Improving Language Models in Speech-Based Human-Machine Interaction

Regular Paper

Raquel Justo1,*, Oscar Saz2,3, Antonio Miguel2, M. Inés Torres1 and Eduardo Lleida2

1 Department of Electricity and Electronics, University of the Basque Country, Leioa, Spain
2 Communications Technology Group, Aragon Institute for Engineering Research, University of Zaragoza, Zaragoza, Spain
3 Language Technologies Institute of Carnegie Mellon University, Pittsburgh, PA, USA
* Corresponding author E-mail: raquel.justo@ehu.es

Received 20 Jun 2012; Accepted 5 Dec 2012

© 2013 Justo et al.; licensee InTech. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. Introduction

Nowadays, a great number of processes in our daily life are carried out automatically by machines. In order to get the machines to complete the desired task, a message providing a given set of acoustic instructions has to be conveyed with precision. That is, an effective interaction between human and machine is needed. Human-Machine Interaction (HMI), sometimes referred to as Human-Computer Interaction (HCI), is a discipline concerned with the design, evaluation and implementation of interactive computing systems for human use and the study of the major phenomena surrounding them. Interaction between users and computers can be carried out using different resources: keyboard and mouse, a graphical user interface (GUI), a haptic environment, human conversation, etc. Advances in Artificial Intelligence have promoted research in the HMI field in order to obtain an intuitive and easy means of communication with machines. Additionally, aspects relating to a socially correct interaction dependent upon cultural criteria are also considered. Within this
framework, the use of speech and gestures in the communication process becomes essential, since these are natural modes of communication for humans.

The research carried out in this work deals with an improvement in HMI by using a conversational interface. To do this, a Spoken Dialogue System (SDS) is considered. SDSs, thoroughly described in [1-3], aim to enable people to interact with robots - or computers in general - using spoken language in a natural way. They consist of a human-machine interface that can recognize and understand a speech request uttered by the speaker and provide an output answer. In this process, different modules are involved in order to recognize the speech input, understand their meaning, manage the dialogue and provide the speech output. Two different kinds of SDS were considered in this work:

- Informants: Systems providing some piece of information to the user.
- Actuators: Systems that have to carry out some action.

One of the most important challenges which must be faced in a SDS is the speech understanding issue, in which an Automatic Speech Recognition (ASR) and a Language Understanding (LU) module are involved in translating the speech signal into a set of commands that can be understood by the system. Indeed, the rest of the modules cannot provide a good response when the utterance of the speaker has not been properly understood. Thus, this work focuses on the improvement of this aspect in order to achieve better performance on the part of the dialogue system.

On the other hand, ASR systems have managed to achieve very good performance when in a controlled environment of use and with a collaborative user. However, their rates of accuracy drop when they leave these conditions or else there is a serious acoustic mismatch between the training and the testing data. This is the case with a robot that has been provided with a dialogue system in order to interact with people. It might be working in a noisy environment where different users try to get information about some issue. In order to tackle these problems, Acoustic Modelling has to be considered. Acoustic Models (AMs) can reduce the variability in the speech signal and obtain a set of models and techniques that are robust to all the sources of variability in the speech. In this work, the impact of using a speaker adaptation strategy in the system’s performance is explored. Speaker adaptation is especially desirable in cases where there are many and repeated spoken interactions with a limited number of users; for instance, a system or robot that helps in home-related tasks.

In this work, we propose to improve the performance of an ASR system included in a dialogue system, taking into account the LM on the one hand and the AM on the other hand. Additionally, we contribute by exploring the
reciprocal impact of these two aspects upon the overall behaviour of the system. Moreover, the use of the proposed LMs allows the integration of the ASR and LU system in just one step. This fact promotes the speech-recognition and natural language components to become tightly coupled so that only the acoustically promising hypotheses that are semantically meaningful are advanced.

This paper is organized as follows: Section 2 describes the proposed LMs and provides some experiments showing the improvements associated with the use of these models in the ASR system. In Section 3, a SDS that could be installed in a robot is considered. This new application is described in detail. Within this task, improvements carried out on AMs and LMs are explored along with their reciprocal impact. The results extracted from a different series of experiments are also given. Section 4 describes the use of the proposed LMs in a language understanding task and Section 5 provides the extracted conclusion and the future work.

2. Impact of the Language Model in ASR systems.

In this section, the impact of an improved language modelling approach will be evaluated.

2.1 Hierarchical Language Models

The use of class-based LMs in ASR systems - specifically class n-gram LMs - has been widely explored by different authors [6, 8]. A class n-gram LM is more compact and generalizes better on unseen events; nevertheless, it only captures the relations between the classes of words, while it assumes that the inter-word transition probabilities depend only upon the word classes. This fact degrades the performance of the ASR system. Various approaches based on different techniques can be found in the literature which attempt to avoid this loss of accuracy by integrating different information sources, such as model interpolation [9, 10], class back-off [11], factorized LMs [12, 13] or LMs based on Latent Dirichlet Allocation [14]. Alternatively, phrases or sequences of words can be integrated into the classes of a class n-gram LM instead of isolated words. In this way, a hierarchical language model is proposed, whereby the cooperation between different models is carried out at two levels. At the higher level, a class n-gram LM is generated to learn the structure of the sentence and to generalize on unseen events. At the second level, a word n-gram LM is generated inside each class to capture the relations between the words. Two different approaches to such a hierarchical LM were employed. The main difference between the two is that in the first one (Msw), the words in a phrase are separately studied and the transition probabilities among them are calculated. In the second approach (Msl), and instead, the words in a phrase are gathered and the whole phrase is treated as a unique token, so that new words need to be considered in the vocabulary. Both approaches were formally defined and formulated in [15], and a brief definition of them is given in Sections 2.1.1 and 2.1.2. These models were compared with a classical word n-gram LM (Mw) employed as a baseline. The probability of a word sequence, \( w = w_1 \ldots w_n \), using this model is given in Equation (2), where \( (n_{w} - 1) \) stands for the maximum length of the word histories considered.

\[
P(w) = P_{M_w}(w) = \prod_{i=1}^{n_w} P(w_i | w_{i-1})
\]  

On the other hand, two different approaches to hierarchical LMs based on classes made up of phrases were proposed.

