Evaluation of several Maximum Likelihood Linear Regression variants for language adaptation

Míriam Luján, Carlos D. Martínez, Vicente Alabau

Departament de Sistemes Informàtics i Computació
Institut Tecnològic d'Informàtica
Universitat Politècnica de València
Camí de Vera, s/n. 46071 València, Spain
{mlujan, cmartine,valabau}@dsic.upv.es

Abstract

Multilingual Automatic Speech Recognition (ASR) systems are of great interest in multilingual environments. We studied the case of the *Comunitat Valenciana* where the two official languages are *Spanish* and *Valencian*. These two languages share most of their phonemes, and their syntax and vocabulary are also quite similar since they have influenced each other for many years. We constructed a system, and trained its acoustic models with a small corpus of Spanish and Valencian, which has produced poor results due to the lack of data. Adaptation techniques can be used to adapt acoustic models that are trained with a large corpus of a language inr order to obtain acoustic models for a phonetically similar language. This process is known as language adaptation. The Maximum Likelihood Linear Regression (MLLR) technique has commonly been used in speaker adaptation; however we have used MLLR in language adaptation. We compared several MLLR variants (mean square, diagonal matrix and full matrix) for language adaptation in order to choose the best alternative for our system.

1. Introduction

Multilingual Automatic Speech Recognition (ASR) systems are of great interest in multilingual environments (Uebler, 2001). In a multilingual environment, where each potential user has a different native language, a Multilingual Automatic Speech Recognition System must deal with different languages and with the inappropriate pronunciation of a language by non-native speakers. Multilingual environments are common in contexts where there are different languages due to politics, tourism, inmigration, and so on.

To build an ASR system, language and acoustic models must be trained. Language models are both task-dependent and language-dependent. For this reason, a multilingual ASR system must include several language models (one for each language that the system must deal with). Acoustic models are also language-dependent because each language defines its own phonemes differently from other languages. In multilingual environments, the influence of the speaker's mother tongue can produce an inappropriate pronunciation of the other languages that are present in these environments. This fact adds a new source of variability and makes accurate speech recognition difficult. Another important problem is that languages are usually influenced by other languages. In addition, the articulation of the same phoneme in different languages may differ in each language. All these facts demonstrate the difficulty of building multilingual speech recognition systems.

We studied the *Comunitat Valenciana*, a multilingual environment that has two official languages: *Spanish* and *Valencian*. Spanish and Valencian have similar phonetic features. They both have a similar set of phonemes, and most acoustic models could be shared by both languages. In this multilingual environment, it is useful to have a multilingual ASR system. In a previous work, we constructed an ASR for this environment (Lujan et al, 2007).

Valencian is a dialect of *Catalan* that is spoken in the *Comunitat Valenciana*. The Valencian dialect has special phonetic features with respect to standard Catalan. This is due to its dialectal variance and the great influence that Spanish has had on it. This influence has been much greater on Valencian than on other Catalan dialects. According to (Vilajoana and Pons, 2001), the total number of people who speak Catalan is 7,200,000, and the number of people who understand it is over 9,800,000. The Valencian dialect is spoken by 27% of all Catalan speakers. The great influence of the Spanish language on the Valencian dialect has modified Valencian phonetics in the average speaker. We constructed an ASR system for this environment even though the training corpora were quite small. However, it is well-know that a large corpus is necessary to build a good ASR and to obtain reliable acoustic models.

One way to obtain better acoustic models would be to use adaptation techniques to adapt acoustic models of a phonetically similar language (Schultz and Waibel, 2001). These models are usually trained with a large acoustic corpus. The resulting adapted models are commonly better than those trained with a small corpus of the real language.

The Maximum Likelihood Linear Regression (MLLR) (Leggetter and Woodland, 1995) technique has been commonly used in speaker adaptation. However, in a few works it has also been applied to language adaptation (Zhao and O'Shaughnessy, 2007). We used MLLR in language adaptation to adapt acoustic models. We tried to obtain acoustic models in Spanish and Valencian from only Spanish acoustic models. These original models were trained with large amounts of training data in Spanish because it was easy to find large Spanish corpora. Unfortunately, we did not have

Work partially supported by VIDI-UPV under PAID06-20070315 program and by the EC (FEDER) and the Spanish MEC under grant TIN2006-15694-CO2-01.

a large Valencian corpus available. However, since Valencian is phonetically very similar to Spanish, we were able to use the small amount of the original training material available in Valencian to adapt the Spanish acoustic models.

