Improving Dialogue Systems in a Home Automation Environment

Raquel Justo
University of the Basque Country
48940-Leioa, Spain
raquel.justo@ehu.es

Oscar Saz
University of Zaragoza
Zaragoza, Spain
oskarsaz@unizar.es

Víctor Guijarrubia
University of the Basque Country
48940-Leioa, Spain
vgga@we.lc.ehu.es

Antonio Miguel
University of Zaragoza
Zaragoza, Spain
amiguel@unizar.es

M. Inés Torres
University of the Basque Country
48940-Leioa, Spain
manes.torres@ehu.es

Eduardo Lleida
University of Zaragoza
Zaragoza, Spain
lleida@unizar.es

ABSTRACT

In this paper, a task of human-machine interaction based on speech is presented. The specific task consists on the use and control of a set of home appliances through a turn-based dialogue system. This work focuses on the first part of the dialogue system, the Automatic Speech Recognition (ASR) system. Two lines of work are taken into account to improve the performance of the ASR system. On one hand, the acoustic modeling required for the ASR is improved via Speaker Adaptation techniques. On the other hand, the Language Modeling in the system is improved by the use of class-based Language Models. The results show the good performance of both techniques to improve the ASR results, as the Word Error Rate (WER) drops from 5.81% using a close-talk microphone to a 0.99% and from 14.53% using a lapel microphone to a 1.52%. Also, an important reduction is achieved in terms of the Category Error Rate (CER), which measures the ability of the ASR system to extract the semantic information of the uttered sentence, dropping from 6.13% and 15.32% to 1.29% and 1.32% for the two microphones used in the experiments.

Categories and Subject Descriptors

J.2 [**Physical Sciences and Engineering**]: Electronics; I.2.1 [**Artificial Intelligence**]: Applications and Expert Systems—*Natural language interfaces*

General Terms

Design, Performance

Ambi-sys 2008, February 11-14, 2008, Quebec, Canada.

1. INTRODUCTION

Ambient intelligence involves the convergence of several computing areas, such as ubiquitous computing, intelligent system research or context awareness. The research carried out in this work deals with intelligent systems and specifically with systems capable of interact with humans in a natural way in order to provide different services.

Human-computer interaction can be carried out using different resources: keyboard and mouse, a graphical user interface (GUI), a haptic environment, human conversation,... This work deals with systems that try to improve human-computer interaction by using a conversational interface, that is, a spoken dialogue system.

Spoken Dialogue Systems, thoroughly described in [18, 16], should enable people to interact with computers using spoken language, that is in a natural way, even when the user is in movement, by using simply a lapel microphone. They consist of the following modules: An Automatic Speech Recognition (ASR) module, an understanding module, a dialogue management module, an answer generator and a Text-To-Speech module.

The aim of this work is to develop a home automation application with a conversational interface. It consists of a virtual butler service that would be installed at home to control electrical appliances and provide information about their conditions [17]. The system was developed in the GENIO project [5], partially supported by FAGOR Home Appliances, and it allows the user to ask for the state of each appliance, to program them or to consult a database of recipes, using spontaneous speech. The virtual butler picks the input voice signal uttered by a speaker and making use of a dialogue system provides a multimodal output that combines speech, dynamical graphics displays and actions such as switching on/off or programming the different appliances.

One of the most important difficulties which must be faced in a dialogue system application is the ASR issue. In fact, the ASR module captures the acoustic signal uttered by the speaker and provides the transcription to the other modules

of the dialogue system that will interpret it and will generate the appropriate response. Thus, a poor performing of an ASR module disables the overall evaluation of the dialogue system since the rest of the modules cannot provide a good response from a sentence recognized with a high error rate. Thus, this work focuses on the improvement of the ASR module in order to achieve a better performance of the dialogue system, under different acoustic conditions.

Automatic Speech Recognition systems have managed to achieve very good performance in presence of a controlled environment of use and a collaborative user. But their rates of accuracy drop heavily when they get out of these conditions or there is a serious acoustic mismatch between the training data and the testing data. This mismatch can be due to the acoustic environment (noise or reverberation present in the recording), as well as to inter-speaker or intra-speaker variability. It is the work of Acoustic Modeling to reduce this 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. Strategies like speaker adaptation might be very useful in a task like the presented in this work, where the environment and the speakers are set once the system is installed at home.

