Intelligent Automation and Soft Computing, Vol. 15, No. 4, pp. 645-664, 2009

SPEECH TECHNOLOGY AT HOME: ENHANCED INTERFACES FOR PEOPLE WITH DISABILITIES

R. SAN-SEGUNDO[1], R. CORDOBA[1], J. FERREIROS[1], J. MACIAS-GUARASA[2], J.M. MONTERO[1], F. FERNÁNDEZ[1], L.F. D'HARO[1], R. BARRA[1] AND J.M. PARDO[1]

[1]*Speech Technology Group*
Department of Electronic Engineering
ETSI Telecomunicación
Universidad Politécnica de Madrid. Spain

[2]*Department of Electronics*
University of Alcalá. Spain

ABSTRACT—This paper presents new advances in speech technology carried out by the Speech Technology Group (GTH) at the Universidad Politécnica de Madrid (UPM) to develop enhanced interfaces at home. These interfaces provide a better interaction for people with disabilities. The speech recognizer includes a speaker identification feature (that makes an acoustic adaptation possible for improving recognition performance) and an emotion classifier (to detect the user emotion). The understanding module, with a bottom-up strategy, increases its flexibility against recognition errors. The dialog manager has been improved by a new dialog control based on Bayesian Networks and a new platform for developing multimodal, multilingual, and user dependent dialog services from scratch. Finally, the speech synthesis module includes new advances for increasing the voice naturalness and incorporating emotions. These advances have been integrated into a new interface for controlling a Hi-Fi audio system, thus significantly increasing its ergonomics.

Key Words: speech technology, enhanced interfaces, Assistive Technology (AT), disable people.

1. INTRODUCTION

There is currently an increasing interest in incorporating Assistive Technology (AT) at home. This interest is based on the significant number of people that need some kind of help to guarantee their autonomy at home. In Europe there are more than 50 million people with some kind of disability (6-7% of the total population) [1]. On the other hand, as a result of an increase in life expectancy and a decrease in the fertility rate, the percentage of elderly people is growing very fast reaching 50% between now and 2025. In Europe, at the moment, 20% of the population is over 65 (more than 100 million people).

Assistive Technology at home has four main targets: people safety (i.e. using gas and water sensors), better use of natural resources, entertainment, and a better interaction with household appliances. In the last area, speech technology can play an important role in developing advanced interfaces that provide a better interaction for people with mobility or vision problems. Nowadays, speech technology has reached a significant level of performance and is being used in end-user applications [2]. The improvements have been achieved thanks to the immense effort in research

carried out by telephone companies and research centers. As a result of this effort, large speech and text databases have been generated and new speech and text processing models have been developed. The advances in this technology have been supported by the significant increase in speed obtained in the hardware.

From 1975, the Speech Technology Group (GTH) at UPM has developed speech technology and enhanced interfaces for people with disabilities. This paper presents the newest advances and their application for controlling appliances (for example a Hi-Fi audio system). The enhanced interface developed allows users to control a Hi-Fi audio system with natural spoken language, against other speech interfaces based on simple commands. In this case, users can speak naturally (i.e. they can request several actions in a spoken sentence) neither do they have to memorize any command list nor use a specific phraseology to control the Hi-Fi audio system. For the prototype developed, the audio system is a commercial system made up of a compact disc (with three discs loader), two tapes and a radio receiver. This system is controlled by an infrared (IR) remote control. Figure 1 shows a block diagram of the speech interface consisting of six modules [3]:

- The speech recognizer, converts natural speech into a sequence of words (text) using both acoustic and language models. This module also allows the speaker to be identified and adapt the models to increase their performance (described in Section 2).
- The Natural Language Understanding module extracts the main semantic concepts from the text. In this process, it uses several semantic rules. This module is described in Section 3.
- The Dialog Manager controls the interaction flow and defines the actions carried out over the Audio System. The dialog control technology is described in Section 4.
- The Execution Module translates the actions defined by the Dialog Manager into Infra Red (IR) commands. This translation is carried out based on a mapping table, where one action is translated into one or several IR commands.
- The Response Generation Module uses response templates to create a natural language sentence as system response.
- Finally, the Speech Synthesis Module converts the response sentence into natural speech. The technology of the new speech synthesis module including emotion is described in Section 5.

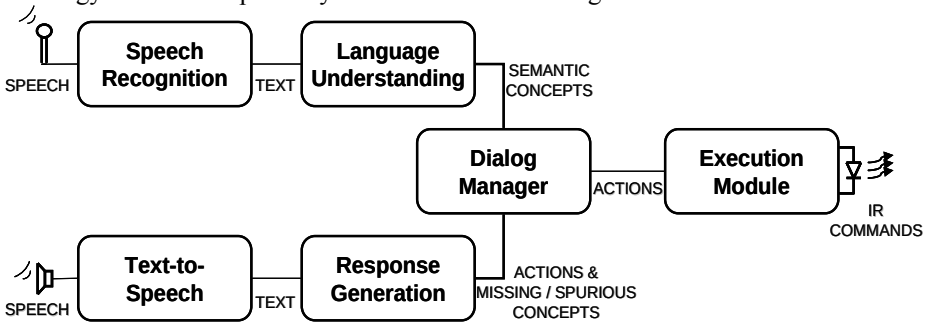

Figure 1. Block diagram of the interface

2. SPEECH AND SPEAKER RECOGNITION

This section describes both the speaker and speech recognition and the acoustic model adaptation.

2.1. Speaker Identification

Speaker identification is crucial as long as you want to introduce more intelligence in the behavior of speech interfaces. Two main targets are aimed at with this technology:

- First, with speaker identification it is possible to use user-adapted acoustic and language models, a characteristic that will lead the system to higher speech recognition and understanding accuracies. This aspect increases the user acceptability and quality perception of the interface.
- In addition, once the personality of the speaker has been identified, user-adapted interpretation rules can be used, dependent on some saved user profile that will enhance the friendliness of the system: the system can adopt some learned behavior that was preferred by the user previously. Of course, the user may correct this choice, but on average, the system saves many dialog turns by carrying out the actions within the restrictions established by the user (without the need to be specified again at each iteration). Even dialog characteristics like the answering expectancies, the complexity of the response messages or the speed or the volume of the speech produced can be adapted to the user.

In speech interfaces at home, the speaker identification task is simpler than a personality verification one (i.e. needed in a banking application). In fact, most of the time, very limited population of identities will access a household interface (only the people living in the house). Nevertheless, it is desirable to recognize and deal with new users appropriately (such as occasional guests). The solution implemented in GTH uses the Bayesian Information Criterion (BIC) for evaluating the distance between two acoustic speech segments. A special clustering algorithm is used for assigning each segment to one of the previously recorded clusters (thus recognizing the speaker's identity) or for generating a new one if none of the preexistent clusters accepts the current intervention (a new person is using the system).

Initial work with this technology was on Acoustic Change Detection (ACD) tasks on the Broadcast News database [4][5], where different acoustic segments from 218 speakers that could appear with several acoustic conditions (such as a telephone line, studio quality, background street noise, voice mixed with music, etc.) had to be segmented, marking the time change points. This experience showed that our ACD and clustering algorithms combined satisfactory giving as low as an 18% EER (Equal Error Rate: the point where false acceptance equals false rejection of true ACD hypothesis) in such difficult work. In the following work [6], similar performances were obtained: one with a similar ACD task on lecture recordings with some short student interventions. In this case, a clustering algorithm was used as a labeler of the different segments cut by the ACD system, building a full speaker identification system. A similar work was carried out on a segmentation and clustering task of a radio news program with about 12 different speakers. The system obtained a very high precision (more than 90%) matching very short segments (1 s. of speech).