Our objective is to compare the different MLLR variants in order to choose the best alternative for this pair of languages.

This work is organized as follows. In Section 2., we describe the MLLR adaptation technique. In Section 3., we describe the corpus that was used for the experiments. In Section 4., we explain the design of the language models and the acoustic models. In Section 5., we detail the different experiments carried out. In Section 6., we present our conclusions and future work.

2. The MLLR adaptation technique

The aim of speaker adaptation techniques is to obtain a speaker-dependent recognition system by using a combination of general speech knowledge from well-trained hidden Markov models and speaker-specific information from a new speaker's data. Speaker adaptation is applied to a speech recognition system in order to obtain a speaker-dependent system with a better performance than the original system for a specific speaker.

MLLR is a technique to adapt a set of speaker-independent acoustic models to a speaker by using small amounts of adaptation material. This technique can also be used in language adaptation by using adaptation material in the language being adapted to.

The MLLR approach requires an initial independent continuous density HMM system. MLLR takes some adaptation data from a speaker to adapt the acoustic models. MLLR updates the model mean parameters to maximize the likelihood of the adaptation data. The means are updated using a transformation matrix, which is estimated from the adaptation data. We applied the formulation presented in (Leggetter and Woodland, 1995) to language adaptation: we take some adaptation data from a language to adapt the acoustic models using MLLR in the same way as in speaker adaptation.

The theory is based on the concept of regression classes. A regression class is a set of mixture components that share the same transformation matrix, which is estimated from the adaptation data. When all the models are in the same regression class, we have a global regression class. However, any set of regression classes can be manually or automatically defined over the gaussians of the HMM. There is no method to analytically determine the optimal number and composition of regression classes (Leggetter and Woodland, 1995).

To perform the adaptation of the means, we computed a transformation matrix $\vec{W}$ for each regression class. This matrix is applied to the extended mean vector of all the mixtures pertaining to the regression class to obtain an adapted mean vector. Given a state q in a HMM, for the ith gaussian of the output distribution, we denote its mean vector as $\vec{\mu}_{qi}$. The adapted mean vector $\widehat{\vec{\mu}}_{qi}$ is obtained by:

$$\widehat{\vec{\mu}}_{qi} = \vec{W} \cdot \vec{\xi}_{qi}$$

where $\widehat{\vec{\mu}}_{qi}$ is the adapted mean and $\vec{\xi}_{qi}$ is the extended mean vector defined as:

$$\vec{\xi}_{qi} = [w, \mu^0_{qi}, \ldots, \mu^n_{qi}]' = [w : \vec{\mu}_{qi}]$$

where n is the number of features, $\vec{\mu_{qi}}$ is the original mean vector and w is an offset term.

If we have a set of adaptation data denoted by the sequence of acoustic feature vectors $\vec{X} = \vec{x}_1\vec{x}_2...\vec{x}_T, \vec{x}_t \in \mathbb{R}^D, t = 1, ..., T$, we can estimate the adaptation matrix $\widehat{\vec{W}}$ using the maximum likelihood approach as:

$$\widehat{\vec{W}} = \max_{\vec{W}} p_{\vec{\theta}}(\vec{X})$$

where $\vec{\theta}$ defines the parameters of the adapted model.

To compute the transformation matrix, we can use several variants: without the same covariances of the distributions (with a full or a diagonal matrix) or with the same covariances of the distributions. Details on the estimation of these variants can be consulted in (Leggetter and Woodland, 1995). The following formulation assumes only one adaptation sample, but it can be easily extended for n adaptation samples.

2.1. Full matrix

Given a state q in a HMM, for the ith gaussian of the output distribution, we denote its mean vector as $\vec{\mu}_{qi}$ and its covariance matrix as $\vec{\Sigma}_{qi}$.

To compute a full matrix, it is necessary to compute an auxiliary tridimensional matrix $\vec{G}$. In this case, $\widehat{\vec{W}}$ must be calculated by rows because $\vec{G}$ is a tridimensional matrix. We calculate the row k of $\widehat{\vec{W}}$ as:

$$\widehat{\vec{w}}'_k = \vec{G}^{(k)-1} \vec{z'}_k$$

where

$$\vec{z} = \sum_t \sum_q \sum_i (\gamma_{qi}(t)) \vec{\Sigma}^{-1}_{qi} \vec{x}_t \vec{\xi'}_{qi}$$

$\gamma_{qi}(t)$ is defined as the posteriori probability of occupying state q at time t given that the observation sequence $\vec{X}$ is generated by the ith gaussian.