The aim of a Language Model (LM) is to capture the way in which the combination of words is carried out in a specific language, therefore it is of a great importance in the framework of Automatic Speech Recognition. Nowadays Statistical Language Models [7] are broadly used in ASR systems, being word n-gram LMs the most widely used approach because of their effectiveness when it comes to minimizing the *word error rate* (WER)[6]. However, when dealing with applications for which the amount of training material available is limited (e.g. specific tasks in dialog systems) the sparsity of the data becomes a problem and an alternative approach such as a class n-gram LM [1] could be used. These kind of models can be extended with phrase-based LMs [15, 8]. In this work we propose the use of a specific extended LM which uses semantic classification in order to improve the speech recognition and understanding rates.

This paper is organized as follows: On Section 2, the task and the acquired corpus recorded for the task are presented. In Section 3, the strategies used for Acoustic Modeling are introduced; while in Section 4 the employed class-based Language Models are described. Posteriorly, the experimental framework and the results are given in Section 5; and the conclusions to this work presented in Section 6.

2. TASK AND CORPUS

In order to complete the preliminary speech recognition results in the home automation scenario, a specific corpus was recorded in the kitchen of the FAGOR Home Appliance facilities: DOMOLAB. It was composed by 48 speakers with 125 utterances per speaker. 3 tasks were considered: 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. In all cases, the frequency sample is 16 KHz and the audio signal is coded with 16 bits. Three acoustic environments were considered in the recording:

		DOMOLAB
Train.	Sentences	44,236
	Different sent.	43,962
	Words	349,890
	Vocabulary	357
Test	Sentences	1,617
	Words	9,660
	Vocabulary	325
	Out Of Vocabulary Words	27
	Perplexity (word-based LM)	9.46

Table 1: Features of the corpus

- 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

Also, 2 speaker positions were defined: P0, in front of the washing machine, and P1, in front of the extractor hood. 15 utterances for every speaker were recorded in every position and acoustic environment.

The features of the employed text corpus are detailed in Table 1. The transcription of the sentences uttered by the first 30 speakers were chosen as the training set and the LMs were generated over this training set. The sentences corresponding to the last 18 speakers were employed as the test set.

3. ACOUSTIC MODELING

Current Acoustic Modeling techniques mostly rely on 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 [2] is used.

But, 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. This is the situation in this work, where the system is to be installed in a home with a limited number of users. Several adaptation techniques are existing these days. The ones more commonly accepted are the Maximum A Posterior (MAP) algorithm [4] and the Maximum Likelihood Linear Regression (MLLR) algorithm [11]. Both algorithms show 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 that maps the mean supervector (θ_{ML}) to the new space (θ_{MLLR}) as seen on Equation 1.

$$\theta_{MLLR} = W_i * \theta_{ML} = (G_i^{-1} * Z_i) * \theta_{ML} \quad (1)$$

In many works, the set of units in the model are clustered according to their proximity in the units space prior to the estimation procedure (i.e. computed like a Kullback-Leibler distance [10]). In this way, units that may not appear in the training data, but are close to units that are in that data, will use the information of the existing units to modify their supervectors.

Contrary to MAP algorithm, a priori information is not used in MLLR, so bad conditioned adaptation data might create some strongly mismatched adapted units. Also, MLLR does not require an iterative procedure like MAP to converge to the optimum values; but, iterative MLLR could be used as a way to improve the performance of the speaker dependent model.

4. LANGUAGE MODELING

The use of class-based LMs in ASR systems, specifically class n-gram LMs, have been widely explored by different authors [1, 14]. 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 assumes that the inter-word transition probabilities depend only on the word classes. This fact degrades the performance of the ASR system. In order to avoid the loss of information associated to the use of class n-gram LMs, we integrate in this work, phrases or sequences of words instead of isolated words, into the classes of a class n-gram LM. In this way, a class n-gram LM is generated to learn the structure of the sentence and a word n-gram LMs is generated inside each class to capture the relations between the words. Two different approaches to such a class-based LM were employed. The main difference between the two is that in the first one (M_{sw}) the words in a phrase are separately studied and the transition probabilities among them is calculated. In the second approach (M_{sl}) instead, the words in a phrase are gathered and the whole phrase is treated as a unique token, so as new words need to be considered in the vocabulary. Both approaches were recently described and formulated in [9].