Finally, for a speech interface at home, it is necessary to implement a procedure that allow dynamic auto-creation of the clusters (models) for new incoming speakers and a new user profile preparation for saving the preferred personal choices. If the voice is not identified as previously known, the dialog system will be informed and it can talk to the person to give initial greetings and obtain personal information like the name (which will be used in future dialogs with this speaker).

2.2. Speech Recognition

The speech recognition process is responsible for converting the speech waveform into the most likely sequence of words corresponding to user utterance (Figure 2). The feature extraction

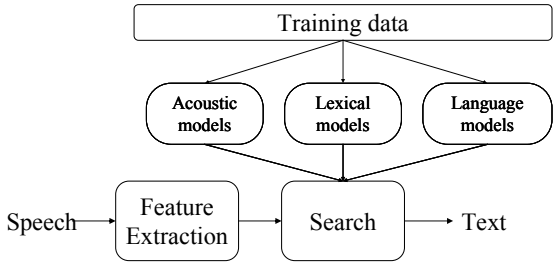

Figure 2. Block diagram of a speech recognition system

module converts the input speech into a sequence of feature vectors, which are used during the search stage under some sort of probabilistic formulation, integrating previously trained sources of information (acoustic, lexical and language models). The system uses a feature vector with Perceptual Linear Predictive (PLP) parameters in the Cepstral domain derived from a Mel-scale filter bank (Mel-PLP), with 13 coefficients including their first and second-order differentials, giving a total of 39 parameters for each 10 ms. frame. As the channel conditions become noisy, the system incorporates two normalization techniques that are specially designed to compensate for channel variations: Cepstral Mean Normalization (CMN) and Cepstral Variance Normalization (CVN). All these aspects are described in [2].

State of the art systems are usually based in some form of progressive search [7], whereby successively more detailed (and computationally expensive) knowledge sources are brought to bear on the recognition search as the hypothesis space is narrowed down. Current commercial Automatic Speech Recognition (ASR) systems are known to perform reasonably well when the speech signals are captured in a noise-free environment using a close-talk microphone located near the mouth of the speaker. Moreover, the user is assumed to speak using a reasonably consistent pronunciation structure, which is unlikely to happen in certain cases, such as people affected from motor related disorders. These two restrictions open up exciting challenges for future research into speech recognition technology.

The restriction related to the use of close-talk microphones can be solved by using fixed microphones placed at some distance from the user (necessary for speech interfaces at home). Unfortunately, as the distance between the user and the microphone grows, the speech signal becomes increasingly degraded by the effects of additive noise and reverberation, which in turn severely degrade speech recognition accuracy. In these distant-talking environments, the use of an array of microphones, rather than a single microphone, and advanced processing techniques offer an increasingly viable alternative which overcomes many of the disadvantages of close talking microphones [8]. Traditional speech recognition systems with microphone arrays use a two-stage strategy: array processing and then recognition [9], in which the array processing is considered as a pre-processing stage for signal enhancement. The first stage typically involves beam forming: filtering and combining the individual microphone signals to enhance signals coming from a particular location [10]. To operate properly, beam forming is usually preceded by speaker location or tracking [11]. Further improvements can be achieved by post-filtering the output of the beam former [12].

Eliminating the second restriction (assuming a consistent pronunciation by the speaker) would allow people with speech impediments to use speech recognition systems [13]. Although the nature of the statistical methods used in speech recognition technology is capable of capturing acoustic variations in speech utterances, impaired speech poses a major challenge for current systems, and their performance is severely reduced in these cases. Automatic speech recognition

of people with speech impediments, such as cerebral palsy patients, requires a robust technique that can handle conditions of very high variability and limited training data. In the literature one can find few references related to this problem and most of them show studies on limited tasks (small to medium vocabulary and isolated word recognition) [14][15].

This problem is related to that of dealing with alternative pronunciations made by speakers using certain dialects. In this area, a number of strategies have been proposed in the literature (an excellent revision can be found in [16]), but it is difficult to extract definitive conclusions on the actual usefulness of modeling pronunciation variations, as few improvements are typically reported [17].

2.3. Acoustic Adaptation

Speech recognition can be considered a very difficult problem in real-life environments because of several factors: the great variability between speakers and even for the same speaker resulting from stress amongst other things, significant variations between channels and/or environments, the presence of background noise, etc. The most effective strategy probably consists of adapting the ASR system to the speaker (when the speaker has been identified), specifically the acoustic model.

2.3.1 Adaptation Techniques

This paper focuses on the adaptation of acoustic models with two main adaptation techniques:

- **Maximum A Posteriori (MAP) estimation [18].** MAP adaptation involves the use of prior knowledge about the model parameter distribution. The idea is to use a previously well-trained model as the prior knowledge. The adaptation formula for the mean vectors of the Gaussian distribution generates an interpolation of the mean vector as the final mean vector for the prior model and the mean of the observed adaptation data (a similar formula is applied to the other model parameters). More details can be found in [18]. When enough adaptation data is available, the MAP estimation converges to the maximum likelihood criterion, which is the optimum for estimating the parameters from scratch. So, it should be the best technique for large adaptation sets. But, MAP does not modify the parameters that do not appear in the adaptation data, so it is a bad choice for small adaptation sets.
- **Maximum Likelihood Linear Regression (MLLR) [19].** In MLLR, a set of transformations for the model parameters is computed which reduces the mismatch between an initial model set and the adaptation data. The effect of these transformations is to modify the mean and variance component of the Gaussian model to model more likely the adaptation data. The mathematics behind the transformation matrix is complex, so the reader should consult [19] to see the details. The main idea is that the transformation matrix is obtained by solving a maximization problem with the Expectation-Maximization (EM) technique that uses the likelihood of the adaptation data as the maximization criterion. One important issue is that it is not feasible to compute a transformation matrix for every unit in the model set. The solution is to group the most similar units given a similarity measure or distance between units and then compute a common transform for all of them. The main advantage of MLLR is that it shares the transforms between similar units, so that every parameter in the model set gets updated in the adaptation process even though it does not appear in the adaptation set. So, this technique should be better than MAP for smaller adaptation sets (medium size).

In [20] there is a detailed explanation and its application in an Air Traffic Control system (ATC).

2.3.2. Speech Database and Experiments

The database is made up of 206 speakers that uttered several different types of sentences: isolated speech (commands to control a robot, city and street names for a Global Positioning System –GPS- navigator), continuous speech (addresses, movie names to control a home media center, orders including street names), and spontaneous speech: addresses and movie names in a very spontaneous style. On average, there are 10 minutes of speech for each speaker. The recording conditions are optimum: clean speech and a close-talking microphone. All systems use context-dependent continuous Hidden Markov Models (HMMs) (these models consider the adjacent allophones to model the current one) built using decision-tree state clustering. The feature extraction module is the same as that presented in section 2.2 for all systems.