The row k of $\vec{G}$ is defined as:

$$\vec{G}^{(k)}_{jq} = \sum_{q,i} \vec{v}^{(qi)}_{ii} \vec{d}^{(qi)}_{jq}$$

where

$$\vec{D}^{(qi)} = \vec{\xi}_{qi} \vec{\xi'}_{qi}$$

and

$$\vec{V}^{(qi)} = \sum_t \gamma_{qi}(t) \vec{\Sigma}^{(-1)}_{qi}$$

2.2. Diagonal matrix

To compute a $\widehat{\vec{W}}$ transformation matrix, we define a diagonal matrix:

$$\vec{W} = \begin{bmatrix} w_{1,1} & w_{1,2} & 0 & \cdots & 0 \\ w_{2,1} & 0 & w_{2,3} & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ w_{n,1} & 0 & \cdots & 0 & w_{n,n+1} \end{bmatrix}$$

For non-zero elements of this matrix, we rewrote the matrix to a transformation vector $\widehat{\vec{w}}$ as:

$$\vec{w} = \begin{bmatrix} w_{1,1} \\ \vdots \\ w_{n,1} \\ w_{1,2} \\ \vdots \\ w_{n,n+1} \end{bmatrix}$$

We defined a matrix $\vec{D_{qi}}$ made up of the elements of the extended mean vector $\vec{\xi_{qi}}$ as:

$$\vec{D_{qi}} = \begin{bmatrix} w & 0 & \cdots & 0 & \mu_{qi1} & 0 & \cdots & 0 \\ 0 & w & \ddots & \vdots & 0 & \mu_{qi2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & w & 0 & \cdots & 0 & \mu_{qin} \end{bmatrix}$$

Then, $\widehat{\vec{w}}$ can be calculated as:

$$\widehat{\vec{w}} = \left[\sum_t \sum_q \sum_i (\gamma_{qi}(t)) \vec{D}'_{qi} \vec{\Sigma}^{-1}_{qi} \vec{x}_t \right]^{-1}$$

$$\left[\sum_t \sum_q \sum_i (\gamma_{qi}(t)) \vec{D}'_{qi} \vec{\Sigma}^{-1}_{qi} \vec{D}_{qi} \right]$$

2.3. Mean Square

We can consider that all covariances of the distributions are the same. We can follow the Viterbi approximation, where each speech frame is assigned to exactly one distribution. Therefore, the adaptation matrix can be computed by:

$$\widehat{\vec{W}} = \left(\sum_t \vec{x}_t \vec{\mu}'_{qi} \right) \left(\sum_t \vec{\mu}_{qi} \vec{\mu}'_{qi} \right)^{-1}$$

The sequence of $\vec{\mu}_{qi}$ is usually defined by a Viterbi alignment of the samples.

3. Corpus

To perform the experiments, we employed a corpus about an Information System task in Spanish and Valencian. The corpus was acquired from the telephone line. The corpus is made up of approximately 4 hours of recording (2 hours for each language). It contains a set of 120 utterances (60 for each language) for each of the 20 speakers. An example sentence can be found in Figure 1.

Half of the speakers (10) were native Spanish speakers and the other half (10) were native Valencian speakers. All the speakers were university students and recorded sentences in the two languages. Table 1 summarizes the statistics of the corpus. The distribution of men and women was equal.

Spanish

- Por favor, quiero saber el e-mail de Francisco Casacuberta, adiós.
- Hola, cuál es el horario de consultas de Enrique Vidal?, muchas gracias.

Valencian

- Per favor, vull saber l'e-mail de Francisco Casacuberta, adeu.
- Hola, quin és l'horari de consultes d'Enrique Vidal, moltes gràcies.

English

- Please, I want to know Francisco Casacuberta's e-mail. Goodbye.
- Hello, what are Enrique Vidal's office hours? Thank you very much.

Figure 1: A selected sentence of the corpus. The English translation is provided for a better understanding of the example.

		Spanish	Valencian
Training	Sentences	240	240
	Running words	2887	2692
	Length	1 h 33 m	1 h 29 m
	Vocabulary	131	131
Test	Sentences	60	60
	Running words	705	681
	Length	23m	21m

Table 1: Corpus statistics.