The main novelty of this work lies in the integration of semantic information into class-based LMs where classes are made up of phrases. The classes employed to generate the LMs of the ASR system were chosen to be the semantic classes used by the understanding module of the dialogue system. These semantic classes are in general dependent on the task and are made up of sequences of words (phrases).

The use of this set of classes involves a partial classification of the training corpus, i.e. only some words of the vocabulary were classified. In this way, the class n-gram LM is actually a mixed LM that can contain n-grams over both words and word classes. On the other hand, a word n-gram LM is generated within each class. When the first approach is used (M_{sw}) the relations among the words of the phrases in a class are considered in the word n-gram LM and different values of n could be explored. When the second approach is used (M_{sl}) a unigram LM must be generated since the phrases are considered as a unique token or new word. Therefore, the proposed class-based LM, take advantage of both a word-based and a class-based LM, by learning the structure of partially classified sentences and by generating a word-based LM within the classes. Thus a improved LM is obtained, which can provide a better performance of the ASR system in terms of *word error rate*.

Moreover, when using semantic classes to generate the proposed class-based LMs in the ASR system, the transcription of the sentence and the semantic information associated to the classes could be obtained at the same time. The extraction of the semantic information from a text sentence is related to the understanding process so that it is the work of the understanding module. The understanding process consists of two phases. In the first one the input sentence is sequentially translated into a sentence of an intermediate semantic language, and, in the second phase, the frame or frames associated to this sentence are generated. In a frame, a user turn of the dialogue is represented as a concept (or a list of concepts) and a list of constraints made over this concept. Thus, in this case, the recognition process and the first phase of the understanding process could be merged in an only one step.This fact could help to speed up the interventions of the dialogue system. Furthermore, the system could even retrieve semantic information lost due to recognition errors. In this work we explore if obtaining the semantic information directly in the recognition process a better *category error rate* (CER) could be obtained. The CER value gives an idea of the performance of the understanding process, since it considers the semantically categorized sentence.

The semantic classes were manually chosen and 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 on the task and could be used in other applications. However all of these classes have relevant information to provide an answer to the user. 40 semantic classes were defined.

An example of some of those classes are 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 programes of the washing-machine and **Negación** is made up of negative clauses.

- **ApagarHorno**: para de cocinar, dejar de cocinar, para el horno,...
- **Tiempo**: durante dos horas y veinticinco minutos, durante cuatro horas y veinticinco minutos, durante veinte minutos, ...

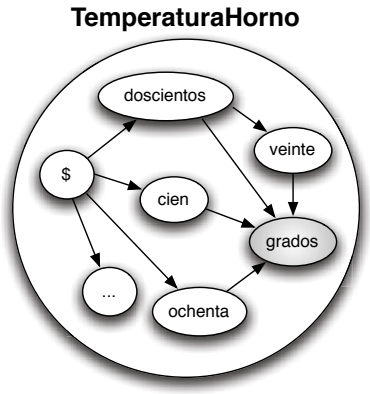

Figure 1: Automaton of the model generated for the TemperaturHorno class according to the M_{sw} model. The "$" state is considered as the initial state and the states in grey are final states

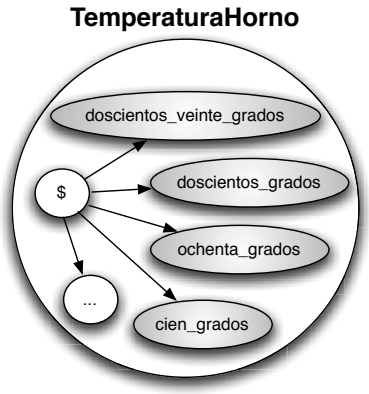

Figure 2: Automaton of the model generated for the TemperaturHorno class according to the M_{sl} model. The "$" state is considered as the initial state and the states in grey are final states

- **TemperaturaHorno**: ochenta grados, cien grados, doscientos veinte grados, doscientos grados, ...
- **ProgramaLavadora**: algodón treinta, delicado frio, lana, centrifugado, prelavado sesenta, ...
- **Negación**: no, no está bien, anular, incorrecto, ...