From the database, 110 speakers have been used to train a general speaker model, and the rest for testing the adaptation. The data available for each speaker has been divided into several sets: 25% for validation and 75% as adaptation material. To evaluate how the amount of adaptation data affects the recognition performance, this set has been subdivided into 4 sets, using 25%, 50%, 75%, and 100% of the adaptation material. A greedy algorithm has been used to select the sentences that best fit the desired phonetic distribution in the system. This way ensures the best possible coverage of phonemes. Tables I and II shows the results in Word Error Rate (WER) and improvement using MAP and MLLR respectively.

Table I. Word Error Rate (WER) and achieved improvements when adapting with MAP.

Adaptation data	Isolated		Continuous		Spontaneous	
	WER	Improv.	WER	Improv.	WER	Improv.
No adaptation	0.41	-	3.16	-	7.51	-
25% set	0.41	0%	3.06	3.2%	6.67	11.2%
50% set	0.32	22%	2.96	6.3%	6.77	9.8%
75% set	0.20	51%	2.98	5.7%	6.93	7.7%
100% set	0.15	63%	3.00	5.1%	7.01	6.8%

Table II. Word Error Rate (WER) and achieved improvement when adapting with MLLR.

Adaptation data	Isolated		Continuous		Spontaneous	
	WER	Improv.	WER	Improv.	WER	Improv.
No adaptation	0.41	-	3.16	-	7.51	-
25% set	0.19	54%	3.00	5.1%	6.04	18.9%
50% set	0.17	59%	3.08	2.5%	6.30	18.1%
75% set	0.13	68%	3.10	1.9%	6.46	13.3%
100% set	0.09	78%	2.97	6.0%	6.52	12.6%

For isolated speech, results improve drastically as expected. As there are more adaptation data, the improvement is more remarkable. MLLR behaves better than MAP, especially with less data. So, with just a few sentences, MLLR should always be selected as the optimum technique.

For continuous and spontaneous speech, results show no improvement or improve very little by using bigger adaptation sets. This is unexpected, but the reason is that the greedy algorithm that is used to select the adaptation sets has included most spontaneous sentences in the first set, so very little improvement is obtained when using the whole set. Another possible explanation is that there are more sentences for isolated speech, and there is an adaptation not only to the speaker

but also to the speaking style, which is quite different for spontaneous speech. In summary, MLLR adaptation is recommended in our case.

2.4. User's Emotion Identification

This section presents the first experiments in user's emotion identification in an enhanced speech interface. In this work, the Spanish Emotional Speech corpus (SES) has been used. It contains three emotional speech-recording sessions played by a professional male actor in an acoustically treated studio. Each recorded session includes thirty words, fifteen short sentences and four paragraphs, simulating four basic or primary emotions (sadness, happiness, surprise and anger) and a neutral speaking style. The text uttered by the actor did not convey any intrinsic emotional content. Finally, the recorded database was phonetically labeled semi-automatically. The assessment of the emotional speech was aimed at judging the appropriateness of the recordings as a model for recognizable emotional speech.

The features used for emotion classification have been some segmental features (thirteen Mel Frequency Cepstrum Coefficients (MFCC)) and supra-segmental features (several statistics calculated from F0 contour of voice segments during a sentence, average F0, F0 standard deviation, F0 average variation, minimum F0, maximum F0 and F0 range). The module uses a 128-VQ Bayes-Classifier for the MFCC based identification task. Table III shows the confusion matrices obtained when identifying the underlying emotion: Sadness, Anger and Happiness are the highest identified emotions (100%, 97.8% and 91.1%, respectively), as opposed to Surprise which is the least identified (68.9%).

Table III. Confusion matrix when using the prosodic features for automatic emotion identification

	IDENTIFIED EMOTION				
INTENDED EMOTION	**Happiness**	**Anger**	**Surprise**	**Sadness**	**Neutral**
Happiness	**91.1%**		6.7%		2.2%
Anger		**97.8%**			2.2%
Surprise	2.2%	28.9%	**68.9%**		
Sadness				**100.0%**	
Neutral		26.7%			**76.3%**

In further experiments, 21 people were used in a perceptual experiment that consisted of identifying the emotion simulated by the actor. Considering these experiments, the emotion classifier developed in this work emulates the perceptual experiment with an average correlation of 81.5%. This automatic classification helps the system to select different dialog strategies (depending on the user emotion) and to know which parameters of the Text To Speech (TTS) must be modified to synthesise emotional speech.

3. SPEECH UNDERSTANDING

This process is responsible for extracting the semantic information or "meaning" (related to the specific application domain) from the speech recogniser output. The semantic information is conveyed using semantics concepts. A semantic concept consists of an attribute (identifier) and a value: there is a concept *VOLUME* while the value is "5" on a 1-10 scale. For controlling a Hi-Fi audio system, the semantic concepts can be actions to be carried out with the hi-fi system (i.e. play) or system parameters that can be configured in the system, (i.e. volume). In the prototype, there are 15 actions and 44 parameters. Internally, the system manages other concepts that carry

the semantic information as it is being developed. The understanding module creates a semantic frame made up of a variable set of concepts, each one with its pair of attributes (identifier) and corresponding value with the appropriate format.

In this prototype, the speech understanding technology is a rule-based technique considering a bottom-up strategy. In this case, the relationships between semantic concepts and word and/or concept sequences are defined by hand by an expert. In a bottom-up strategy, the semantic analysis starts from each word individually and extending the analysis to neighbourhood context words or already-formed concepts. This extension is carried out to find specific combinations of words and/or concepts that generate another concept. Not all of the words contribute to the formation of the final interpretation. This strategy is more robust against speech recognition errors and it is frequently preferred when an N-gram language model is used in the recognizer, as in our case. Depending on the scope of the word relationships defined by the rules, it is possible to achieve different compromises between reliability of the concept extracted (greater with greater lengths) and the robustness against recognition errors (greater with smaller lengths).

The understanding process is carried out in two steps (Figure 3): semantic tagging and rule application. In the first step, one or several syntactic-pragmatic tags are assigned to every word in the vocabulary (i.e. "cero, uno, dos,… (one, two, three,…) " are assigned the "numero (number)" tag or the verb "reproducir (play)" is assigned the "acc_repr (action)" tag). The understanding module works by applying different rules that convert the tagged words into semantic concepts and values.

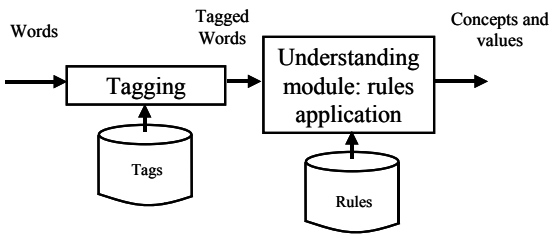

Figure 3. Structure of the natural language understanding module

3.1 Concept Confidence Measure

The prototype developed is one of the first understanding modules that generates one confidence value for every concept obtained [6]. The confidence measure is a value between 0.0 (lowest confidence) and 1.0 (highest confidence). This concept confidence is computed by an internal procedure that is coded inside the GTH proprietary language interpreter that executes each rule. In this internal engine, there are "primitive functions," responsible for the execution of the rules. Each primitive has its own way of generating the confidence for the elements it produces (considering the confidence of the input elements). One common case is for the primitives that check for the existence of a sequence of semantic blocks to generate new ones, where the primitive usually assigns to the newly created concepts the average confidence of the block, which it has relied on. In other more complex cases, the confidence for the new concepts depends on a combination of confidences from a mixture of words and/or internal or final concepts. In order to generate concept confidence every word from the speech recognizer output must be assigned a confidence value [21]. The concept confidence measures can be used for incorporating confidence filters to avoid generating wrong concepts when their confidence values are very low. This strategy improves the Concept Accuracy, percentage of concepts correctly

extracted, from 76.6% to 80.5%. Secondly, the Dialog Manager can use concept confidence measures to define the interaction flow: i.e. reducing the confirmation turns when the concept confidence is very high.

