Cross-Task Adaptation and Speaker Adaptation in Air Traffic Control Tasks

R. Córdoba, J. Ferreiros, J.M. Montero, F. Fernández, J. Macías-Guarasa, and S. Díaz

Grupo de Tecnología del Habla. Departamento de Ingeniería Electrónica
Universidad Politécnica de Madrid

{cordoba, jfl, juancho, efhes, macias, sdiaz}@die.upm.es

Abstract

When we want to develop a recognition system for a new environment, we have to decide which is the best option in what respect to the acoustic modeling: developing acoustic models from scratch using the data available for the new environment or to do cross-task adaptation starting from reliable HMM models.

In this paper, we show the performance of several alternatives, comparing cross-task MAP and MLLR adaptation, in two speech recognizers related to air traffic control tasks, one for spontaneous speech and the other one for a command interface.

For the spontaneous speech recognizer, we also include the comparison between MAP and MLLR for speaker adaptation using a variable amount of adaptation data.

We show how MAP outperforms MLLR when there is enough data available and the threshold points where both techniques provide similar performance or MLLR is better than MAP. In all cases, we show the effectiveness of variance adaptation.

1. Introduction

To develop a speech recognition system in a new environment we have to take into account that the usual speech recognition systems often perform well when tested on data similar to that used in training, but give much higher error rates when tested on data from a new task.

So, we have to consider two options. In first place, we can begin from scratch, but collecting a large amount of task-specific data needs a great effort, it is very costly, and is often impractical. And, worst of it, may be the result is 'I would need more data to have a robust system'.

The second option is to do cross-task adaptation as we did in a previous work [1]. We need a generic recognition system that works well over a range of tasks, it has to be very robust and trained with a lot of data. Then, with a small set of adaptation data, we adapt it to the new environment.

This work has been done under the project INVOCA, for the public company AENA, which manages Spanish airports and air navigations systems [2]. We have worked with two different systems and two languages, Spanish and English, the first system is a command interface, used to control the air traffic controller position, and the second one is a spontaneous speech system with conversations between controllers and pilots. Therefore, we have worked with four different databases in total.

We have considered the two main adaptation techniques that can be applied to cross-task adaptation: maximum a posteriori (MAP) estimation [3] and maximum likelihood linear regression (MLLR) [4, 5]. We will show the behavior of each technique in all these systems with varying sizes and characteristics. In both techniques, we will see the effect of adapting the means alone or the means and variances together.

We will also see the effect of speaker adaptation in the command interface for Spanish, using the same techniques and varying the size of the adaptation set to find the point where MAP outperforms supervised MLLR.

The paper is organized as follows. In section 2 we present the database used in the experiments and the general conditions of the experiments. In section 3, the results for the command interface and the spontaneous speech systems are described. The conclusions are given in Section 4.

2. System Setup

2.1. Databases used

We have used two different databases:

- An isolated speech database, used in a command interface to control the air traffic controller position. In fact, it contains some compound words.
- A spontaneous speech database, which consists of conversations between controllers and pilots. It is a very difficult task, noisy and very spontaneous.

2.2. General conditions of the experiments

The system uses a front-end with PLP coefficients derived from a Mel-scale filter bank (MF-PLP), with 13 coefficients including c_0 and their first and second-order differentials, giving a total of 39 parameters for each 10 msec. frame.

As the channel conditions are noisy, we decided to apply CMN plus CVN. CMN plus CVN meant a 15% improvement in average over CMN alone in preliminary experiments.

For all experiments, we have considered very detailed sets of allophones. In Spanish, we used a set of

45 units [6]: we differentiate between stressed/unstressed/nasalized vowels, we include different variants for the vibrant 'r' in Spanish, different units for the diphthongs, the fricative version of 'b', 'd', 'g', and the affricates version of 'y' (like 'ayer' and 'cónyuge').

In English, we defined a very detailed set of 61 units: we have 19 vowels and diphthongs plus 16 of them stressed. The remaining units are consonants.

All systems use context-dependent continuous HMM models built using decision-tree state clustering. We have developed our own rules using phonetic relevant information in Spanish and English.

2.3. Isolated word recognition experimental setup

A database specific to the project Invoca was recorded. The vocabulary of the task consists of 228 different commands (words or compound-words) for the Spanish experiments and 242 commands for the English case. We had a total of 30 different speakers, all identified, so we could do speaker adaptation experiments. Table 1 shows the database details for isolated speech recognition.