In this way, for a word sequence such as "a cien grados de temperatura durante dos horas y veinticinco minutos", the corresponding classified sentence is "a **TemperaturaHorno** de temperatura **Tiempo**". The automata associated to the word ngram LM generated for the class **TemperaturaHorno** are shown in Figure 1 and Figure 2 corresponding to the approaches M_{sw} and M_{sl} respectively.

5. EXPERIMENTAL FRAMEWORK

The techniques used in this work for the improvement of the Acoustic and Language Modeling have been tested over 18 of the 48 speakers of the Domolab database. The signals used for the experiments were the recorded through the close-talk microphone (referred to as m0 microphone) and through the left lapel microphone (referred to as m1 microphone).

The set of features used for the ASR system extracted from every one of the signals in the database is a 39-feature set. Features are obtained every 10 milliseconds in the signal, with a window length of 25 millisecond. The features are based on the Mel-Frequency Cepstrum Coefficients (MFCC), 12 features represent the first 12 cepstral coefficients of the signals, 12 features represent the first derivative of the cepstrum and 12 features represent the second derivative of the cepstrum. Also, the log-energy, as well as the first and second derivative of the log-energy are used in the feature set. Cepstral Mean substraction (CMS) [3] was used as a first way to avoid mismatching between the channel features of the signals used for training the speaker independent model and the channel of the Domolab utterances.

5.1 Acoustic Modeling

The speaker independent model in this work was trained via the Maximum Likelihood algorithm. A total of 38,798 utterances were used for the training of this model; 13,600 of these utterances were taken from the training and testing set of the Albayzin database [13] and 25,378 utterances from the training and testing set of the Spanish SpeechDat-Car database [12]. No signals from the Domolab database were used for the training of the speaker independent model in order to keep total independence between the train and the testing sets for this Domolab task. This model contained 744 acoustic units, each one of them modeling a context-dependent part of the speech (i.e. the transition between a given pair of phonemes). Also two more models were created to model the silence at the beginning and at the end of the utterance and to model the inter-word silence. All of this 746 units were modeled as one-state models, each one of them described by a Gaussian Mixture Model (GMM) containing 32 gaussians per state.

For the evaluation of speaker adaptation techniques, a MLLR strategy was taken, due to its ability to achieve convergence in one single iteration. The experiments carried out in this work followed a structure similar to a leave-one-out experiment, in order to keep the independence between the utterances used for adaptation and the ones used for the evaluation of the experiments. Considering that the amount of data per speaker is 90 sentences, four different experiments were made. 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 last 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.

Only the mean vectors in the GMMs where modified during the adaptation process. No clustering was finally made prior

	Speaker Independent Acoustic Models					
	M_w		M_{sw}		M_{sl}	
	WER	CER	WER	CER	WER	CER
m0	5.81	6.13	4.92	4.36	4.70	3.39
m1	14.53	15.32	14.31	13.92	13.34	12.69

Table 2: WER and CER for the m0 and m1 microphones, using speaker independent acoustic models and different LMs: a word n-gram LM and two different approaches to class-based LMs (M_{sw} and M_{sl})

	Speaker Independent Acoustic Models			
	M_{sw}		M_{sl}	
	WER	CER	WER	CER
m0	15.3%	28.9 %	19.1 %	44.7%
m1	1.5%	9.1%	8.2%	17.2%

Table 3: Percentage of improvement in WER and CER for the two class-based LMs compared to the word n-gram LM using speaker independent models

to the adaptation, as there is enough data for all the most used units to appear in the training data.

5.2 Language Modeling

Different series of experiments were carried out over the DOMOLAB corpus in order to evaluate the ASR system performance when different LMs are considered. Firstly, the speaker independent acoustic models were employed considering the following LMs: A classical word n-gram LM (M_w) and two different approaches to class-based LMs (M_{sw} and M_{sl}), where semantic classes made up of phrases were used. These models were evaluated in term of WER and CER. When the M_w was considered the test set was recognized and the WER obtained, then, both recognized and reference sentences were classified in order to measure the value of CER. However, when M_{sw} and M_{sl} models were employed the WER was obtained in the same way from the output of the ASR system and, the classified sentence, obtained at the same time, was compared to the previously classified reference sentence, in order to obtain the CER. The results are shown in Table 2. Two different values of WER and CER are given for each model, one for the acoustic signal obtained with a close-talk microphone (m0) and the other one for the signal obtained with a lapel microphone (m1). Then, the same experiments using the same LMs were repeated but using the speaker dependent models. The results are shown in Table 4.