The complete description of the corpus can be found in (Alabau and Martínez, 2006).

There were no out-of-vocabulary (OOV) words in the Spanish test corpus, and only 2 OOV words were observed in the Valencian test corpus.

Due to the small size of the speech corpus (both in signal and vocabulary), we can expect low-perplexity language models but badly estimated acoustic models.

4. Language and acoustic modeling

4.1. Language models

Language models define what type of sentences are allowed by a system. Therefore, our language models must accept all the sentences of the training corpus.

In our case, all the sentences of the corpus have a common structure: greeting, question, information, title, person, and farewell. Some examples of sentences are shown in Figure 1. A sentence is not required to have all the fields. In accordance with this idea, we constructed an automaton using blocks as the language model.

We developed two separate language models: an automaton was built for each language using the acceptor automaton (in the corresponding language) of each block. The automaton was made by joining the acceptor automata in a series. For every two consecutive automata, we merged the final states of the first acceptor automaton with the initial states of the second one. Figure 2 shows an example of the serialization process.

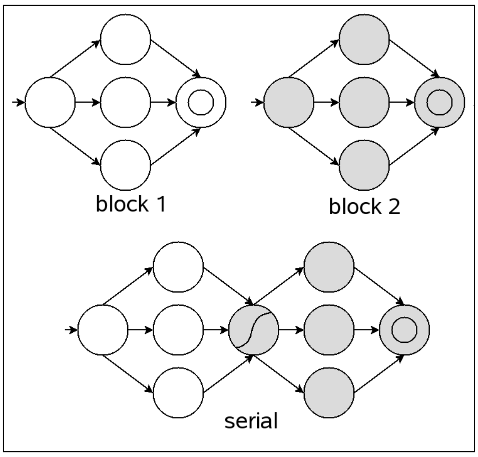

Figure 2: Illustration of the serialization process.

We computed the perplexity of these language models with the following results: Spanish 5.98 and Valencian 6.46.

Note that the perplexity of the models is very low. This is in accordance with the size of the speech corpus, which is also small.

4.2. Acoustic models

Each acoustic model is associated to a phoneme (i.e., monophones) in order to make a comparison of the features of an acoustic sequence with the acoustic models. The acoustic models were hidden Markov models (HMM) that were trained using the HTK toolkit (Young et al., 2004). The HMMs followed a three-state, left-to-right topology without skips. We tested models with 32 Gaussians per state. Each gaussian modeled a 33-component feature vector (10 cepstrals coefficients plus energy with the first and second derivatives).

To obtain reliable acoustic models, they must be trained with huge amounts of training data. It was easy to find large Spanish corpora, but we did not have a large Valencian corpus available for this study. However, since Valencian is phonetically very similar to Spanish, we were able to use the small amount of training material available in Valencian to adapt acoustic models that were trained from a large Spanish corpus. We used the *Senglar* corpus as the initial training corpus. This corpus has been successfully used in other tasks (Casacuberta et al., 2004). The recording conditions for the *Senglar* corpus were different from those of our corpus. For this reason, we obtained adapted acoustic models from the *Senglar* acoustic models for both Spanish and Valencian so that they could be used in our task.

Adapted acoustic models were obtained with the MLLR technique by estimating a global adaptation matrix for each language (i.e., only one regression class). We used our training corpus as adaptation material to obtain the adapted acoustic models. This technique has been used before to obtain acoustic models for multilingual speech recognition. The quantity of signal that we used to adapt the models was similar to the quantity of signal used in previous works (Schultz and Waibel, 2001) and (Zhao and O'Shaughnessy, 2007).

5. Experiments and Results

To analyze the results, we used the Word Error Rate (WER) as the evaluation measure. This measure computes the edit distance between a reference sentence and the recognized sentence.

To perform the experiments, we used two language models: a Spanish language model and a Valencian language model. We used adapted acoustic models for Spanish and Valencian. These acoustic models were adapted with the adaptation data in Spanish and Valencian from the acoustic models trained with the Senglar corpus (Casacuberta et al., 2004).

The adaptation data was small, so we adapted the acosutic models with only one regression class and, therefore, we only computed one transformation matrix.

We implemented the three variants of MLLR presented above: full matrix, diagonal matrix and mean square.