4. DIALOG MANAGEMENT

This section presents the two main advances corresponding to the Dialog Manager. The first one is the dialog control based on Inference Bayesian Networks, and the second is a complete platform for developing multimodal, multilingual and user-dependent dialog services from scratch based on a Finite State Automata.

4.1. Management Based on Bayesian Networks

The Dialog Manager module is the responsible for identifying the user's goals from the processed utterances (in the context of controlling electronic domestic devices, such goals could be mainly the execution of some commands, i.e. actions carried out on the HIFI System), and for detecting missing, wrong, spurious and required concepts given an identified goal. This information is used to drive the dialog prompting for missing concepts, clarifying wrong concepts and ignoring the spurious concepts thus allowing more flexible and natural dialogs. This paper proposes Bayesian Belief Networks approach [3][22] for dialog modelling. This section describes how this approach carries out these two main tasks.

4.1.1 Goal Detection from Processed Utterances

From the semantic concepts extracted from the user's utterances, as well as those retrieved from the Dialog History (i.e. a record of the evolving dialog in the form of a stack of referred concepts), it is possible to infer the dialog goal(s). A goal is considered as a specific action on the Hi-Fi system, e.g. to set the volume to a specific level. In the prototype developed, a set of 20 goals and a set of 70 concepts (i.e. actions, parameters and values) have been defined by an expert in the application domain. When using Belief Networks (BNs) for dialog modeling, the first solution is to develop one BN per goal.

A BN is a directed acyclic graph with nodes and arcs where the direction of the arcs represents the probabilistic dependency between two nodes. The arrows of the acyclic graph are drawn from cause to effect. Assuming the basic topology (Figure 4), the BN models the causal relation between the goal and some concepts. This topology assumes conditional independence between concepts. Each BN is defined by a specific goal G_i and a set of input concepts $\{C_1,...C_M\}$. In this approach, the goals and the concepts are all binary, so the concept C_j is true (C_j=1) when it is extracted from the user utterance. In order to avoid too complex models, the expert can select only the concepts with the strongest dependency for each goal as its inputs. Hence, N (N=20) binary decisions are made (considering N BNs) in the presence or absence for each goal. From observations extracted from the user's sentence, i.e. $C=\{C_1=0, C_2=1,\ldots, C_M=1\}$, the Bayesian Inference obtains the later probability $P(G_i|C)$ for each goal (see Equation 1, it simply applies Bayes' Theorem assuming marginal and conditional independence; M is the number of input evidences).

$$P\left(G_i = 1 \middle| \underline{C}\right) = P\left(G_i = 1\right) \prod_{k=1}^{M} \frac{P\left(C_k = c_k \middle| G_i = 1\right)}{P\left(C_k = c_k\right)} \tag{1}$$

$$where\ \underline{C} = \left\{C_1 = c_1, C_1 = c_2, ..., C_M = c_M\right\}$$

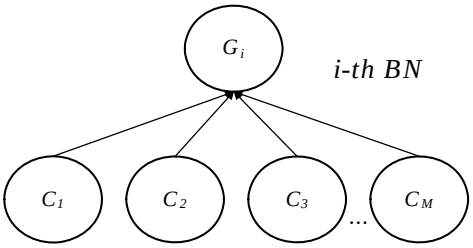

Figure 4. Basic Topology for a BN.

Comparing this probability with a defined threshold makes a binary decision: one goal is present or active if the later probability of the corresponding BN is greater than the threshold; otherwise the goal is absent. For simplicity, the threshold may be set to 0.5 since $P(G_i=1|C)+P(G_i=0|C)=1$.

This approach assumes a multiple goal evaluation scheme, thus multiple goals can be active if they vote positive. Moreover, it is possible to identify Out Of Domain (OOD) sentences: those sentences for which all BNs vote negative. In this paper, an enhanced topology is presented adding some links between concept nodes (according to the expert's criterion) to model the inter-concept dependencies. These links introduce certain variations in the probability propagation for goal inference [23]. In our prototype, the probabilities involved in the inference process (left-hand side of Equation 1) have been hand-assigned by the expert too. In [24], Meng et al. present a Minimum Description Length approach for learning topologies automatically. The conditional probabilities for each BN are estimated by tallying the counts from training data.

4.1.2 Detection of Missing, Wrong, Spurious and Required Concepts

This process is carried out by the "Backward Inference" technique [22]. In this case, considering the inferred goals (i.e. those which are present), it is necessary to test, for the corresponding BNs, their probability for each input concept. The inferred result is assumed, i.e. $G_i=1$, as a new evidence to add to the observations' vector. Then Bayesian inference is applied again but this time aimed at the estimation of $P(C_i|C')$ (the updated concept's probability) where $C'=\{G_i=1, C_1=0, C_2=1,\ldots, C_M=1\}$. Now, two thresholds are defined, Θ_{LOW} and Θ_{HIGH}, resulting in three different levels for that probability: low, medium and high. These thresholds are estimated from previous experiments by analyzing the probability distribution. Just comparing the probability with the defined thresholds, the concepts are classified according to the Backward Inference result. Based on the value of $P(C_i|C')$, the system checks whether this concept should be present (i.e. $P(C_j|G_i, C)>\Theta_{HIGH}$), absent (i.e. $P(C_j|C')<\Theta_{LOW}$) or neither of them (i.e. optional or spurious).

Then the result of this decision is compared to the actual occurrence of the concept in the observations' vector and its confidence value. The system checks that every concept obtained from the understanding process is assigned with a confidence value, otherwise if a particular concept is not observed no confidence value is available. In the analysis, a predefined pair of thresholds is used in order to classify the confidence level. For simplicity, the same thresholds are considered (Θ_{LOW} and Θ_{HIGH}).

The whole analysis resulting from the comparison between later and confidence values is summarized in Table IV. As a result of this analysis, the system automatically detects missing, wrong or spurious concepts so that the system can drive the dialog prompting for missing concepts, clarifying for wrong concepts, and ignoring the spurious (i.e. optional) ones.

