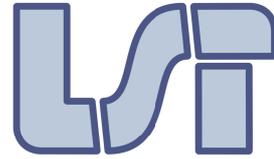




ugr | Universidad
de Granada



On the development of Adaptive and Portable Spoken Dialogue Systems: Emotion Recognition, Language Adaptation and Field Evaluation

Zoraida Callejas Carrión

Dissertation submitted for the degree of
Doctor of Philosophy in Computer Science
with European Doctorate Mention

Supervised by:
Dr. Mr. Ramón López-Cózar Delgado

*Departamento de Lenguajes y Sistemas Informáticos
Universidad de Granada*

Granada, April 2008

Acknowledgements

Some years ago I was introduced to the area of dialogue systems, and I was lucky it was by somebody who enjoys working with them. Since then, in his enthusiasm I have had a splendid referent. If the role of a supervisor is to guide and advice, it can not be in any other way rather than encouraging to develop the own ideas freely. Thank you Ramón.

And to work freely it is necessary to go out of our laboratory, stop staring at our screen and find out what the others do, so that we can enrich our own contributions. Pasteur once said that science does not have any nationality, but that we scientists do. I would like to thank Jan Nouza and all the members of the Laboratoř počítačového zpracování řeči for demonstrating that this is not an obstacle, and that the willing to work towards shared objectives is the only important thing to create a team. Thank you for teaching me so many things, professionally and personally, and for inviting and receiving me in that spring I will never forget.

Also many thanks to Michael McTear. I remember when we discussed my first research results when I began my Ph.D. studies. I started the venture of carrying out my Doctoral Thesis in a large extent because of his encouragement and interest. Thank you for offering me your valuable opinion and advice, which have made me improve, learn and progress.

I have also been lucky to share my day by day with wonderful people. I am very grateful to the colleagues in the department who have cheered me up, encouraged me in my work and congratulated me for my achievements. You they have brighten up the way. Specially, I would like to thank my friends from Labs 1 and 2, including those who are no longer here, with whom I have lived the best moments and drank a lot of “coffee”.

There are also some other coffees which, although sometimes “virtual”, have rescued me from drowning. Those from my last resorts, from my friends who despite the distance have always demonstrated their affection. Thank you Bea, M^aPaz and Vane, without you, who understand me like nobody else, nothing would have been the same. In the distance my gratitude is also for Hynek, my guide in Hrabal’s country, with your great sense of humour you have reduced the highest pitfalls: I will always remember your saviour bramborák.

My greatest thanks go to my conviction and security in the uncertainty moments: my family. To my parents and my sister Clara for helping and believing in me constantly and unconditionally. Thank you for bearing the “side effects” of the thesis, the deadlines, the exhausting working days... This work is also yours. Thanks also to Carmen for sweetening my English classes since I was a child.

Finally, thank you David for your infinite patience, care and affection, for reminding me that it was worth and giving me wings to continue. Thank you for smoothing the way and holding my hand along it. Although you are the last I mention, *in my life ...*

Summary

This Thesis presents the work done on three of the most challenging topics in the area of spoken dialogue systems: recognition of non-acted emotions, cross-lingual adaptation of speech recognizers and field (not laboratory restricted) evaluation with both “objective” and “subjective” criteria. The research described constitutes a novel contribution to what the experts have established as the major research trends in the area of SDS: adaptiveness and portability of the systems.

Firstly, regarding emotion recognition, a detailed study is supplied on how to calculate and interpret reliability coefficients for the annotation of corpora of real emotions. A new efficient approach is proposed that considerably enhances inter-annotator agreement and machine emotion recognition by the use of several context information sources. On the one hand, human annotation is facilitated by achieving inter-annotator agreement values closer to the maximum attainable even with non-expert annotators. On the other hand, a machine-learned emotion recognition method is proposed which automatically extracts the contextual information at run time, obtaining results which provide a 40% improvement on the state-of-the-art approaches.

Secondly, the research on cross-lingual adaptation of speech recognizers was carried out during a three-month stay at the Technical University of Liberec (Czech Republic). An approach is presented to cost-efficiently (in terms of time and effort) adapt a speech recognizer to work in another language. The proposal has been used to adapt a Czech speech recognizer to a language which is acoustically very similar (Slovak) and another with a completely different origin (Spanish). It obtained a recognition accuracy around 70% for Spanish and 80% for Slovak in tasks demanding rich vocabularies (around 150,000 words).

Thirdly, several statistical studies were carried out on a field evaluation of a spoken dialogue system. New empirical evidence is provided on the relationships between evaluation criteria. The study includes both interaction parameters and quality judgments, paying special attention to user satisfaction and task success, studying the impact of the dialogue management initiatives employed, as well as the users expertise and collaboration during the interaction with the system.

