Experiments on Acoustic Model supervised adaptation and evaluation by K-Fold Cross Validation technique

Daniel R.S. Caon*, Asmaa Amehraye†, Joseph Razik‡, Gérard Chollet*, Rodrigo V. Andreão§ and Chafic Mokbel¶
*Institut TELECOM, CNRS LTCI TELECOM ParisTech, Paris, France
Email: caon,chollet@telecom-paristech.fr
†ESIGETEL, LRIT laboratory, Avon, France
Email: asmaa.amehraye@esigetel.fr
‡Université du Sud-Toulon-Var, Toulon, France
Email: Joseph.Razik@univ-tln.fr
§Coordenadoria de Eletrotécnica, CEFETES, Vitória, ES, Brazil
Email: rodrigova@ifes.edu.br
¶Mathematics Department, University of Balamand, 100 El-Koura, Lebanon
Email: chafic.mokbel@balamand.edu.lb

***Abstract*—This paper is an analysis of adaptation techniques for French acoustic models (hidden Markov models). The LVCSR engine Julius, the Hidden Markov Model Toolkit (HTK) and the K-Fold CV technique are used together to build three different adaptation methods: Maximum Likelihood *a priori* (ML) , Maximum Likelihood Linear Regression (MLLR) and Maximum *a posteriori* (MAP). Experimental results by means of word and phoneme error rate indicate that the best adaptation method depends on the adaptation data, and that the acoustic models performance can be improved by the use of alignments at phoneme-level and K-Fold Cross Validation (CV). The very known K-Fold CV technique will point to the best adaptation technique to follow considering each case of data type.**

***Keywords*-speech recognition;maximum likelihood linear regression;maximum a priori;k-fold cross validation;**

I. Introduction

The speech recognition systems have faced many challenges. The bigger ones were the hardware capacities for data processing. Nowadays, the main problem is adaptation to the speaker [1], to the speech type (spontaneous conversations, read speech, debates etc), to the channel (microphone), to the environment and to the dialect. This work is conducted within the CompanionAble Project [1], with the aim of acquiring the best acoustic models for very specific scenarios. The objective of this work is to find the best adaptation technique (a common research in the area of automatic speech recognition, e.g. [2]) for the French Acoustic Models (HMM's) trained on ESTER [3] broadcasting news database. The French acoustic models are the adaptation targets in this paper. In section II the objective of this work, the acoustic models and all databases used in the experiments are described. Different databases are tested and the adaptations are supervised. The use of a K-Fold CV aims to provide more reliable evaluation of the results. The experiments are conducted in a lower level language unit using forced phoneme alignment by Viterbi. This is known as a good tool for identifying the actual pronunciation [4] contained in the utterances (although the best matching pronunciation must be previously listed in the lexicon). The Maximum Likelihood (ML) reestimation is used with two different configurations: The first updates transitions, means, variances and weights (ML tmvw). The second configuration doesn't update the transitions (ML mvw). The Maximum *a posteriori* (MAP) and The Maximum Linear Likelihood Regression (MLLR) adaptation methods are the other tested techniques. It is known that HTK's MAP implementation [4] does not update transition probabilities while the means (m), the variances (v) and the weights (w) are acceptable options. The explanations about the HTK's implementations are documented in [4]. MLLR is used in a static two-pass adaptation approach, which is described further. A K- Fold CV is used to evaluate the experiments. More details about this technique will be introduced later in subsection IV-A. As it is not only a matter of adaptation, but also to have a good evaluation method, the phone or the word aligned comparison with references are studied in V-A. The Section III explains the main features of MAP and MLLR adaptation approaches. The experimental protocol is described in section IV and section V opens a discussion towards the level of alignments (recognition output) for the validation of supervised adaptations, also showing the experimental results. The conclusion comes in section VI.

[1]http://www.companionable.net/

II. Databases, Models and Objectives

Three different types of adaptation databases are employed in this work: Readings, Interviews and Distress Situations. The ESTER [3] database is recorded on French broadcasting news and around 40 hours were used to train the acoustic models (hidden Markov models composed by monophones

978-1-4244-5997-1/10/$26.00 ©2010 IEEE

containing 5 states and 256 mixture components per state). The language model is based on 3-gram probabilities from large newspaper data (”Le Monde”) and the dictionary composed of 65k words.

For the adaptation databases, the speakers are non-native French, which permits some analysis on non-native speech adaptation. The Speaker Dependent Interviews (SDI) are recorded by two non-native speakers (SDIm by one male speaker and SDIf by one female). The FDE contains French Distress Expressions [2] recorded by 19 native and 2 non-native speakers in distress situations. More information are summarised in table I.

Database	Utterances	Words	Speakers	Speech Type
SDR	162	1572	1	Read Text
SDIf	103	521	1	Interview
SDIm	103	521	1	Interview
FDE	2646	10080	21	Distress Exp.