Full matrix and diagonal matrix were calculated with one iteration of Expectation-Maximization because more iterations provided worse results. In general, when the initial models provide good Gaussian frame alignments, only a single iterarion of EM is required to estimate the transformation matrix (Woodland, 2001).

To obtain baseline results, we performed experiments for Spanish and Valencian with the *Senglar* acoustic models without adaptation. The *Senglar* Valencian models were built by cloning the most similar Spanish model for each Valencian phoneme.

	Spanish	Valencian
Baseline	11.0%	16.5%
Full Matrix	6.4%	11.5%
Diagonal Matrix	7.6%	11.7%
Mean Square	**5.6%**	**10.4%**

Table 2: Results of experiments. (The best results are in boldface)

Table 2 shows the best results of experiments. In the same conditions, another standard MLLR tool (HTK) provided similar results for the full matrix case (6.0% in Spanish, 11.0% in Valencian) (Young et al., 2004). As the results show, in this case, it is best to use Mean Square because the WER improved 5 points in Spanish (from 11.0% to 5.6%) and 6 points in Valencian (from 16.5% to 10.4%). Mean Square obtains the best results because the difference between the Senglar acoustic models and the adaptation data is large enough to make the different covariances a source of errors when computing the state occupancy ($\gamma_{qi}(t)$). The difference between a full matrix and a diagonal matrix is small, but with a full matrix the results are better than with a diagonal matrix.

6. Conclusions and Future Work

We have implemented three variants of MLLR for language adaptation in order to choose the best alternative for our system, which deals with two languages: Spanish and Valencian. Our proposal is to employ language adaptation in these languages. We used acoustic models (trained with a large corpus in Spanish) and our training corpus as adaptation material to obtain the adapted acoustic models. The results show that, in this case it is better to use Mean Square. The results with language adaptation are better than without language adaptation. In conclusion, we think that MLLR is a good option for language adaptation.

Nevertheless, these conclusions should be confirmed with a larger corpus and a more realistic task. In future work, we plan to adapt our acoustic models with more appropriate initial acoustic models, i.e., a set of Spanish acoustic models that is closer to our conditions. It would be interesting to obtain the adapted Valencian acoustic models from a set of standard Catalan acoustic models (Moreno et al., 2006). Moreover, we plan to test different quantities of regression classes and other adaptation techniques such as MAP (Gauvain and Lee, 1992).

7. References

Alabau, V. and C.D. Martínez, 2006. Bilingual speech corpus in two phonetically similar languages. In *Proc. of LREC'06,* pp. 1624–1627.

Casacuberta, F., H. Ney, F. J. Och, E. Vidal, J. M. Vilar, S. Barrachina, I. Garcia-Varea, C. Martinez D. Llorens, S. Molau, F. Nevado, M. Pastor, D. Pico, and A. Sanchis., 2004. Some approaches to statistical and finite-state speech-to-speech translation. *Computer Speech and Language*, 18:25–47.

Leggetter, C.J. and P.C. Woodland, 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech and Language*, 9:171–185.

Moreno, A., A. Febrer, and L. Márquez, 2006. Generation of language resources for the development of speech tecnologies in Catalan. In *Proc. of LREC'06,* pp. 1632–1635.

Schultz, T. and A. Waibel, 2001. Language-independent and language-adaptive acoustic modeling for speech recognition. *Speech Communication*, 35(Issues 1-2):31–51.

Uebler, U., 2001. Multilingual speech recognition in seven languages. *Speech Communication*, 35:53–69.

Vilajoana, Jordi and Damià Pons, 2001. *Catalan, language of Europe*. Generalitat de Catalunya.

Woodland, P.C., 2001. Speaker Adaptation for Continuous Density HMMs: A Review. *ITRW on Adaptation Methods for Speech Recognitionn*, 11–19.

Luján, M. and C. D. Martínez and V. Alabau, 2007. A study on bilingual speech recognition involving a minority language. *33rd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, 138–42.

Young, S., G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, July, 2004. *The HTK Book*. CUED, UK, v3.2 edition.

Xufang Zhao and Douglas O'Shaughnessy, August 27-31, 2007 An Evaluation of Cross-Language Adaptation and Native Speech Training for Rapid HMM Construction Based on Very Limited Training Data. *Interspeech 2007*.

J. Gauvain and C. Lee, February, 1992 MAP Estimation of Continuous Density HMM: Theory and Applications. *In Proc. DARPA Speech and Natural Language Workshop*, 185–190.