2.1.1 First approach, Msw

In this approach, classes are made up of phrases constituted by unlinked words. The probability of a sequence of words, using this Msw model, can be computed by means of Equation (3):

\[
P(w) = \sum_{c \in C} \sum_{\tilde{c} : w \in \tilde{c}} P(w | s, \tilde{c}) P(s) P(\tilde{c})
\]  

being C the set of all the possible sequences of T classes given a predefined set of classes, where classes are made up of phrases constituted by not linked words. \( S_c(w) \) is a set of segmentations of the word sequence which were compatible with the possible sequences of classes (\( \tilde{c} \)) associated with each sequence of words. Next, we assume the following approaches: a class n-gram model to estimate \( P(c) \), \( P(s | c) = \alpha \) and \( P(w | s, c) \), estimated with zero order models. Equation (3) is rewritten as Equation (4):

\[
P(w) = P_{M_w}(w) = \alpha \sum_{\tilde{c} : w \in \tilde{c}} \prod_{i=1}^{n_{w}} P(w_i | w_{i-1}). P(\tilde{c}_{i-1} | \tilde{c}_{i})
\]

where \( n_{w} \) stands for the maximum length of the word histories considered inside each class \( c \) and \( n_{c} \) is the maximum length of the class history considered.

2.1.2 Second approach, Msl

In the second approach, classes are made up of phrases constituted by linked words, \( \bar{I} \). The probability of a word sequence model is given by Equation (5), being C the set of all the possible sequences of T classes, given a predetermined set of classes:
\[ P(w) = \sum_{c \in C} \sum_{l \in L_c(w)} P(w | l, c) P(l | c) P(c) \]  

(5)

\[ L_c(w) \) is the set of all the possible sequences of phrases compatible with the given sequence of words and the possible sequences of classes. Assuming that \( P(c) \) is estimated using a class n-gram model, \( P(l | c) \) is estimated using zero-order models and, finally, \( P(w | l, c) \) is equal to 1 for \( l \in L_c(w) \) and 0 otherwise. Accordingly, Equation (5) is rewritten as Equation (6), considering that \( P(w) = P_{M_d}(w) \):

\[ P_{M_d}(w) = \sum_{c \in C} \sum_{l \in L_c(w)} \prod_{i=1}^{t} P(l_i | c_i) P(c_i | c_{i-1}^{l-1}) \]  

(6)

### 2.2 Acoustic Models

Nowadays the most widely employed AMs are based on Hidden Markov Models (HMMs) [16]. An HMM is a statistical model in which the system being modelled is assumed to be a Markov process with unobserved (hidden) states. The probability \( P(x | w) \) in Equation (1) can be estimated by the concatenation of different HMMs where the sequence of observable events comprises the parameters extracted from the acoustic signal and it is assumed that they have been generated by a sequence of hidden states. Usually, a three state HMM like the one shown in Figure 2 is used in order to represent phonetic units (phonemes in general) in which the acoustic signal is divided. The concatenation of the different HMMs associated with the phonemes of a word composes the HMM associated with a word.

![Figure 2. HMM associated with the word “no”](image)

### 2.3 Task and Corpus

The experiments - to be presented in this section - were carried out over a task-oriented corpus that consists of human-machine dialogues in Spanish, DIHANA (acquired by a consortium of Spanish Universities) [16]. In this corpus, 225 speakers asked for information about long-distance train timetables, fares, destinations and services by telephone (8 KHz). Thus, an informant dialogue system is needed. A total of 900 dialogues were acquired using the Wizard of Oz technique. Each speaker took part in 4 dialogues. In order to obtain more realistic dialogues, the speakers had to reach a certain goal in each dialogue, while they were entirely free to express themselves as desired. The features of the training and test corpora are given in Table 1. The DIHANA task has an intrinsically high level of difficulty due to the spontaneity of the speech and the problems related to the acquisition of a large amount of transcriptions of human-machine dialogues for training purposes. Therefore, it is well-suited for studying the improvements that result from modifications to the LM. This task is employed in order to show that the presented models can improve the results of an ASR system, as can be seen in [15].

|          | DIHANA   |
|----------|----------|
| Training |          |
| Sentences| 8,606    |
| Different Sentences| 5,590 |
| Running Words    | 77,476 |
| Vocabulary       | 865     |
| Audio Length     | 8.5 hours |
| Test             |          |
| Sentences        | 1,349   |
| Running Words    | 12,365  |
| Vocabulary       | 580     |
| Audio Length     | 2.3 hours |
| Out of Vocabulary Words| 72 |

Table 1. Features of the DIHANA corpus.

### 2.4 Experimental Framework

The ASR system employed in this work was developed in the University of the Basque Country. A set of 32 features (10 Mel-frequency cepstral coefficients with delta and acceleration coefficients, energy and delta-energy) to be processed by the ASR system was extracted from the acoustic signals recorded in the DIHANA corpus. The features were obtained using an analysis window size of 25 ms and with a window shift of 10 ms. In this work, the vocabulary was divided into 821 triphone-like units, each of which is modelled by a typical left-to-right non-skipping self-loop three-state HMM, with 32 Gaussian mixtures per state.