5.3 Results

The system was evaluated in terms of WER and CER in order to evaluate its ability to transcript and understand the sentence uttered by the speaker. The results obtained with the speaker independent acoustic models are shown in Table 2. It should be noticed that better values of WER were obtained when any of the class-based approaches was used, reaching an improvement of 19.1% with respect to the M_w model when the M_{sl} approach and the signal of the m0 microphone is used, as shown in Table 3. That means that the class-based LMs proposed in Section 4 take advantage of different information sources and improve the performance of the ASR system in terms of WER.

	Speaker Dependent Acoustic Models					
	M_w		M_{sw}		M_{sl}	
	WER	CER	WER	CER	WER	CER
m0	2.31	2.58	1.21	1.74	0.99	1.29
m1	2.47	2.79	1.73	1.41	1.52	1.32

Table 4: WER and CER for the m0 and m1 microphones, using speaker dependent acoustic models and different LMs: a word n-gram LM (M_w) and two different class-based LMs (M_{sw} and M_{sl})

	Speaker Dependent Acoustic Models			
	M_{sw}		M_{sl}	
	WER	CER	WER	CER
m0	47.6%	32.5%	57.1%	50.0%
m1	29.9%	49.4%	38.4%	52.7%

Table 5: Percentage of improvement in WER and CER for the two class-based LMs compared to the word n-gram LM using speaker dependent models

Comparing the values obtained with the close-talk (m0) and lapel (m1) microphones even though the same tendency is observed there is a significant difference in the rate of improvement between the two cases (1.5% vs. 15.3% for the M_{sw} approach and 8.2% vs. 19.1% for the M_{sl} approach as shown in Table 3). This difference may lie on the acoustic conditions, that is, when the m1 microphone is used the influence of the LM is not so significant because the acoustic signal has a higher level of noise. Regarding the CER results, the same tendency observed in the WER values is kept, however the improvements are much more significants (9.1% vs. 1.5% for M_{sw} approach and 17.2% vs. 8.2% for the M_{sl}). It can be concluded that some errors related to the semantic information of the understanding module are caused by recognition errors and they could be avoided by using the class-based LMs proposed in this work and by obtaining the semantic classes directly in the recognition process.

The results obtained with the use of Speaker Dependent Acoustic Models (Table 4) show a great improvement over the results with the speaker independent acoustic models. This reveals the good performance of the MLLR algorithm of speaker adaptation in this task. Improvements range from 5.81% of WER for the m0 microphone to 2.31% (60.2% less WER) and from 14.53% of WER for m1 microphone to 2.47% (82.6% less WER) when using the traditional word n-gram Language Model. Furthermore, comparing the WER and CER values achieved with the class-based LMs and word-based model, better results were attained with the class-based approaches again. Nevertheless, in this case the improvements are more significant than the ones obtained with speaker independent acoustic models, as shown in Table 3 and Table 5. This could be because of the acoustic conditions again, in this case the acoustic issue is better solved so the LM becomes more important and its effect is more noticeable.

The lower WER value is obtained with the M_{sl} approach and the m0 microphone. This result improves the one obtained with the M_w model and the same acoustic conditions by 57.1%. Moreover, comparing the best WER values (the

ones achieved with the M_{sl} model) with the ones obtained using the baseline acoustic and language models (speaker independent acoustic model and a word n-gram language model) the improvement reaches a 82.9% with the m0 microphone and a 89.5% with the m1.

It can be concluded from the obtained results that both the WER and CER results obtained with m1 microphone and the improved acoustic and language models (speaker dependent acoustic models and class-based LMs) are better than the ones obtained with the m0 microphone and baseline models (word-based LM and speaker independent acoustic models). Thus, the proposed models are appropriate to integrate in the ASR module that is inside the dialogue system employed in the home automation virtual butler. That is, the system has been adapted in order to work efficiently in the home automation environment where the speaker uses simply a lapel microphone (noisy acoustic signal) that allows to be in movement and speaks in a natural way.