Table 1. Database for isolated speech (words / hours)

	Spanish	English
HMM training set	20,380 / 10.4	11,589 / 5.6
Test set	10,220 / 5.2	9,097 / 4.4

For cross-task adaptation, we have used the SpeechDat database for Spanish, the isolated speech part, with 4,000 speakers who utter the following items: application words, isolated digits, cities, companies, names, and surnames. There are a total of 44,000 files for training (41.8 hours).

2.4. Spontaneous speech recognition experimental setup

Another database was created for these experiments. It consists of recordings on five air traffic control real positions (Arrivals, Departures, Madrid Barajas North taxing, Madrid Barajas South taxing and Clearances). As Barajas is an International Airport, both Spanish and English utterances have been obtained interleaved. The recordings proceeded for about one week per position on a channel where only the controller speech was captured. During these recordings, a group of about 30 different controllers for each position contributed with their voices to the database. Although they knew, for legal requirements, that they were being recorded, they were doing their real work and the speech produced was fully spontaneous. In fact, recording equipment was in a different room from the actual controlling facility, and thus, no disturbance has been produced on their work.

The only drawback for our purposes is that they did not allow us to control the identity of each speaker, so we could not do speaker adaptation experiments in this task, as we would have liked, in a similar way to the command interface.

Expert labelers that marked each sentence with relevant information regarding both the correct grapheme and the artifacts that actually appeared in the speech realization processed the recordings.

To train the HMMs from scratch, we used speech from the Clearances position. Table 2 shows the database details. The same training set was used to do cross-task adaptation.

Table 2. Database for continuous speech (sentences/h.)

	Spanish	English
HMM training set	4,588 / 8.0	2,700 / 5.7
Test set	503 / 0.9	453 / 0.9

Table 3 shows the size of the vocabularies for each position (vocab. size) along the test set perplexity of the bigram language model used (LM perp.) and the number of sentences used for testing (# test senten.) As the sizes of the tasks show, more emphasis and work has been applied to Spanish sentences and on the Clearances position.

We decided to use a bigram language model for two reasons: the phraseology used by the controllers is very regular, so a bigram could be enough, and the text that we had available was clearly insufficient to train a trigram language model.

Table 3. Vocabulary / grammars in spontaneous speech

Task	Spanish			English		
	vocab. size	LM perp.	# test senten.	vocab. size	LM perp.	# test senten.
Clearances	1104	15.2	503	793	23.2	453
Departures	835	11.3	233	425	12.1	71
Arrivals	501	19.5	211	322	16.7	57
North taxing	1624	23.9	349	573	17.9	70
South taxing	1716	29.5	235	641	42.4	123

For cross-task adaptation, we have used the SpeechDat database, the continuous speech part, with 4,000 speakers who utter 9 phonetically rich sentences. Taking out sentences with mistakes and 500 sentences for test, we used a total of 31,393 sentences for training (43.2 hours).

3. Experiments and Results

3.1. Isolated word recognition - Spanish

3.1.1. New system from scratch

We used the train set mentioned in Table 1 to create HMM models from scratch. First, we estimated context independent (CI) models with 10 mixture components per state: we got 2.6% error rate with the vocabulary of 228 commands. Then, we estimated context dependent (CD) models with 1509 states after the tree-based clustering, each state with 6 mixture components. The error rate with that system was **0.95%**.

3.1.2. Cross-task adaptation

We use robust context-dependent HMM models trained with the SpeechDat database. The optimum error rate obtained in that environment was 3.8% with a 500 words dictionary, using a total of 1509 states in the HMMs and 6 mixture components per state.

Using those models without adaptation, the result is 2.1% error rate, so they are worse than the system from scratch, showing that there is a mismatch between both environments.

Beginning from those models, we have considered two types of adaptation: MAP [3] and supervised MLLR [4, 5], as we know the transcription of the adaptation data. The results for MAP can be seen in Table 4. We can see that the results are similar to the ones obtained beginning from scratch, probably because the database is big enough for this task and we need very little information from the original model.

Table 4. MAP adaptation (Isolated-Spanish)

	% error rate
Means adaptation	1.12
Means and variances adaptation	**0.91**

For MLLR [4, 5], we have considered regression class trees of different sizes (64, 128, and 256 nodes), block-diagonal linear transformations and several iterations were run. We can see the results in Table 5.