The LMs presented in this work \( M_{LM} \) and \( M_{a} \) were fed into the ASR system to be integrated into the dialogue system. To do this, the different sets of classes that made up phrases - represented in Table 2 - were used. These classes were statistically obtained using a free tool - mkcls - which employs a clustering algorithm based on a maximum likelihood criterion, as developed by [17]. The mkcls must be given the number of classes as a precondition and each word is assigned to only one class. The set of phrases was also obtained using a statistical criterion based on the most frequent n-grams. This process is thoroughly described in [18] and it does not allow ambiguities - that is, only one sequence of phrases can be associated with each sequence of words.

The performance of the ASR system was evaluated when using the proposed LMs and it was compared with that obtained with a standard LM, like a classic word n-gram LM \( M_{a} \). The set of classes for \( M_{a} \) and \( M_{d} \) models was
selected in a tuning stage. The comparison was carried out in terms of the Word Error Rate (WER) and the results are shown in Table 3. The obtained results show that both $M_{ow}$ and $M_{i}$ models significantly outperform the results obtained with the classical $M_{o}$. The percentage of improvement related to $M_{ow}$ is 6.5% ($M_{ow}$-WER - $M_{o}$-WER) · 100 / $M_{o}$-WER), while that associated with $M_{i}$ reaches 8%. In this way, it has been demonstrated that, with a single improvement in the LM, the results of the speech understanding process involved in the dialogue system could provide better results. This fact promotes the HMI to become easier because fewer correction turns have to be introduced by the dialogue manager and the risk of misunderstanding decreases.

| Classes of phrases | 50, 100, 200, 300 |
|--------------------|-------------------|
| Phrase Vocabulary  | 1,289             |
| Total no. classes  | 43,553            |
| Total no. phrases  | 43,553            |

Table 2. Different sets of classes of phrases for DIHANA.

|          | M_{ow} | M_{ow} | M_{i} |
|----------|--------|--------|-------|
| WER(%)   | 19.74  | 18.45  | 18.15 |

Table 3. WER results for the state-of-the-art word n-gram LM and the two LMs proposed in this work.

3. Impact of the AM and the LM

In this section, a new human-machine interaction application that can be directly applied to robotics is described [19]. The final goal will be the design of a butler robot. This robot should be able to control electrical appliances and provide information about their conditions by using an actuator dialogue system. This work focuses on the design of the needed SDS that was developed in the GENIO project [20] and partially supported by FAGOR Home Appliances. It allows the user to ask for the state of each appliance, to program them or else to consult a database of recipes, using spontaneous speech. The butler robot should pick the input voice signal uttered by a speaker and, making use of the dialogue system, provide a multimodal output combining speech, dynamical graphics displays and actions such as switching on/off or programming the different appliances.

In this section we explore the impact of improvements in the acoustic and language modelling, and in the performance of the ASR and the overall dialogue system. That is, two different issues were considered in order to explore their influence in the ASR system: Language and Acoustic Modelling. Specifically, our goal is to explore whether the use of more accurate AMs could influence the improvement rates obtained with hierarchical LMs.

3.1 Acoustic Modelling

Current Acoustic Modelling techniques mostly rely upon the ability of Hidden Markov Models (HMM) theory for describing the speech signal as a random process variable in time. Robust acoustic models based on HMMs can be obtained with large databases that contain speech from different speakers in different environments. In this situation, when a speaker and channel independent model is to be achieved the Maximum Likelihood (ML) algorithm [21] is used.

However, when it is possible to have some speech data from the speaker and in the environment in which the system is going to be used, the best results are achieved by using speaker adaptation over that data. In this case, our butler robot’s workspace is the home environment, where it is going to receive orders from a few different people that can be easily controlled and used to adapt the ASR system. Several adaptation techniques exist these days. Those that are most commonly accepted are the Maximum A Posterior (MAP) algorithm [22] and the Maximum Likelihood Linear Regression (MLLR) algorithm [23]. Both algorithms display good performance in the adaptation to a given speaker and channel conditions when a small portion of data is available. The MLLR adaptation makes a linear regression over the supervector of a given model to map the supervector towards the desired speaker space. The linear regression is computed with the use of the Expectation Maximization (EM) algorithm to calculate matrices $G$ and $Z$ and then matrix $W$, which maps the mean supervector ($\theta_{as}$) to the new space ($\theta_{as,la}$), as seen in Eq. (6):

$$\theta_{mllr} = W_l \times \theta_{as} = (C_{l}^{-1} \times Z_{l}) \times \theta_{as} \quad (6)$$

Contrary to the MAP algorithm, a priori information is not used in MLLR and so badly conditioned adaptation data might create some strongly mismatched adapted units. In addition, MLLR does not require an iterative procedure like MAP to converge on the optimum values; however, iterative MLLR could be used as a way to improve the performance of the speaker-dependent model. In this way, a MLLR strategy was taken for the evaluation of speaker adaptation techniques due to its ability to achieve convergence in a single iteration.

3.2 Language Modelling

The proposed $M_{ow}$ and $M_{i}$ LMs were also considered in this case. However, in order to build these models, semantic classes were employed instead of the
statistically motivated ones that were used with DIHANA corpus. Thus, in this application, semantic information is integrated into class-based LMs where classes are made up of phrases. The classes employed to generate the LMs were chosen to be the semantic classes used by the understanding module of the dialogue system. These semantic classes are, in general, dependent upon the task and the use of this set of classes involves a partial classification of the training corpus - i.e., only some word sequences of the vocabulary were classified. In this way, when the class n-gram language model is built for both the $M_a$ and $M_b$ models, a mixed model is actually obtained - i.e., a model that considers n-grams made up of a combination of classes and words. This fact does not entail any difference in the training or decoding processes because words can be considered as classes made up of a single word.