Table IV. Concept analysis used to drive the dialog

		Posteriors		
Evidences	Confidence	$P(C_j\|C') < \Theta_{LOW}$	$\Theta_{LOW} \leq P(C_j\|C') \leq \Theta_{HIGH}$	$P(C_j\|C') > \Theta_{HIGH}$
C_j absent	-	**No action (C_j unnecessary)**	**No action (C_j optional)**	**Prompt to request C_j (C_j missing)**
C_j present	Conf. $C_j < \Theta_{LOW}$	**Ignore C_j (C_j wrong)**	**Ignore C_j (C_j optional)**	**Explicit confirmation of C_j (C_j required)**
	$\Theta_{LOW} \leq$ Conf. $C_j \leq \Theta_{HIGH}$	**Ignore C_j (C_j wrong)**	**Ignore C_j (C_j optional)**	**Explicit confirmation of C_j (C_j required)**
	Conf. $C_j > \Theta_{HIGH}$	**Prompt to clarify C_j (C_j wrong)**	**Ignore C_j (C_j optional)**	**Implicit confirmation of C_j (C_j required)**

- Those concepts whose latter probabilities indicate that they should be neither absent nor present (i.e. $\Theta_{LOW} < P(Cj|C') \leq \Theta_{HIGH}$) are going to be regarded as spurious or optional.
- Concepts that should be absent according to their latter probabilities, the dialog manager identifies them as unnecessary concepts (i.e. when the concept is actually absent) and wrong concepts (i.e. when the concept is available or present). Only if the concept is actually present in the input sentence, is its confidence value high, and it has been labelled as a wrong one, the dialog manager will invoke a clarification act. For any other confidence level, the concepts will be ignored.
- Regarding concepts that should be present according to their latter probabilities, missing concepts (i.e. when the concept is not available or absent), and required concepts (i.e. when the concept is available or present but with low confidence) can be identified as well. Depending on the confidence level, the system distinguishes the condition of these concepts. The dialog manager uses implicit confirmation (the system informs the user about this concept) just for those concepts with a high confidence level. On the other hand, if the corresponding confidence value is not high enough, then that concept has to be confirmed through an explicit confirmation procedure (the system asks the user about the concept correctness). Finally, if the concept is not available (i.e. missing), the dialog manager decides to prompt the user for it.

Once the user has provided and confirmed all the required information, the system is able to complete the goal: to carry out a specific action or sequence of actions on the Hi-Fi audio System. Finally, all confirmed concepts are stored in the Dialog History as consolidated knowledge and the state of the audio system is updated. From this point, the user can start a new dialog to carry out another action on the system.

4.2. Management Based on Finite State Automata

Given the growing interest in spoken dialogue systems, a large number of commercial and non-commercial tools have been developed in recent years: CSLU's RAD toolkit from Oregon University [25], Smartkom [26], SpeechBuilder [27] from MIT, OpenSpeech from Nuance, WebSphere Voice Server from IBM or Audium studio. Most of them support the creation of multimodal dialogue systems, and allow a quick development thanks to the use of libraries, a user-friendly graphical interface and a full integration with proprietary run-time platforms (TTS and ASR); however, they present difficulties when creating multilingual services, for handling multiple user profiles, and for including 'intelligent' assistants to help designers to define the

dialogue flow, the backend integration and for solving specific modality issues (e.g. presentation of lists of results or system confirmations).

4.2.1 Semiautomatic Application Generation Platform (AGP)

Given these restrictions, GTH was involved in the GEMINI (IST -2001-32343) European project for creating a semiautomatic platform called AGP [28] which allows the simulatively creation of dialogue applications in four languages and two modalities (Web and Voice). The main contributions were:

- A modular architecture of the platform (Figure 5), separating the high level dialogue flow from the specific characteristics of the service and from the multimodal, multilingual and user specific information. Thus, the designer can easily specify a common flow for all the languages and modalities and in the following assistants include the specific information for each one.
- Several accelerating strategies for speeding up the design in the most critical modules.
- The use of several standard languages, such as xHTML for Web and VoiceXML [29] for speech, which allows the portability and simplifies the use of different devices and execution platforms.

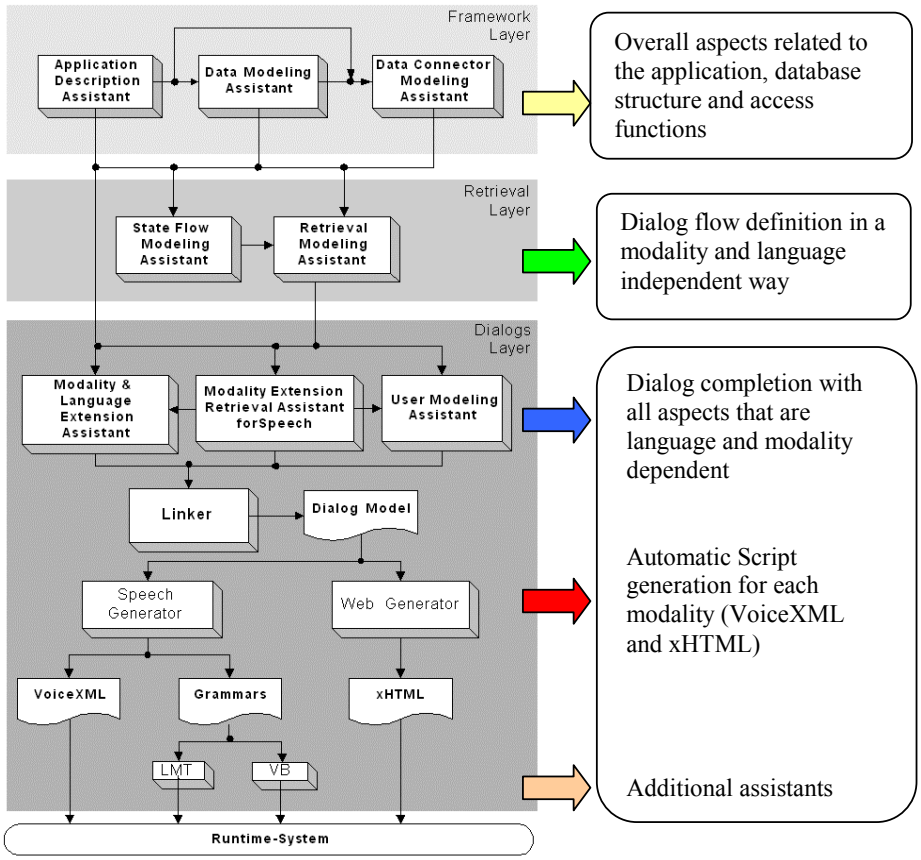

Figure 5. AGP architecture

To rate the acceptance and friendliness of the platform, a subjective evaluation was carried out with 41 subjects (24 novices in dialog application design, 11 intermediate and 6 experts) from Greece, Germany and Spain, with ages ranging from 21 to 49 and, in general, with basic knowledge of programming. Their overall score had an average of 8.4 (on a 0-10 scale) with the maximum scores in the following aspects: speeding up the development time of an application, over-answering and mixed initiative functionality, and list handling for speech applications. The AGP also proved to be both flexible and comprehensive, offering a 75% saving in time/costs over training a human operator for the same task.

4.2.2 Multimodality

One of the main problems with the human-machine dialogue applications is the limitations of the system when the environment is very noisy or the user has physical limitations. One solution is to use several complementary modalities to show or retrieve information to/from the user. With several modalities, the system increases the probability of understanding the user input and the probability that the user receives the message from the system. Moreover, the inclusion of several modalities allows more natural and shorter interactions. Some modalities are visual (Using images, webcams, animated agents, field forms on a Web page), speech (Text-To-Speech, automatic recognition), gestures (Pointers, mousse, graphical tables, etc.) and writing (through keyboards, handwriting or character recognition, etc.)[30].