Table 5. MLLR adaptation (Isolated-Sp.) (% error rate)

	# nodes	Iteration number				
		1	2	3	4	5
Means adaptation	64	1.80	**1.77**	1.82	1.82	1.80
	128	1.52	1.54	**1.50**	1.55	1.53
	256	1.44	1.32	1.24	1.24	**1.22**
Means and variances adaptation	64	1.69	1.59	**1.53**	**1.53**	1.55
	128	1.33	1.29	1.27	1.27	**1.21**
	256	1.20	1.14	1.09	**1.06**	1.08

We can see that MAP outperforms MLLR in all cases, as could be expected as the adaptation database is quite big. In any case, the difference is small. To confirm that, all systems with 256 nodes give better results, showing that there is enough data to model all 256 transforms (and probably more).

3.1.3. Speaker adaptation

For speaker adaptation we begin from the best models so far, obtained using MAP with means and variances adaptation (0.91% error rate).

In this case, we are going to vary the amount of data dedicated to the adaptation. In this database, every speaker uttered five times the list of 228 commands defined for the application. We are going to dedicate up to three of those repetitions for speaker adaptation and do the test with the other two repetitions (there are a total of 4,086 files for the test set in these experiments). Considering this new test set the error rate is 0.73%. The results for MAP and MLLR speaker adaptation are shown in Table 6. For MLLR, several iterations were run again, but just the best result is shown.

Table 6. MAP & MLLR speaker adaptation

	Adaptation set (words)	MAP	MLLR
Means adaptation	50	0.56	0.61
	228	0.29	0.47
	456	**0.17**	0.39
	684	**0.17**	0.29
Means and variances adaptation	50	0.56	0.54
	228	0.27	0.27
	456	**0.17**	0.27
	684	**0.17**	**0.15**

We can extract some interesting conclusions from this results:

- Variance adaptation has very little effect on MAP, but for MLLR the improvement is obvious (30% error rate reduction in average).
- Using variance adaptation, both techniques provide very similar results. We are probably very close to the maximum performance of the system.
- With only 50 words of speaker adaptation MLLR outperforms MAP slightly (as could be expected), and the relative improvement is a remarkable 26% for MLLR.
- With 456 words, MAP outperforms MLLR, but surprisingly with 684 words MLLR is slightly better than MAP. In any case, both techniques are close to a limit in performance. 0.15% equals 6

mistakes (from 4,086 files) that could be considered as impossible to recover.

3.2. Isolated word recognition - English

3.2.1. New system from scratch

Again, we used the train set mentioned in Table 1 to create HMM models from scratch. First, we estimated context independent (CI) models with 10 mixture components per state: 8.2% error rate with the vocabulary of 270 commands. Then, we estimated context dependent (CD) models with 1400 states after the tree-based clustering, each state with 8 mixture components. The error rate was **2.7%**.

The error rate was clearly worse than in the Spanish system with a similar dictionary. The reason for that behavior is that the speakers were in fact Spanish (non-native) and we observed that many pronunciations were quite different from the phoneme transcriptions we had used (native English). We included some alternative pronunciations in the dictionary trying to cover the different possibilities, but we could not get a performance similar to the Spanish system.

In this system, we did not do cross-task adaptation because we did not have a previous robust and general system trained for English. We did not do either speaker adaptation because error rates were low enough to fulfill the project specifications.

3.3. Spontaneous speech recognition - Spanish

3.3.1. New system from scratch

We used the train set with 8 hours (see Table 2) to create HMM models from scratch. All adaptation results refer to the Clearances task. So, they correspond to 503 test sentences with a vocabulary of 1,104 words. In first place, we created context independent (CI) models with 10 mixture components per state: 16.7% error rate. Then we created context dependent (CD) models with 1506 clustered states, each state with 8 mixture components. The error rate with that system was **12.7%**. We created another two systems using 1203 and 1803 states, but results were slightly lower for them.

3.3.2. Cross-task adaptation

Again, we used context-dependent HMM models trained with the SpeechDat database (43.2 hours). The optimum error rate obtained in that environment was 4.2% with a 3,065 words dictionary, using a total of 1,807 states in the HMMs and 7 mixture components per state.

Using those models without adaptation, the result is **19.5%** error rate, so they are even worse than CI models beginning from scratch. There is a clear mismatch between both tasks; the most remarkable aspect is the spontaneity of the Invoca database, whereas SpeechDat is telephone read speech.

After the experience with the isolated database, we decided to do means and variances adaptation, as means only adaptation was worse in all cases.

We can see the results (% error rate) in Table 7. Again, we ran several iterations for MLLR but we only present the best results, always in the third or fourth iteration.

Table 7. MAP & MLLR cross-task adaptation (Spontaneous-Spanish)

	# nodes in MLLR	% error rate
MAP	-	**12.4**
MLLR	64	14.8
	128	14.2
	256	**13.1**

We can extract the following conclusions from these results:

- MAP outperforms MLLR, as could be expected due to the big size of the adaptation set.
- There is enough data to train 256 transforms in MLLR, and probably we could have used 512. In any case, it seems that the performance would be below MAP but very close to it.
- Cross-task MAP adaptation is slightly better that beginning from scratch in this case.