In this work, the sets of phrases and classes were not obtained by a statistical criterion but they were semantically motivated and manually obtained. These classes were made up of the different sequences of words employed to switch on/off, to program or to ask for information about the state of the electrical appliances. There are also classes made up of affirmative or negative phrases that are not dependent upon the task and could be used in other applications. 40 semantic classes were defined.

An example of some of the employed semantic classes is given below. $ApagarHorno$ corresponds to the sentences used to switch off the oven, $Tiempo$ to sentences related to a period of time, $TemperaturaHorno$ to sentences related to the temperature of the oven, $ProgramaLavadora$ to sentences related to the programs of the washing-machine and $Negación$ is made up of negative clauses.

- $ApagarHorno$: para de cocinar (stop cooking), para el horno (stop the oven), . . .
- $Tiempo$: durante dos horas y veinticinco minutos (for two hours and twenty-five minutes), durante cuatro horas y veinticinco minutos (for four hours and twenty-five minutes), durante veinte minutos (for twenty minutes), . . .
- $TemperaturaHorno$: ochenta grados (eighty degrees), cien grados (a hundred degrees), doscientos veinte grados (two hundred and twenty degrees), doscientos grados (two hundred degrees), . . .
- $ProgramaLavadora$: algodón treinta (cotton thirty), delicado frío (delicate cold), lana (wool), centrifugado (spin), prelavado sesenta (preshaw sixty), . . .
- $Negación$: no (no), no está bien (it is not ok), anular (cancel), incorrecto (incorrect), . . .

In this way, for a word sequence such as “a cien grados de temperatura durante dos horas y veinticinco minutos,” the corresponding segmented sentence would be labelled as follows: “a cien-grados de temperatura durante-dos-horas-y-veinticinco-minutos” and the classified sentence as “a TemperaturaHorno de temperatura Tiempo.” Let us note that ambiguities are not allowed in the segmentation or classification processes. That is, each sequence of words is associated with only one sequence of phrases and each sequence of phrases to only one sequence of classes.

### 3.3 Task and Corpus

The experiments were carried out over a specific corpus that was recorded in the kitchen of the FAGOR1 Home Appliance facilities: Domolab. It was composed of 48 speakers with 125 utterances per speaker. 3 tasks were considered: the control of appliances in the kitchen (90 utterances per speaker), continuous digits (15 utterances per speaker) and 20 phonetically balanced utterances per speaker. On the other hand, 8 audio channels were recorded: 3 located in the kitchen (freezer, extractor hood and washing machine), 3 placed on the speaker, a close talk and 2 lapel microphones and finally 2 channels were recorded with a dummy (right and left ears) placed close to the speaker. Additionally, 2 speaker positions were defined: P0, in front of the washing machine, and P1, in front of the extractor hood. A scheme of the kitchen with the location of the microphones and speaker is given in Figure 3. Additionally, the photograph in Figure 4 shows the recording environment.

1 http://www.fagor.com/
Three acoustic environments were considered in the recording:

- **E0**: no appliances on, with 45 dBA of typical Sound Pressure Level (SPL) of noise.
- **E1**: extractor hood on with a 60 dBA of typical SPL of noise.
- **E2**: washing machine on with a 62 dBA of typical SPL of noise.

15 utterances for every speaker were recorded in every position and acoustic environment. The features of the text corpus employed are detailed in Table 4. A total of 38,798 utterances were used for the training of the speaker-independent model (via the Maximum Likelihood Table).

| Training          | Domolab |
|-------------------|---------|
| Sentences         | 44,236  |
| Different Sentence| 43,982  |
| Running Words     | 349,890 |
| Vocabulary        | 357     |

| Test              |         |
|-------------------|---------|
| Sentences         | 1,617   |
| Running Words     | 9,660   |
| Vocabulary        | 325     |
| Out of Vocabulary Words | 27 |

Table 4. Features of the Domolab corpus.

Likelihood algorithm); 13,600 of these utterances were taken from the training and testing set of the Albayzin database [24] and 25,378 utterances from the training and testing set of the Spanish SpeechDat-Car database [25]. No signals from the Domolab database were used for the training of the speaker-independent model in order to maintain total independence between the train and the testing sets for this Domolab task.

For the evaluation of speaker adaptation techniques, a MLLR strategy was taken. The experiments carried out in this work followed a structure similar to a leave-one-out experiment in order to ensure the independence of the utterances used for adaptation from those used for the evaluation of the experiments. Considering that the amount of data per speaker is 90 sentences, four different experiments were performed. In the first experiment, the first 45 (1, 2, ..., 45) utterances of every speaker were used for training and the last 45 (46, 47, ..., 90) were used for evaluation. The second experiment had the last 45 utterances (46, 47, ..., 90) for training and the first 45 (1, 2, ..., 45) for evaluation. The third experiment took the utterances with an even utterance number (2, 4, 6, ..., 90) for training and the odd (1, 3, 5, ..., 89) utterances for evaluation. While the fourth and final experiment used the odd (1, 3, 5, ..., 89) utterances for adaptation and the even (2, 4, 6, ..., 90) utterances for the testing and evaluation results. The final results were obtained as a statistical average of the four experiments run for every speaker.