In AGP [28], the designer has the possibility of creating the same service for two modalities simultaneously: Web and Speech. The Web modality allows the user to fill out web forms using a conventional internet browser. From the xHTML platform for the final script, the designer can include more functionalities such as: animations, interactive maps, videos, etc. that are common in other web pages. For voice modality, the platform automatically generates a script in VoiceXML, which is very common for creating call centers. The current work focuses on the integration of the two current modalities using the X+V language [31] that is a W3C standard which combines visual and spoken interaction using Web technologies (xHTML, XML events and vocabulary) and VoiceXML. Another future work is the integration of an animated agent for helping Deaf people [32].

4.2.3 User Modeling and Multilingualism

In order to provide more natural and friendly dialogue services, the dialogue manager has to be able to adapt its behaviour according to the language, communication style, experience and users' age. Obviously, the inclusion of all this customization implies an additional step in the design. The AGP currently supports the creation of the dialog flow for four languages: English, Greek, German and Spanish. The platform has a specific module where the designer can customize the grammars, prompts, error correction and help dialogues according to the particular language and user level. The system can automatically detect the spoken language (using PPRLM [33]) and user skill (considering several parameters such as user and language identification, number of errors and time duration of the interaction, etc). With this information, the system sets the corresponding specific overall variables in the VoiceXML running script.

5. HIGH QUALITY SPEECH SYNTHESIS

Although speech communication bandwidth is lower than in graphical or visual interfaces, when freedom of movement and far-interaction are needed, speech synthesis is the must-be choice. If the application domain is static and very constrained (only a few patterns to synthesise) and if the same voice can be used in many implemented systems (without any personalization), the speech synthesis module can use a simple but effective technique: concatenation of natural

speech recordings of a professional speaker or actor. However, when the domain is dynamic or not so limited, fully automatic text-to-speech (TTS) conversion is necessary. Artificial TTS systems are made up of three main modules:

- A linguistic and prosodic processor: which analyses the input text to label the most relevant linguistic features that affect the way of speaking (stress, phonetic transcription, etc). These features allow suprasegmental information to be generated (intonation, rhythm and intensity curves).
- A segmental synthesiser: which generates speech from the linguistic and prosodic information, using the speaker database and an algorithm for prosodic and articulatory modification.
- A speech adaptation module: to alter some characteristics of the basic voice (or the modified speech) such as intended emotion/attitude, sex or age of the artificial speaker.

The most important evaluation parameters of synthetic speech are intelligibility (how understandable the utterances are) and naturalness (how similar are human and artificial utterances). Although every module affects both characteristics, intelligibility is mostly affected by the synthesis technique (for a wide range of prosodic curves and non-phonetic features). The most intelligible technique is based on waveform concatenation of variable-length recorded segments (from demiphones or diphones to poliphones or even multi-word units), with a restricted amount of prosody modification (or without any modification at all) and unit-selection process to minimize concatenation artifacts. This concatenation technique also provides the most natural and distinctive timbre, because the segmental component is copied from a natural source with a minimum amount of transformation. Emotion-related phenomena such as laughter and sighs can easily be synthesised. However, the range of prosodic modifications is rather limited, thus reducing the expressivity.

In a highly interactive domestic environment with only a few users and a few locatable noise sources, naturalness and variability are even more important than standard listener-independent segmental intelligibility. To maximize naturalness, the TTS system must provide synthetic speech with accurate domain-dependent prosodic modelling and context-adapted emotional capabilities.

5.1. Restricted Domain Prosody Experiments

The main objective has been to get good predictors for both the F0 curve and phoneme duration by minimizing the model estimation error in a Spanish text-to-speech system. To achieve this, the factors that mostly influence prosodic values in Spanish need to be determined. To minimize the cost of adapting the system to a set of new domains or sub-domains, a machine learning technique must be used [34]. Artificial Neural Networks (ANNs) and k-nearest-neighbour (k-NN) are the techniques that have been able to model both duration (rhythm) and F0 (intonation) successfully. In our previous experiments, ANNs outperform k-NN in prosody-related tasks [35]. In a restricted-domain, the number of syntactic patterns is small, and more training instances per feature are available. Under this condition, an ANN has proven to be an excellent tool for modelling. Therefore, the experimentation has included several combinations of features that yield the minimum prosody estimation Root Mean-Square Error (RMSE) using perceptrons. The database used in this paper has been described in [36].

The resulting system predicts prosody with very good results (for duration: 15.5 ms RMSE, for F0: 19.80 Hz in RMSE), that significantly improves our previous rule-based system (28.5 ms RMS duration error) and a general domain database (19 ms RMS duration error).

- Regarding F0, the best features were: a one-of-N coding of the sub-domain (carrier sentence) and the final punctuation mark, a window of 11 syllables for coding stress and the position

of the syllable in the phrase (in relation to the first and last stressed syllable). There was a significant improvement (41.6 %) when compared to just using stress information.

- The best features for modelling duration were: phoneme identity, phoneme stress, a window of 5 phoneme identities around the target phoneme, the number of words in the phrase, the position of the word in the phrase and the position in phrase in relation to first/last stress (the significant improvement in the whole set was 18.71% when compared to using only the first two features).

5.2. Emotional Speech Experiments

Standard neutral synthetic voices are monotonous for a domestic environment, even if the timbre sounds natural. It is necessary to incorporate variety (emotions and attitudes) into synthesised speech to make it more familiar. To study how emotional state affects voice characteristics, a set of experiments have been carried out that range from natural-voice perceptual tests to fully-automatic emotion-identification experiments. Using the database described in [37], that includes recordings from a professional actor simulating several emotional states (neutral, happy, sad, surprised and angry). In a twenty-people identification test, the recognition results range from 74.6% for happiness to 90.3% for anger.

However, this baseline experiment provides no clue on how to simulate each emotion by means of speech synthesis. Therefore, a new experiment was conducted with a set of prosody-modified recordings: neutral recordings modified to have the same prosody as emotional recordings of the same text, and emotional recordings were adapted to have a neutral prosody. Neutral recordings were identified as emotional when the prosody of surprised (76.2%) and sad (66.6%) was applied. Angry (95.2%), happy (52.4%) and sad (45.2%) recordings were correctly identified as emotional when neutral prosody was used. The remaining prosodically modified recordings were not identified as emotional with a score significantly greater than that of a chance level (20.0%). A similar objective automatic-identification-experiment on natural speech has confirmed most of the results of the perceptual test [38].

Finally, in an emotional copy-synthesis identification test that combines segmental and prosodic information, every emotion was significantly identified (above chance level), ranging from 61.9% for happy (the most difficult emotion) to 95.2% for anger (the easiest one).

6. CONCLUSIONS

Speech technology provides new possibilities for disabled people at home. Physically handicapped people can handle appliances through voice or blind people can receive instructions thorough synthetic speech. But the mass-use of speech interfaces by disabled people has not been possible yet for several reasons, the acoustics of the environment, the high variety of users, the difficulties in making a natural dialog, the cost of adapting the dialog and vocabulary to new application domains etc.

This paper has presented several contributions to this task carried out in several modules of an enhanced speech interface for household appliances. As a pilot demonstration, all of these advances have been incorporated and evaluated in a new speech interface for controlling a Hi-Fi audio system increasing significantly its ergonomics and friendliness.