6. CONCLUSIONS AND FUTURE WORK

In this paper, the improvements in the ASR part of a dialogue system for a speech-based human-machine interaction environment have been shown. The use of MLLR-based speaker adaptation for the 18 speakers used for evaluation in the task has achieved a significant reduction in the WER. On the other hand, the use of class-based LMs has also reduced the WER and CER providing not only a better recognized but also a better understood sentence.

Therefore, the proposed models are appropriate to adjust the dialogue system to a home automation environment where the acoustic conditions are not optimum and should consider the poblems derived from a user in movement and speaking in a natural way.

For further work it could be interesting to explore if the employed set of semantic classes could be replaced by other set of classes related to a different task. Thus, assuming that the words that do not belong to any semantic class are employed in the same way for different tasks, different applications could be managed by simply considering the semantic classes relevant to the new task and the phrases belonging to them without acquiring a new corpus and without training a new language model for it.

7. ACKNOWLEDGMENTS

This work has been supported by the national project TIN-2005-08660-C04-(01 and 03) from MEC of the Spanish government and by the University of the Basque Country under grant 9/UPV00224.310-15900/2004.

8. REFERENCES

[1] P. F. Brown, V. J. D. Pietra, P. V. d. Souza, J. C. Lai, and R. L. Mercer. Class–based n-gram Models of Natural Language. *Computational Linguistics*, 18(4):467–480, 1992.

[2] A.-P. Dempster, N.-M. Laird, and D.-B. Rubin. Maximum likelihhod for incomplete via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–21, 1977.

[3] S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech and signal Processing Society*, 29(2):254–272, February 1981.

[4] J.-L. Gauvain and C.-H. Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Trans. on Speech and Audio Proc.*, 2(2):291–298, April 1994.

[5] GENIO. Gestor Embebido Natural de Interfaz Oral. INTEK project. Industry Department. Basque Government. 2006.

[6] V. Gupta, M. Lenning, and P. Mermelstein. A language model for very large-vocabulary speech recognition. *Computer Speech and Language*, 6(2):331–344, 1992.

[7] F. Jelinek. *Statistical Methods for Speech Recognition*. The MIT Press, January 1998.

[8] R. Justo and M. I. Torres. Phrases in category-based language models for spanish and basque asr. In *Proceedings of Interspeech 2007*, pages 2377–2380, Antwerp, Belgium, Aug 2007.

[9] R. Justo and M. I. Torres. Two approaches to class-based language models for asr. In *Proceedings of IEEE International Workshop on Machine Learning for Signal Processing MLSP 07*, pages 235–240, Thessaloniki, Greece, Aug 2007.

[10] S. Kullback and R.-A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[11] C.-J. Legetter and P.-C. Woodland. Maximum likelihood linear regression for speaker adaptation of the parameters of continous density hidden markov models. *Computer Speech and Language*, 9:171–185, 1995.

[12] A. Moreno, A. Nogueira, and A. Sesma. Speechdat-car: Spanish. *Technical Report SpeechDat*.

[13] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J.-B. Marino, and C. Nadeu. Albayzin speech database: Design of the phonetic corpus. In *Proceedings of the Third Eurospeech*, Berlin, Germany, September 1993.

[14] T. R. Niesler and P. C. Woodland. A variable-length category-based n-gram language model. In *IEEE ICASSP-96*, volume I, pages 164–167, Atlanta, GA, 1996. IEEE.

[15] K. Ries, F. D. Buo, and A. Waibel. Class phrase models for language modelling. In *Proc. ICSLP '96*, volume 1, pages 398–401, Philadelphia, PA, oct 1996.

[16] S. Seneff and J. Polifroni. Dialogue management in the mercury flight reservation system. In *ANLP-NAACL 2000 Satellite Workshop.*, pages 1–6, 2000.

[17] A. Uria, A. Ortega, M.-I. Torres, A. Miguel, V. Guijarrubia, L. Buera, J. Garmendia, E. Lleida, O. Aizpuru, A. Varona, E. Alonso, and O. Saz. A virtual butler controlled by speech. In *Proceedings of the III Jornadas en Tecnologías del Habla*, Zaragoza, Spain, November 2006.

[18] V. Zue, S. Seneff, J. Glass, J. Polifroni, C. Pao, T. Hazen, and L. Hetherington. Jupiter: A telephone-based conversational interface for weather information. In *IEEE Trans. on Speech and Audio Proc.*, pages 8(1):85–96, 2000.