3.4 Experimental Framework

In this case, the database was acquired at 16 KHz; thus, it was parameterized into 38 features, 12 Mel-frequency cepstral coefficients with delta and acceleration coefficients, energy and delta-energy. The vocabulary was divided into 24 phoneme-like units corresponding with the Spanish phonemes. These units were also modelled by a typical left-to-right non-skipping self-loop three-state HMM, with 32 Gaussian mixtures per state.

Different series of experiments were carried out in order to evaluate the ASR system’s performance with speaker-dependent and independent AMs and hierarchical LMs (\(M_{	ext{asw}}\) and \(M_{\text{a}}\)) when using the Domolab corpus. The signals used for the experiments were those recorded through the close-talk microphone (referred to as microphone ml0) and through the left lapel microphone (referred to as microphone ml1). That is, acoustic conditions were selected so as to allow consideration in what is not an especially noisy environment. Nevertheless, the use of the lapel microphone involves the inclusion of more noise than the close-talk one. Furthermore, it allows the user to be in movement while talking, promoting in this way a more natural speech. We can think of a person that is walking around the house while ordering the robot butler to carry out an action.

For the evaluation of speaker adaptation techniques, a MLLR strategy was taken. The experiments carried out in this work followed a structure similar to a leave-one-out experiment in order to ensure the independence of the utterances used for adaptation from those used for the evaluation of the experiments. Considering that the amount of data per speaker is 90 sentences, four different experiments were performed. In the first experiment, the first 45 (1, 2, ..., 45) utterances of every speaker were used for training and the last 45 (46, 47, ..., 90) were used for evaluation. The second experiment had the last 45 utterances (46, 47, ..., 90) for training and the first 45 (1, 2, ..., 45) for evaluation. The third experiment took the utterances with an even utterance number (2, 4, 6, ..., 90) for training and the odd (1, 3, 5, ..., 89) utterances for evaluation. While the fourth and final experiment used the odd (1, 3, 5, ..., 89) utterances for adaptation and the even (2, 4, 6, ..., 90) utterances for the testing and evaluation results. The final results were obtained as a statistical average of the four experiments run for every speaker.

3.4 Experimental Framework

In this case, the database was acquired at 16 KHz; thus, it was parameterized into 38 features, 12 Mel-frequency cepstral coefficients with delta and acceleration coefficients, energy and delta-energy. The vocabulary was divided into 24 phoneme-like units corresponding with the Spanish phonemes. These units were also modelled by a typical left-to-right non-skipping self-loop three-state HMM, with 32 Gaussian mixtures per state.

Different series of experiments were carried out in order to evaluate the ASR system’s performance with speaker-dependent and independent AMs and hierarchical LMs (\(M_{\text{asw}}\) and \(M_{\text{a}}\)) when using the Domolab corpus. The signals used for the experiments were those recorded through the close-talk microphone (referred to as microphone ml0) and through the left lapel microphone (referred to as microphone ml1). That is, acoustic conditions were selected so as to allow consideration in what is not an especially noisy environment. Nevertheless, the use of the lapel microphone involves the inclusion of more noise than the close-talk one. Furthermore, it allows the user to be in movement while talking, promoting in this way a more natural speech. We can think of a person that is walking around the house while ordering the robot butler to carry out an action.

For the evaluation of speaker adaptation techniques, a MLLR strategy was taken. The experiments carried out in this work followed a structure similar to a leave-one-out experiment in order to ensure the independence of the utterances used for adaptation from those used for the evaluation of the experiments. Considering that the amount of data per speaker is 90 sentences, four different experiments were performed. In the first experiment, the first 45 (1, 2, ..., 45) utterances of every speaker were used for training and the last 45 (46, 47, ..., 90) were used for evaluation. The second experiment had the last 45 utterances (46, 47, ..., 90) for training and the first 45 (1, 2, ..., 45) for evaluation. The third experiment took the utterances with an even utterance number (2, 4, 6, ..., 90) for training and the odd (1, 3, 5, ..., 89) utterances for evaluation. While the fourth and final experiment used the odd (1, 3, 5, ..., 89) utterances for adaptation and the even (2, 4, 6, ..., 90) utterances for the testing and evaluation results. The final results were obtained as a statistical average of the four experiments run for every speaker.

3.4 Experimental Framework

In this case, the database was acquired at 16 KHz; thus, it was parameterized into 38 features, 12 Mel-frequency cepstral coefficients with delta and acceleration coefficients, energy and delta-energy. The vocabulary was divided into 24 phoneme-like units corresponding with the Spanish phonemes. These units were also modelled by a typical left-to-right non-skipping self-loop three-state HMM, with 32 Gaussian mixtures per state.

Different series of experiments were carried out in order to evaluate the ASR system’s performance with speaker-dependent and independent AMs and hierarchical LMs (\(M_{\text{asw}}\) and \(M_{\text{a}}\)) when using the Domolab corpus. The signals used for the experiments were those recorded through the close-talk microphone (referred to as microphone ml0) and through the left lapel microphone (referred to as microphone ml1). That is, acoustic conditions were selected so as to allow consideration in what is not an especially noisy environment. Nevertheless, the use of the lapel microphone involves the inclusion of more noise than the close-talk one. Furthermore, it allows the user to be in movement while talking, promoting in this way a more natural speech. We can think of a person that is walking around the house while ordering the robot butler to carry out an action.