The new speech recognition module allows the identification of the speaker, the emotion of the speaker, and the adaptation of several system modules to the user capabilities. The user's emotion classifier has been incorporated and evaluated; obtaining results that they correlate 81.5% to those results obtained from a perceptual experiment. The acoustic adaptation module for improving the speech recogniser performance has achieved an 18.9% error reduction for spontaneous speech. The natural language processing module proposes a rule-based module with

a bottom-up strategy that increases its flexibility and robustness against speech recognition errors. The concept confidence measure obtained from this module also allows the concept accuracy increase from 76.6% to 80.5%. Two main advances are presented in the dialogue manager: the dialog control based on Bayesian Networks and a set of new tools for helping in the process of developing a new dialog manager from scratch. Proposing Bayesian Networks for dialog modelling allows the dialog process to be dealt automatically (no dialog script is needed). The new speech synthesis module includes advances for increasing the voice naturalness and incorporating emotion into the speech.

Further work needs to be done for this system to be used at home: the first and most important issue is to test it in the field. Although the new modules are based on experience has lead us to improve previous methods, the need to test in a real environment is not eliminated. Other modules need to be improved as well, particularly the adaptation to the acoustics of the room, the possible use of distant microphone and the development of on-line dialog learning methods that could adapt the dialog to new unseen users.

The prospects and benefits of including speech technology in systems to help handicapped people are enormous since speech is the most natural way used to communicate. By continuous research in the area and assistance to people at home using speech interfaces will be a reality in the coming years.

ACKNOWLEDGEMENTS

This work has been supported by the following projects TINA (UPM and DGUI-CAM. ref: R05/10922), ATINA (UPM and DUI-CAM. ref: CCG06-UPM/COM-516), ROBINT (MEC ref: DPI2004-07908-C02) and EDECAN (MEC ref: TIN2005-08660-C04). Authors also want to thank the English style revision by Mark Hallett. The work presented here was carried out while Javier Macías-Guarasa was a member of the Speech Technology Group (Department of Electronic Engineering, ETSIT de Telecomunicación. Universidad Politécnica de Madrid).

REFERENCES

1. R. Bakker “Assistive Technology” *Intensive course in assistive technology/med.engineering.* 5.3.-18.3.2001 Jyväskylä Polytechnic (Finland) and Högskolan Dalarna (Sweden). 2001.
2. X. Huang, A. Acero, and H. Hon “Spoken Language Processing: A Guide to Theory, Algorithm and System Development*.” Prentice Hall PTR*. 2003.
3. F. Fernández, J. Ferreiros, V. Sama, J.M. Montero, R. San-Segundo, J. Macías-Guarasa, and R. García. “Speech Interface for controlling an Hi-Fi Audio System Based on a Bayesian Belief Networks Approach for Dialog Modeling.” *INTERSPEECH 2005*. Lisboa. Portugal. 2005.
4. J. Ferreiros and D.P.W Ellis. “Using Acoustic Condition Clustering to Improve Acoustic Change Detection on Broadcast News,” *ICSLP’00*, 2000.
5. J. Ferreiros. “Detección de puntos de cambio de locutor y reconocimiento del locutor sobre la base de datos Broadcast News basada en el BIC (Bayes Information Criterion*),” I Congreso de la Sociedad Española de Acústica Forense,* 2000.
6. J. Ferreiros, J. Macías-Guarasa, J.M. Montero, R. San-Segundo, L.F. D´Haro, and F.J. Yuste, “Blind Segmentation and Labeling of Speakers via the Bayesian Information Criterion for Video-Conference Indexing,” *III Jornadas en Tecnología del Habla*, 2004.
7. J.L. Gauvain and L. Lamel, “Large-vocabulary continuous speech recognition: advances and applications.” *Proceedings of the IEEE*, Vol: 88, Issue: 8, pp. 1181-1200. Aug. 2000.

8. M. L. Seltzer. "Microphone Array Processing for Robust Speech Recognition." *Ph.D. Thesis. Carnegie Melon University.* 2003.
9. D. Ward and M. Brandstein. "Microphone Arrays: Signal Processing Techniques and Applications." *Springer*, 2001.
10. B. Van Veen and K. Buckley, "Beamforming: A Versatile Approach to Spatial Filtering," *IEEE ASSP Magazine, vol. 5, pp. 4-24,* 1988.
11. M. Wolfel, K. Nickel, and J. McDonough. "Microphone array driven speech recognition: Influence of localization on the word error rate." *MLMI,* 2005.
12. C. Marro, Y. Mahieux, and K. U. Simmer. "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering." *IEEE Transactions on Speech and Audio Processing,* 6(3):240–259, 1998.
13. J. M. Noyes and C. R. Frankish. "Speech recognition technology for individuals with disabilities." *Augmentative & Alternative Communication.* Vol 8, N 4, pp. 297-303. 1992.
14. J.R. Deller, D. Hsu, and L.J. Ferrier. On the use of hidden Markov modelling for recognition of dysarthric speech. *Computer Methods and Programs in Biomedicine,* vol. 35(2):125-39. 1995.
15. P. D. Polur and G.E. Miller. Experiments with fast Fourier transform, linear predictive and cepstral coefficients in dysarthric speech recognition algorithms using hidden Markov model. *IEEE Tran. on Neural Systems and Rehabilitation Engineering.* Vol. 13, n 4, pp 558-561. 2005
16. H. Strik and C. Cucchiarini, C. "Modeling pronunciation variation for ASR: a survey of the literature." *Speech Communication,* vol 29, p. 225-246. 1999.
17. R. Barra, J. Macías-Guarasa, F. Fernández, L.F. D.Haro, J.M. Montero, and J. Ferreiros. "A Proposal of Metrics for Detailed Evaluation in Pronunciation Modeling" *IV Jor. en Tecnología del Habla,* 2006.
18. J.L. Gauvain and C.H. Lee, "Maximum A-Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains*," IEEE Trans. SAP*, Vol. 2, pp. 291-298, 1994.
19. M.J.F. Gales and P.C. Woodland, "Mean and Variance Adaptation Within the MLLR Framework," *Computer Speech & Language*, Vol. 10, pp. 249-264, 1996.
20. R. Córdoba, J. Ferreiros, R. San-Segundo, J. Macías-Guarasa, J.M. Montero, F. Fernández, LF D'Haro, and J.M. Pardo, "Cross-Task and Speaker Adaptation in a Speech Recognition System for Air Traffic Control." *IEEE Aerospace and Electronics Systems Magazine*. Vol. 21-9pp. 12-17. 2006
21. J. Ferreiros, R. San-Segundo, F. Fernández, LF. D'Haro, V. Sama, R. Barra, and P. Mellén. "New Word-Level and Sentence-Level Confidence Scoring Using Graph Theory Calculus and its Evaluation on Speech Understanding." *INTERSPEECH 2005.* Lisboa. Portugal. 2005.
22. H.M. Meng, C.Wai, and R.Pieraccini. "The Use of Belief Networks for Mixed-Initiative Dialog Modeling," *IEEE Tran. on Speech and Audio Processing,* Vol.11, No.6, pp. 757-773, 2003.
23. C. Huang and A. Darwiche, "Inference in belief networks: a procedural guide." *International Journal of Approximate Reasoning* Vol 11 pp 1-158. USA. 1994.
24. H.M. Meng, W. Lam, and K.F. Low. "Learning Belief Networks for Language Understanding." *ASRU* 1999.
25. M. McTear, "Software to Support Research and Development of Spoken Dialogue Systems" *Eurospeech'99,* pp. 339-342. 1999.
26. W. Wahlster, N. Reithinger, and A. Blocher "SmartKom: Multimodal communication with a life-like character." *Eurospeech,* pp. 1547–1550. 2001
27. J. Glass and E. Weinstein "SPEECHBUILDER: Facilitating Spoken Dialogue System Development" *Eurospeech,* pp. 1335-1339. 2001.

