptation process itself. This fact might explain the differences observed in the literature concerning performances of confidence-measure-driven unsupervised adaptation (e.g. [4] reports improvement, [2] reports a very small improvement and [1] reports degradation).

6. SEMI-SUPERVISED INCREMENTAL ADAPTATION

For unsupervised adaptation with selection of data according to their confidence measures, some data are discarded from the adaptation process. However, these data may be very useful to adapt model parameters. Indeed, if the confidence measure is effective, the data with the lowest confidence measures are the data more likely to lead to recognition errors, because they don't fit with the model parameters. Hence, it may be interesting to take them into account in a supervised adaptation. The adaptation data that require manual transcriptions for supervised adaptation are

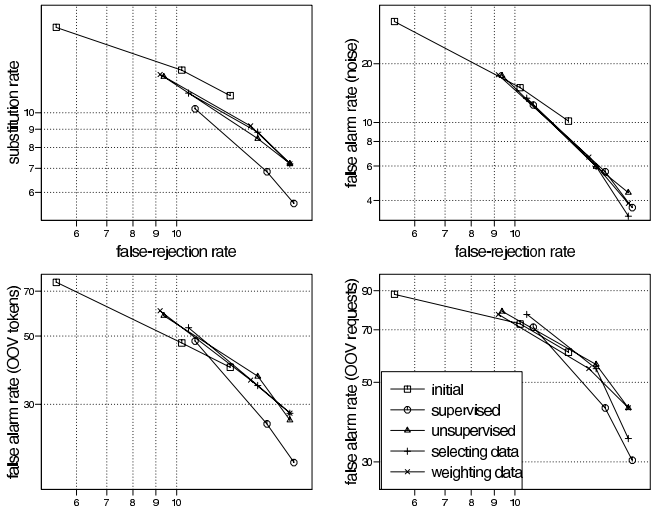

Fig. 2. Recognition performances for different adaptation process, with no threshold on prior data

the discarded data of the unsupervised adaptation process. In our experiments, this amount represents 10% of the total amount of data, and it remains feasible to get manual transcriptions for this little part of data. The following semi-supervised adaptation is thus proposed:

1. Do confidence-measure-driven unsupervised adaptation with selection of data
2. For the data discarded from adaptation process because of their low confidence measure, do manual transcriptions
3. Do supervised adaptation of the model obtained after step 1

Figure 3 presents results for substitution rate with semi-supervised adaptation, when unsupervised adaptation is made with a threshold on prior data ($N_{MaxInit} = 100$). The supervised adaptation step with few data that follows the confidence-measure-driven unsupervised adaptation step does not change significantly the model accuracy (results on false alarm rate for the other corpora confirm this conclusion). This may be due to the relative weight of prior and adaptation data. Other ways to merge supervised adaptation with few data and unsupervised adaptation with a lot of data should be investigated.

7. CONCLUSION

This paper focuses on the integration of a confidence measure to control unsupervised incremental adaptation. Integration based on a ranking of the adaptation data according to their confidence measure is proposed. Evaluations show that significant improvement compared to unsupervised adaptation can be obtained with a selection of adaptation data according to their ranked confidence measure.

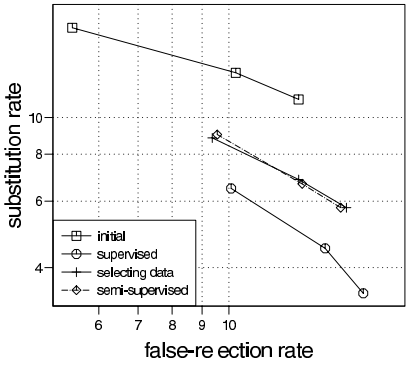

Fig. 3. Recognition performances for semi-supervised adaptation (with unsupervised adaptation done with $N_{MaxInit} = 100$)

This work experimentally shows that the influence of the confidence measure depends also on the adaptation rate: in our work, influence of the confidence measure depends on the relative weight of the prior data with respect to the adaptation data.

8. REFERENCES

[1] S. Goronzy, K. Marasek, A. Haag, and R. Kompe, "Prosodically motivated features for confidence measures," in *ASR2000*, Paris, France, 2000, pp. 207–212.

[2] Y. Gao, B. Ramabhadran, and M. Picheny, "New adaptation techniques for large vocabulary continuous speech recognition," in *ASR2000*, Paris, France, 2000, pp. 107–111.

[3] Shigeru Homma, Kiyoaki Aikawa, and Shigeki Sagayama, "Improved estimation of supervision in unsupervised speaker adaptation," in *ICASSP'97*, Munich, Germany, 1997, pp. 1023–1026.

[4] T. Kemp and A. Waibel, "Unsupervised training of a speech recognizer: Recent experiments," in *Eurospeech'99*, Budapest, Hungary, 1999, pp. 2725–2728.

[5] N. Moreau, D. Charlet, and D. Jouvet, "Confidence measure and incremental adaptation for the rejection of incorrect data," in *ICASSP'00*, Istanbul, Turkey, 2000, pp. 1807–1811.

[6] C. Vair, M. Mercogliano, and Fissore L., "Incremental training of cdhmms using bayesian learning," in *Eurospeech'99*, Budapest, Hungary, 1999, pp. 2753–2756.

[7] D. Willet, A. Worm, C. Neukirchen, and G. Rigool, "Confidence measures for hmm-based speech recognition," in *ICSLP'98*, Sidney, Australia, 1998, pp. 525–528.