For the evaluation of speaker adaptation techniques, a MLLR strategy was taken. The experiments carried out in this work followed a structure similar to a leave-one-out experiment in order to ensure the independence of the utterances used for adaptation from those used for the


| m0  | WER | Msw  | Msl | Impr. | WER | Msl  | Impr. |
|-----|-----|------|-----|-------|-----|-----|-------|
| m0  | 5.81| 14.33| 15.3| 4.70  | 19.1|
| m1  | 14.53| 14.31| 1.5 | 13.34 | 8.2 |

Table 5. WER results for microphones m0 and m1, using speaker-independent acoustic models and different LMs: a word n-gram LM and two different approaches to hierarchical LMs (Msw and Msl). The improvement associated with hierarchical models with regard to the Msl model is also given.

Let us note that these improvements are even more noticeable, taking into account the low WER taken as a starting point - i.e., the WER values obtained with the baseline LM. When the results obtained with different microphones are compared with each other, the same tendency can be observed although a significant difference is observed in the improvement percentages (1.5% vs. 15.3% when using the Msw model and 8.2% vs. 19.1% for the Msl model). This could be due to acoustic conditions. That is, when using microphone m1 the LM influence is not so significant due to the high level of the noise of the acoustic signal. Thus, the LM cannot overcome acoustic restrictions and the improvement in the ASR system’s performance is less significant.

Alternatively, speaker-dependent AMs were employed and the experiments were repeated. The results obtained can be seen in Table 6. Looking at these results, it can be noticed that very important improvements have been obtained with these AMs. That is, the MLLR algorithm increases the ASR system’s performance when dealing with the present task. On the other hand, better WER values are observed when comparing hierarchical LMs with the baseline Msw and the improvements are slightly better for Msl, like in the previous case. However, here the improvements are much more significant, reaching 50% with microphone m0 and 30% with m1. Thus, it can be concluded that, when using better AMs, the improvement associated with hierarchical LMs is much more noticeable. Again, this might happen due to acoustic conditions. When using speaker-dependent models, the AMs can better solve the problems associated with acoustic issues - thus, the LM becomes more important and its influence is more noticeable in the ASR system’s results.

| m0  | WER | Msw  | Msl | Impr. | WER | Msl  | Impr. |
|-----|-----|------|-----|-------|-----|-----|-------|
| m0  | 2.31| 2.12 | 47.6| 0.99  | 57.1|
| m1  | 2.47| 1.73 | 29.9| 1.52  | 38.4|

Table 6. WER results for microphones m0 and m1, using speaker-dependent acoustic models and different LMs: a word n-gram LM and two different approaches to hierarchical LMs (Msw and Msl). The improvement associated with hierarchical models with regard to the Msl model is also given.

The best WER result was obtained with speaker-dependent AMs, the Msl model and m0 microphone. This result outperforms the value obtained with the baseline Msw model and the same acoustic conditions by 57.1%. When comparing it to speaker-independent models, the improvement reaches 82.9%.

The obtained results can be interpreted in another way. Specifically, we can wonder if the loss of accuracy related to the use of a lapel microphone - which introduces more noise into the system but gives the user more freedom to walk around the house - could be offset by the integration of improved acoustic and language models. The answer is in the affirmative, because the WER values obtained with microphone m1, speaker-dependent AMs and both Msw and Msl, are better than the result obtained with microphone m0, speaker-independent AMs and Msw LM.

In summary, it can be concluded that the integration of hierarchical LMs, on its own, could benefit the ASR system, as shown in Section 2. On the other hand, a significant difference is observed in the improvement percentage over this task when using hierarchical LMs. This might be motivated by two factors: firstly, the use of semantic ad-hoc classes. This kind of classification has proven to be useful even when using classical class n-gram LMs [26]. Secondly, the power of hierarchical Msw and Msl models can take advantage of a class-based and a word-based LM. Furthermore, when combining these LMs with certain improvements in the AMs, the impact of using more accurate LMs is much more noticeable in the overall system. That is, improvements in each of the models can receive the benefit from the improvements in the other one, leading to an efficient dialogue system that can improve HMI. Notice that improvements carried out in the ASR system can be directly translated into the dialogue system’s performance because, when using an efficient ASR, the number of confirmation turns can be reduced and the number of turns dedicated to error recuperation decreases, leading to much more pleasant interaction for the user. When using a poor ASR instead, it is very difficult for the dialogue manager to go on with a meaningful dialogue and the experience becomes frustrating to the user.

4. Hierarchical LMs in Language Understanding.

When dealing with Language Understanding (LU), different strategies can be found in the literature, aiming to extract the semantic information from an input acoustic sentence [27, 28]. First, LU systems have performed syntactic analysis on the best sequence obtained from an ASR system and used non-probabilistic rules for mapping syntactic structures onto semantic ones [29]. Alternatively, different approaches based on stochastic finite-state models have been developed [30, 31]. There are also some approaches that transform acoustic signals directly into basic semantic constituents [32-34]. In this
kind of system, ASR and LU processes are carried out together by searching at the same time for the best sequence of words and concepts.

The LU module is devoted to extracting semantic information from a text sentence. Let us note that we have considered a LU process consisting of two phases, like the one described in [31]. In the first phase, the input sentence is sequentially translated into an intermediate semantic language; in the second phase, the frame or frames associated with this intermediate sentence are generated. In a frame, a user turn of the dialogue is represented as a concept (or as a list of concepts) and a list of constraints made over this concept.

When hierarchical LMs are considered, the ASR system provides, at the same time, the recognized sentence and the associated classified sentence. Taking into account the fact that in this case semantic classes are involved in the hierarchical models, the semantic information associated with the word sequence can also be retrieved during the recognition process. Therefore, when using these models, the recognition process and the first phase of the understanding module can be merged into a single step, thus speeding up the interventions of the dialogue system. Moreover, when using a LM that includes semantic information, the recognition process will be influenced by this information and only semantically meaningful sentences will be obtained by the ASR system. In this way, the LU process would take the benefit from the proposed LMs, leading to a better dialogue system.