TABLE I
DATABASE INFORMATION

III. MAP *versus* MLLR

A. MAP

The MAP adaptation has a capability of achieving performance near to speaker dependent) systems [1]. It uses effective combination of prior knowledge, i.e. the initial model parameters, and ML estimates obtained on the adaptation data [1][5]. The MAP adaptation can also deal with foreign accents where often some phonemes differ a lot from the usual pronunciation while other phonemes don’t [1][5]. The main disadvantage of MAP is the large amount of adaptation data needed before all phonemes can be updated. This adaptation may be hard for those phonemes which do not appear very frequently in the adaptation database.

B. MLLR

Unlike MAP, the MLLR adaptation updates many (or even all) gaussians at once. One global regression class concerning all gaussians can be used (the fast case of MLLR). The basic principle is to calculate one or several transformation matrices from the adaptation data. Clustering several gaussians into regression classes that share the same transformation or regression matrix creates the possibility to update even the not observed parameters (i.e. a parameter that is not observed will join the group of an observed one, the one which has nearer acoustic features, and use the same transformation). The MLLR can quickly adapt the acoustic model to new speakers, environments, channel etc requiring only few adaptation data. One limitation of MLLR is about the foreign accents: Some phonemes will differ a lot from the usual pronunciation and the use of a transformation matrix in the specific regression class of this phoneme may not be appropriated for acquiring better results.

[2]Authors thank ESIGETEL (http://www.esigetel.fr/) for providing the French Distress Expressions database.

IV. EXPERIMENTAL PROTOCOL

A. Validation Technique

Validation Techniques have two main problems in the pattern recognition research area: Model selection and performance validation. This paper aims to validate the performance (recognition accuracy) of adapted acoustic models by taking into consideration the transcription level of references (detailed in V-A). A K- Fold CV was implemented. The variable K is equal to 20 to make the experiments, which means that every time about 5% of the adaptation data is tested while 95% (the other K- 1 parts) are used for adaptation. The K- Fold CV technique is commonly used to give more accurate evaluation results. K has to be chosen accordingly to the database size due to the desired computational time issues and the expected bias for the true error rate.

B. Two-Pass MLLR

The MLLR adaptation can be supervised (using labeled adaptation data) or not (labeling adaptation data by recognition before adapting the target). In this work, we used a supervised MLLR, with two-pass static adaptation by means of HERest (HTK tool) [3]. The first one of the two-pass steps builds a global transformation class, while the second pass builds a (multiple) regression transformation class [4].

V. EXPERIMENTAL RESULTS

The experiments start by analyzing the best forced alignment options. The alignment type will directly affect the observed results A 20-Fold CV Technique (for the chosen forced alignment option) is done for the adapted acoustic models. The figure 1 explains how the experiments use a supervised K-Fold CV adaptation.

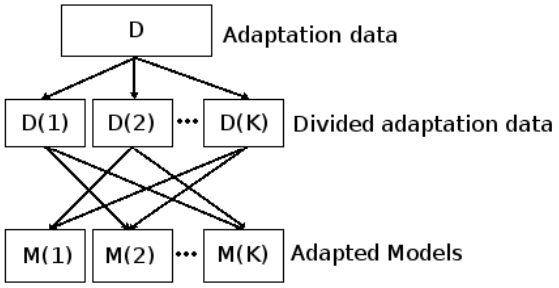

Fig. 1. Supervised K-Fold CV adaptation. D is the adaptation database, D(k) is the k-th data subset, M(k) is the adapted acoustic model obtained using all the adaptation data available in D except the k-th part D(k) which is a test database.

The 20-Fold CV is used to give the mean of recognition accuracy for the not-adapted (original models trained on ESTER corpus) and the adapted acoustic models (with ML, MLLR and MAP options).

[3]HERest is the main HTK’s training tool. It performs a single re-estimation of all HMMs, simultaneously. It uses the Forward-Backward algorithm to store statistics of state occupation, means, variances etc.

A. Forced Alignment Options

To choose the best language unit level to evaluate the adaptation, the SDR database is used to do a ML re-estimation on the ESTER's acoustic models. The other databases could be used to evaluate the forced alignment options too, but SDR was chosen because it is not as large as FDE and not as short as the SDI. Also, a ML re-estimation should give good results with the SDR database, due to the utterance's repetitions and the sufficient data to cover all the acoustic model set. For building the phoneme reference transcriptions from word reference transcriptions, the procedure is simple: The first occurrence of possible pronunciation for each word found in the dictionary is chosen. This makes the reference files not always compatible with what will be truly spoken. For example the French *liaisons* are spoken in an random mood (the speaker sometimes makes the *liaisons*, and sometimes it doesn't). This problem could be solved by a manual review of the reference transcriptions.

In this work, the experiments employ an adaptation procedure of transcriptions which permits to fix some mispronunciations (if provided in the lexicon, the best pronunciation can be chosen by HTK's Viterbi tool), but the reference file construction (considering only the first pronunciation on the lexicon of each word) is not very flexible. This way, the adaptation is conducted with no problems unlike the evaluation results which may be affected by the variability of the lexicon's pronunciations for a specific word. At the same time, for evaluating results with a higher language unit level like words, the chance of error is higher in the hypothesis, due to the impact of the language model probabilities (n-grams) to make the system fail when putting the phonemes together to compose the words, even if it recognizes the right phonemes before word aligning. If an utterance contains "there for" (spoken fasten as "therefore", with no short silence interval between the words), the pronunciation will be "DH EH R F AO R", the same for the word "therefore". Then, the hypothesis depends on the language model probabilities for choosing if the recognized output (word aligned) will be "there for" or "therefore".