28. L. F. D'Haro, R. Córdoba, J. Ferreiros, S.W. Hamerich, V. Schless, B. Kladis, V. Schubert, O. Kocsis, S. Igel, and J.M. Pardo. "An advanced platform to speed up the design of multilingual dialog applications for multiple modalities." *Speech Communication* Vol. 48, No 8, pp.863-887. 2006.
29. Voice Extensible Markup Language. http://www.w3.org/TR/voicexml21/. 2006.
30. M. Johnston, L.F. D'Haro, M. Levine, and B. Renger. "A Multimodal Interface for Access to Content in the Home." Proc. 45th Annual Meeting of the ACL. pp. 376-383, 2007.
31. XHTML+Voice Profile. http://www.w3.org/TR/xhtml+voice/. 2006.
32. R. San-Segundo, R. Barra, L. F. D'Haro, J.M. Montero, R. Córdoba, and J. Ferreiros "A spanish speech to sign language translation system for assisting deaf-mute people" *Interspeech,* Pittsburgh, USA. 2006.
33. R. Cordoba, R. San-Segundo, J. Macias, Juan M. Montero, R. Barra, L. F. D'Haro, J. C. Plaza, and J. Ferreiros. "Integration of Acoustic Information and PPRLM Scores in a Multiple-Gaussian Classifier for Language Identification." *IEEE Odyssey 2006:* 2006.
34. R. Córdoba, J.M. Montero, J. Gutiérrez-Arriola, J.A. Vallejo, E. Enríquez, and J.M. Pardo. "Selection of the most significant parameters for duration modeling in a Spanish text-to-speech system using neural networks." *Computer Speech & Language,* Vol 16 Nº 2, pp. 183-203. 2002.
35. R. San Segundo, J.M. Montero, R. Córdoba, and J.M. Gutiérrez-Arriola, "Stress Assignment in Spanish Proper Names." *Intern. Conference on Spoken Language Processing*, pp. 346-349. 2000.
36. J.M. Montero, R. Córdoba, J.A. Vallejo, J. Gutiérrez-Arriola, E. Enríquez, and J.M. Pardo, Restricted-Domain Female-Voice Synthesis in Spanish: from Database Design to ANN Prosodic Modeling. *Proceedings of ICSLP,- pp. 621-624.* 2000.
37. J. M. Montero, J. Gutiérrez-Arriola, R. de Córdoba, E. Enríquez, and J. M. Pardo "The role of pitch and tempo in Spanish emotional speech: towards concatenative synthesis" in "Improvements in speech synthesis" *John Wiley & Sons, Ltd., pp. 246-251.* 2002.
38. R. Barra, J.M. Montero, J. Macias-Guarasa, L.F. D'Haro, R. San-Segundo, and R. Cordoba. "Prosodic And Segmental Rubrics In Emotion Identification." *ICASSP'2006* pp 1085-1088. 2006.

ABOUT THE AUTHORS

R. San-Segundo received his MSEE (1997) and Ph.D. (2002) degrees from Universidad Politécnica de Madrid (UPM), with highest distinctions. During 1999 and 2000, Ruben did two summer stays at The Center of Spoken Language Research (CSLR), University of Colorado (Boulder). From Sep. 2001 through Feb. 2003, Rubén worked at the Speech Technology Group of Telefónica I+D.

R. de Cordoba Herralde received his MSEE (1991) and Ph.D. (1995) degrees from Universidad Politécnica de Madrid (UPM) with highest distinctions. He has been a member of the Speech Technology Group since 1990, teaching in the UPM since 1993, now working as Associate Professor in the Department of Electronic Engineering. He worked as Research Associate in Cambridge Univ. (UK), Speech, Vision and Robotics Group, in 2001.

J. Ferreiros López received his MSEE (1990) and Ph.D. (1996) degrees from Universidad Politécnica de Madrid (UPM) with highest distinctions. Since 1988 Javier has been member of the Speech Technology Group at UPM, where he holds an associate professor position and is currently the associate director of the Department of Electronic Engineering. From Oct 1999 to Apr 2000, Javier stayed at ICSI, Berkeley, CA as visiting researcher. His research interests focus on spoken dialog systems.

J. Macias-Guarasa received his MSEE degree (1992) and Ph.D. (2001) degrees from Universidad Politécnica de Madrid (UPM), with highest distinction. From 1990 to 2007 he was a member of the Speech Technology Group and associate professor at UPM and is currently associate professor in the Department of Electronics of the University of Alcalá. He spent six months in the Speech Group of the ICSI in Berkeley, California.

J. M. Montero Martínez. Received his MSEE (1992) and Ph.D. (2003) degrees from Universidad Politécnica de Madrid (UPM) with highest distinctions. He spent seven months in the Speech Group of the ICSI in Berkeley, California. Currently, Juan M. is associate professor in the Department of Electronic Engineering at UPM and member of the Speech Technology Group since 1990.

F. Fernández Martínez received his MSEE degree from Universidad Politécnica de Madrid (UPM) in 2002 (with highest distinction) and he is currently assistant professor and Ph.D. candidate at UPM. During 2006, Fernando had a summer stay at The IDIAP Research Institute affiliated with the "Ecole Polytechnique Fédérale de Lausanne" (EPFL) and the University of Geneva (Switzerland).

L. F. D'Haro Enríquez received his degree as Electronics Engineer in 2000, from Universidad Autónoma de Occidente in Cali, Colombia. He is currently assistant professor and Ph.D. student at UPM, Spain. In 2005 he stayed at Computer Science VI, RWTH Aachen University (Germany) working in machine translation and language modeling, and in 2006 at AT&T labs research in Florham Park, NJ (USA), working in multimodal dialogue interaction and interfaces.

R. Barra Chicote received his MSEE degree from Technical University of Madrid in 2005 (with highest distinction). Since 2003 he is a member of the Speech Technology Group. In 2006 he was a visitor researcher of the Center for Spoken Language Research (CSLR) at Colorado University. In 2008 he was a visitor researcher of the Center for Speech Technology Research CSTR) at Edinburgh University. His main research interests are related to emotional speech synthesis and automatic emotion identification.

J. M. Pardo Muñoz got his MSEE Degree (1978) and Ph.D. (1981) from Universidad Politécnica de Madrid. He got a best graduate national award (1980) and a best Ph.D. Thesis national award (1982). He has been Head of the Speech Technology Group since 1987 and Full Prof. since 1992 being head of the Electronic Engineering Dept. from 1995-2004. He has been a Fulbright Scholar at MIT, a visiting scientist at SRI International and a visiting fellow at the ICSI. He was chairman of EUROSPEECH 1995, member of the ISCA Advisory Council, ELSNET Executive Board, and NATO RSG 10 & IST 3.