4.1 Experimental Framework

Different experiments were carried out in order to assess the performance of our LU task. The evaluation was carried out in terms of the Category Error Rate (CER). The CER is measured in the same way as the WER is, but considering the class sequences provided by the system and the classified reference sentences. Regarding the \( M_w \) model, the CER values were obtained by classifying reference sentences on the one hand and, on the other hand, classifying the sentences obtained with the ASR system and an \( M_w \) model. The results thus obtained are summarized in Table 7 (speaker-independent) and Table 8 (speaker-dependent).

These CER results show the same tendency observed previously for WER values. However the improvement percentages are much higher here: 9.1% vs. 1.5% for the \( M_w \) model and 17.2% vs. 8.2% for \( M_s \) when microphone \( m1 \) is used and speaker-independent acoustic models are considered. Differences are even higher for microphone \( m0 \). Thus, it can be concluded that the CER improvement is not only due to a better recognized sentence with a lower WER but also due to the semantic information involved in the hierarchical LMs, which is helping in order to provide a higher CER percentage improvement.

When considering speaker-dependent AMs, the same tendency is observed when using microphone \( m1 \), but the CER improvement percentage is lower than the WER improvement percentage for microphone \( m0 \). This happens because of the very low error rates and some ambiguity problems in the reference sentence's classification. When the error rates are higher, these problems are not noticeable. Let us give an example with the user sentence “repite”, which means “repeat”. This word can be found in two different classes: one made up of phrases related to an utterance that has not been understood by the system and another one related to phrases associated with the steps in a recipe. When classifying the reference, the second class was assigned. Let us note that there was no ambiguity in the classification process and a word sequence was labelled with an only one sequence of phrases and classes. The ASR system, instead, assigns the first class to the sentence “repite”.

|     | \( M_w \) | \( M_{sw} \) | \( M_s \) |
|-----|----------|----------|--------|
| %   | CER      | CER      | Impr.  |
| m0  | 6.13     | 4.36     | 28.9   |
| m1  | 15.32    | 13.92    | 9.1    | 12.69  | 17.2   |

Table 7. CER results for microphones m0 and m1, using speaker-independent acoustic models and different LMs: a word n-gram LM and two different approaches to hierarchical LMs (\( M_{sw} \) and \( M_s \)). The improvement associated with hierarchical models with regard to the \( M_w \) model is also given.

|     | \( M_w \) | \( M_{sw} \) | \( M_s \) |
|-----|----------|----------|--------|
| %   | CER      | CER      | Impr.  |
| m0  | 2.58     | 1.74     | 32.5   |
| m1  | 2.79     | 1.41     | 49.4   | 1.31   | 52.7   |

Table 8. CER results for microphones m0 and m1, using speaker-dependent acoustic models and different LMs: word n-gram LM and two different approaches to hierarchical LMs (\( M_{sw} \) and \( M_s \)). The improvement associated with hierarchical models with regard to the \( M_w \) model is also given.

5. Concluding Remarks and Future Work

The goal of this work was to explore speech-based human-machine interaction by considering a dialogue system installed in a robot. Since one of the most difficult issues that must be solved in a dialogue system is the ASR, this work has a bearing on this system. For doing this, a hierarchical LM is employed. It can be concluded that the dialogue system can take advantage of the use of hierarchical LMs in different ways: more accurately recognized sentence will be obtained from the ASR system, on the one hand. This improvement will be even more noticeable when using any method to obtain better AMs, such as speaker adaptation techniques. On the other hand, when using semantic classes in the proposed models, speech recognition and language understanding processes could be merged in just one step, thus speeding
up the dialogue system. Moreover, the use of these semantic classes in the LM only allows the ASR system to provide semantically meaningful sentences, thus improving the LU results.

For future work, we propose to consider the underlying idea of hierarchical language models in the dialogue manager module in order to provide a better dialogue system strategy.

6. Acknowledgments

This work has been partially supported by the Spanish MICINN under grants CSD2007-00018 and TIN2011-28169-C05-04, as well as by the Basque Government under grants GIC10/158 IT375-10.