The results considering word and phoneme forced alignments are presented in figure 2. The Phoneme Error Rate(PER) is used for phoneme alignment and the Word Error Rate(WER) is used for the word alignment. The results confirm a better recognition rate for phoneme aligning (PER is lower than WER). It cannot prove that phone alignment is better than word alignment for evaluation though. The last column shows the error rate mean of all folds together. The goal is to provide a more reliable information by the use of the K-Fold CV, instead of taking just one random fold to validate. Although the inflexibility of the reference transcriptions at phoneme level (as described before) the experiments results are given in Phone Error Rates (PER) and should provide good observations about adaptation gains. The PER makes us more independent from the language model probabilities. This is a very good aspect as we assume to analyze only the acoustic model adaptation.

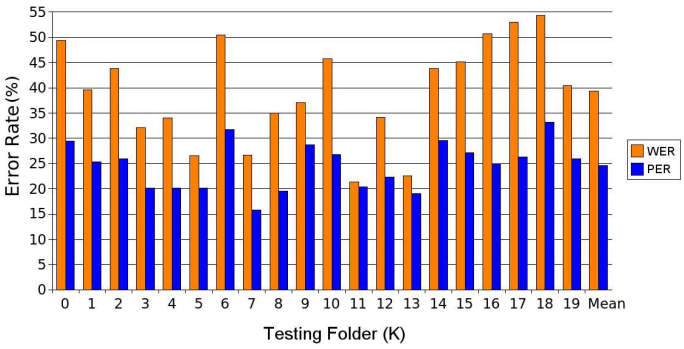

Fig. 2. SDR's comparison of error rates after ML re-estimation with phoneme or word alignment

B. PER 20-Fold Cross Validation

The tests are conducted using the original HMMs trained on the ESTER database (Not-Adapted) and the HMMs adapted with different techniques (ML, MAP and MLLR).

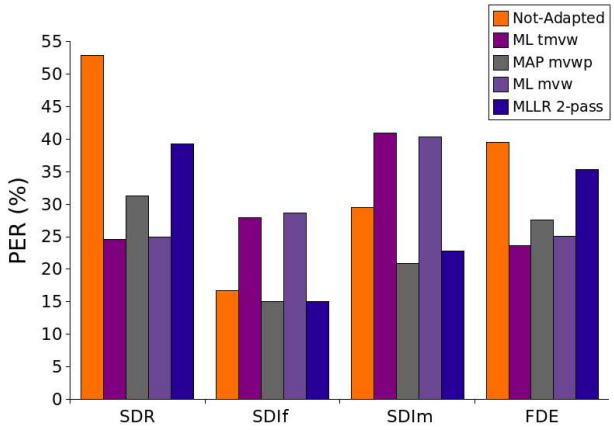

Fig. 3. Adaptation results for different techniques.

The results are shown by means of the 20 folds (results added and divided by 20, the total fold number) for easier visualization. Due to the high variation of the recognition accuracy results for each part, the mean of the K recognition tests is an information more useful than taking only a random part which explains the use of K-Fold CV. It can be noticed from figure 3 that ML has a better impact in the SDR and FDE databases than MAP or MLLR. This is explained by the fact that the SDR and the FDE have many utterance's repetitions (around 5 for SDR and 21 for FDE). For the SD Interview databases , a better accuracy is observed when doing MAP or MLLR than when doing ML re-estimation.

VI. Conclusion

The use of phoneme's alignments is recommended for the evaluation of acoustic models adaptation. It is still better than considering, for example, the comparison of two words with the same pronunciation as being mismatched. The adaptation method should be chosen according to the data available. If the data is sufficient for covering the acoustic space or there

are mispronunciations (like in foreign accents), MAP is better. With enough statistical information about the acoustic space and not too much mispronunciations, MLLR is effective even with short adaptation data. The K-Fold CV points to the best technique to use in each case and solves the choice of method for further adaptation iterations.

ACKNOWLEDGMENT

This work was conducted in the framework of the IST, FP7 Integrated Project CompanionAble (http://www.companionable.net/).

REFERENCES

[1] S. Goronzy and R. Kompe, in *A Combined MAP + MLLR Approach for Speaker Adaptation*, vol. 9, 2000, pp. 9–14.

[2] Z. Wang, T. Schultz, and A. Waibel, "Comparison of acoustic model adaptation techniques on non-native speech," in *in ICASSP 2003. IEEE*. IEEE, 2003, pp. 540–543.

[3] S. Galliano, E. Geoffrois, G. Gravier, J.-F. Bonastre, D. Mostefa, and K. Choukri, "Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news." LREC, 2006, pp. 315–320.

[4] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department, 2006.

[5] G. Zavaliagkos, R. Schwartz, and J. McDonough, "Maximum a posteriori adaptation for large scale hmm recognizers," *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 2, pp. 725–728, 1996.