The reason for the improvement in cross-task adaptation is that the adaptation set is much smaller than the train set in SpeechDat, so we can take advantage of some information from the original system. The improvement can be considered low because there is a clear mismatch between tasks: Invoca is very spontaneous and SpeechDat is read speech.

3.4. Spontaneous speech recognition - English

3.4.1. New system from scratch

We used the train set with 5.7 hours (see Table 2) to create HMM models from scratch. All adaptation results refer to the Clearances task (453 test sentences) with a vocabulary of 793 words. As before, we created context independent (CI) models with 9 mixture components per state: we obtained 28.7% error rate.

Then we created context dependent (CD) models with 901 clustered states, each state with 8 mixture components. The error rate with that system was **22.2%**. We created another three systems using 599, 1205 and 1499 states, but the optimum was using only 901.

We can see that the results are clearly worse than in Spanish. We have found two reasons for that: first, the train set is almost half the size and is clearly too small,

as the optimum was found for only 901 states; second, the controllers are non-native speakers and their pronunciation is quite Spanish, especially in airline, airport and city names, and even some greetings and goodbyes are in Spanish. In fact, first results were even worse, so we included alternative pronunciations with a remarkable improvement.

3.5. Spontaneous speech recognition – the other positions

In Table 8, the results for the other controller positions are shown. The language models are specific to the task, but the acoustic models correspond to the Clearances task. So, there is a certain mismatch between some of them and, in some cases, results are not statistically significant, as the number of test sentences is too small. We present these results for information purposes, as the data set available to train both acoustic models and language models was clearly too small.

Table 8. Results for the other controller positions

Task	Spanish			English		
	LM perp.	# test senten.	Error rate	vocab. size	LM perp.	Error rate
Departures	11.3	233	14.1	12.1	71	19.0
Arrivals	19.5	211	23.4	16.7	57	20.5
North taxing	23.9	349	29.4	17.9	70	22.5
South taxing	29.5	235	26.2	42.4	123	33.7

4. Conclusions

We have shown a whole set of adaptation experiments using MAP and MLLR in two different tasks.

For the isolated speech task, the cross-task experiments show that MAP outperforms MLLR in all cases, but the difference is small, and the results with MAP are similar to beginning from scratch. Therefore, the adaptation set is big enough and little information from the original model is used. In the speaker adaptation experiments, we showed that: 50 words are enough for a remarkable improvement; with 50 words, MLLR outperforms MAP slightly; using more words, both techniques have similar results; the best result means a 79.5% relative improvement over no speaker adaptation with a negligible error rate.

For the spontaneous speech system, the cross-task experiments show again that MAP outperforms MLLR, and now cross-task adaptation is slightly better than beginning from scratch (2.4% relative improvement).

In summary, we have created very robust recognition systems with very good error rates in all cases.

5. Acknowledgements

The authors wish to thank the people at Human-Computer Technology Lab (Universidad Autónoma de Madrid) who recorded the Isolated Speech Database; and AENA staff who participated in the recordings of the Spontaneous Database.

6. References

[1] Cordoba, R, Woodland, P.C., Gales, M.J.F., "Improved Cross-Task Recognition Using MMIE Training", *IEEE ICASSP 2002*, pp. 85-88.

[2] *INVOCA Project Synopses*. Eurocontrol. Analysis of Research & Development in European Programmes. Available at http://www.eurocontrol.int/eatmp/ardep-arda/servlets/SVLT014?Proj=AEN043

[3] Gauvain, J.L., Lee, C.H., "Maximum A-Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains", *IEEE Trans. SAP*, Vol. 2, pp. 291-298, 1994.

[4] Gales, M.J.F., Woodland, P.C., "Mean and Variance Adaptation Within the MLLR Framework", *Computer Speech & Language*, Vol. 10, pp. 249-264, 1996.

[5] Leggetter, C.J., Woodland, P.C., "Flexible Speaker Adaptation Using Maximum Likelihood Linear Regression", *Proc. ARPA SLT Workshop*, pp. 104-109. Morgan Kaufmann. 1995.

[6] Córdoba, R., Macías-Guarasa, J., Ferreiros, J., Montero, J.M., Pardo, J.M., "State Clustering Improvements for Continuous HMMs in a Spanish Large Vocabulary Recognition System", *ICSLP 2002*, pp. 677-680.