7. References

[1] Zue V, Seneff S, Glass J, Polifroni J, Pao C, Hazen T, Hetherington L (2000) Jupiter: A telephone-based conversational interface for weather information. IEEE Trans. on Speech and Audio Proc. (2000) 8(1):85–96.
[2] Lamel L, Rosset S, Gauvin J, Bennacef S, Prouts G (1998) The limsi arise system. In: IEEE 4th Workshop on Interactive Voice Technology for Telecommunications Applications. 209–214.
[3] Seneff S, Polifroni J (2000) Dialogue management in the mercury flight reservation system. In: ANLP-NAACL 2000 Satellite Workshop. 1–6.
[4] Jelinek F. Statistical Methods for Speech Recognition. The MIT Press, January 1998.
[5] Gupta V, Lenning M, Mermelstein P (1992) A language model for very large-vocabulary speech recognition. Computer Speech and Language, 6(2):331–344.
[6] Brown P.F, Pietra V.J.D, Souza P.V.D, Lai J.C, Mercer R.L (1992) Class-based n-gram Models of Natural Language. Computational Linguistics 18(4):467–480.
[7] Ries K, Buo F.B, Waibel A. (1996) Class phrase models for language modelling. In: Proc. ICSLP ’96, volume I, pages 398–401, Philadelphia, PA.
[8] Niesler T.R, Woodland P.C. (1996) A variable-length category-based n-gram language model. In IEEE ICASSP’96, volume I, pages 164–167, Atlanta, GA.
[9] Jelinek F, Mercer R. L (1980) Interpolated estimation of Markov source parameters from sparse data. In: Workshop on Pattern Recognition in practise, The Netherlands: North-Holland. pp. 381–397.
[10] Klakow D (1998) Log-linear interpolation of language models, in: Proceedings of ICSLP ’98, pp. 1695–1699.
[11] Zitouni I (2007) Backoff hierarchical class n-gram language models: effectiveness to model unseen events in speech recognition. Computer Speech and Language 21, 99–104.
[12] Bilmes J,A, Kirchhoff K (2003) Factored language models and generalized parallel backoff, in: Hearst, M., Ostendorf, M. (Eds.), HLT-NAACL 2003: Short Papers, Association for Computational Linguistics, Edmonton, Alberta, Canada. pp. 4–6.
[13] Kirchhoff K, Vergyri D, Bilmes J, Duh K, Stolcke A (2006) Morphology-based language modeling for conversational Arabic speech recognition. Computer Speech & Language 20, 589–608.
[14] Chien J.T, Chueh C.H (2011) Dirichlet class language models for speech recognition. IEEE Transactions on Audio, Speech, and Language Processing, 19, 482 – 495.
[15] Justo R, Torres M.I (2009) Phrase classes in two-level language models for ASR. Pattern Analysis & Applications 12, 427–437.
[16] Benedit J.M, Lleida E, Varona A, Castro-Bleda MJ, Galiano I, Justo R, López de Letona I and Miguel A Design and acquisition of a telephone spontaneous speech dialogue corpus in Spanish: DIHANA. In: Proc. of LREC 2006. pp. 1636-1639. Genoa (Italy). May 2006.
[17] Och F.J (1999) An efficient method for determining bilingual word classes. In: Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics, Morristown, NJ, USA, Association for Computational Linguistics. 71–76.
[18] Guijarrubia V.G, Torres M.I (2010) Text- and speech-based phonotactic models for spoken language identification of Basque and Spanish. Pattern Recogn. Lett., 31, 523–532.
[19] Justo R, Szaz O, Guijarrubia V.G, Miguel A, Torres M.I, Lleida E (2008) Improving Dialogue Systems in a Home Automation Environment. In: International Conference on Ambient Media and Systems. Quebec city Canada, 11-14 February 2008.
[20] GENIO. “Gestor Embebidio Natural de Interfaz Oral,” INTEK project. Department of Industry of the Basque Government 2006.
[21] Dempster AP, Laird N.M, Rubin D.B (1977) Maximum likelihood for incomplete via the EM algorithm. Journal of the Royal Statistical Society, 39(1):1–21.
[22] Gauvin J.L, Lee C.H (1994) Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Trans. on Speech and Audio Proc., 2(2):291–298.
[23] Legetter C.J, Woodland P.C (1995) Maximum likelihood linear regression for speaker adaptation of the parameters of continuous density hidden Markov models. Computer Speech and Language, 9:171–185.
[24] Moreno A, Poch D, Bonafonte A, Lleida E, Listerri J, Marino J.B, Nadeu C (1993) Albayzin speech database: Design of the phonetic corpus. In: Proceedings of the Third Eurospeech, Berlin, Germany.
[25] Moreno A, Nogueira A, Sesma A. Speechdat-car: Spanish. Technical Report SpeechDat.

[26] Justo R, Torres M.I, Benedi J.M (2006) Category-based Language Models in a Spanish Spoken Dialogue System. In: C. Inchaurrealde and I. Ibarretxe editors. Procesamiento del Lenguaje Natural. no 37, pp: 19-24. ISSN: 1135-5948.

[27] Mori D, Bechet R, Hakkani-Tur F, McTear D, Riccardi M, Tur G (2008) Spoken language understanding. IEEE Signal Processing Magazine 25, 50–58.

[28] Hahn S, Dinarelli M, Raymond C, Lefevre F, Lehnen P, de Mori R, Moschitti A, Ney H, Riccardi G (2011) Comparing stochastic approaches to spoken language understanding in multiple languages. IEEE Transactions on Audio, Speech & Language Processing 19, 1569–1583.

[29] Woods W.A (1975) What’s in a link: Foundations for semantic networks, in: Bobrow, D.G., Collins, A. (Eds.), Representation and Understanding. Academic Press, pp. 35–82.

[30] Johnston M, Bangalore S (2000) Finite-state multimodal parsing and understanding, In: Proceedings of the 18th conference on Computational linguistics - volume 1, Association for Computational Linguistics, Stroudsburg PA, USA. pp. 369–375.

[31] Segarra E, Sanchis E, Galiano M, Hurtado F.G.L (2002) Extracting semantic information through automatic learning techniques. International Journal of Pattern Recognition and Artificial Intelligence IJPRAI 16, 301–307.

[32] Lin B.S, Chen B, Wang H.M, Lee L.S (2002) A hierarchical tag-graph search scheme with layered grammar rules for spontaneous speech understanding. Pattern Recogn. Lett. 23, 819–831.

[33] Bangalore S, Johnston M (2004) Balancing data-driven and rule-based approaches in the context of a multimodal conversational system, in: Susan Dumais D.M, Roukos S. (Eds.), HLT-NAACL 2004: Main Proceedings, Association for Computational Linguistics, Boston, Massachusetts, USA. pp. 33–40.

[34] Raymond C, Bechet F, Camelin N, Mori R.D, Damnati G (2007) Sequential decision strategies for machine interpretation of speech. IEEE Transactions on Audio, Speech, and Language Processing 15(1):162-171.