All the methods proposed in the Thesis have been tested with real systems, for which the UAH spoken dialogue system was developed. The system has been available to the public on the phone from June 2005, since when all the interactions have been recorded. A year of user calls was semi-automatically annotated with de-facto standard evaluation criteria (e.g. word-error-rate). This corpus was extended with annotation of emotions by nine non-expert annotators, who tagged each utterance as “neutral”, “angry”, “doubtful” or “bored”. Both the emotion recognition methods and the evaluation studies contained in this Thesis were tested with the UAH corpus. The proposed method for cross-lingual adaptation was evaluated using the MyVoice system, which was developed in the Technical University of Liberec. The translation of its commands for interaction in Spanish is another of the contributions of the Thesis.

The empirical results obtained with the dialogue systems have been rigorously tested and their significance calculated using different statistical significance studies. The results of the research described have been published in several prestigious national and international conferences and journals. They have also formed part of oral presentations, posters and demonstrations internationally.

Contents

1. Introduction	17
1.1. Spoken dialogue systems	17
1.1.1. Speech recognition	17
1.1.2. Natural language processing	19
1.1.3. Dialogue management	20
1.1.4. Natural language generation	22
1.1.5. Text to speech synthesis	23
1.2. Applications of SDSs	24
1.3. Evolution of SDSs	28
1.4. Objectives of the Thesis	39
1.5. Structure of the Thesis	40
2. The UAH spoken dialogue system	43
2.1. Introduction	43
2.2. Modular architecture	44
2.2.1. Automatic speech recognition	46
2.2.2. Dialogue management	46
2.2.3. Database access	48
2.2.4. Oral response generation	49
2.3. Automatic grammar generation	50
2.4. The UAH speech corpus	57
2.5. Conclusions	57
3. Recognition of non-acted emotions	59
3.1. Introduction	59
3.2. Related work	61
3.3. Human annotation of the UAH corpus	71
3.3.1. Calculation of the agreement between annotators	72
3.3.2. Discussion of human annotation results	78
3.4. Automatic classification of the UAH corpus	89
3.4.1. Automatic classification based on standard acoustic features	91

3.4.2.	Automatic classification based on normalized acoustic features	93
3.4.3.	Automatic classification based on dialogue context	97
3.4.4.	Automatic classification based on normalized acoustic features and dialogue context (two-step method)	102
3.4.5.	Previous version of the two-step method	105
3.5.	Conclusions	108
4.	Cross-lingual adaptation of speech recognizers	111
4.1.	Introduction	111
4.2.	Related work	113
4.3.	The MyVoice system and Czech speech recognizer	116
4.4.	Cross-lingual adaptation	118
4.4.1.	Phoneme mapping between Czech and Slovak	119
4.4.2.	Phoneme mapping between Czech and Spanish	120
4.4.3.	Speaker adaptation	122
4.5.	Experimental set-up	123
4.6.	Experimental results	126
4.6.1.	Interaction with MyVoice	126
4.6.2.	Impact of speaker adaptation	127
4.6.3.	Effect of the size of the recognition dictionary	128
4.7.	Conclusions	130
5.	Field evaluation of spoken dialogue systems	133
5.1.	Introduction	133
5.2.	Related work	135
5.3.	Evaluation criteria	142
5.3.1.	Interaction parameters	142
5.3.2.	Quality judgments	145
5.4.	Statistical studies employed for evaluation	149
5.5.	Evaluation results	152
5.5.1.	Impact of the interaction performance on the user decision to answer the subjective test	156
5.5.2.	Criteria with highest impact on user satisfaction and task success	158
5.5.3.	Criteria with highest number of significant relations	162
5.5.4.	Impact of user's knowledge and experience	163

5.5.5. Impact of dialogue management initiative	164
5.6. Conclusions	168
6. Conclusions and future work	171
6.1. Summary of contributions	171
6.2. Future work	175
6.2.1. Recognition of non-acted emotions	175
6.2.2. Cross-lingual adaptation of speech recognizers	175
6.2.3. Field evaluation of spoken dialogue systems	176
 A. Publications	 177

List of Tables

3.1.	Summary of Spanish emotional speech corpora	70
3.2.	Distance between the emotions considered	76
3.3.	Values of the Kappa coefficients for unordered and ordered annotation schemes	78
3.4.	Kappa values for the different annotator types	83
3.5.	Observed agreement values	88
3.6.	Acoustic features used for classification	91
4.1.	Mapping of the Slovak phonemes that do not exist in Czech to the closest Czech ones	120
4.2.	Mapping of the Spanish phonemes to the closest Czech ones .	121
4.3.	WER [in %] for the command-and-control task	126
4.4.	Effect of dictionary size on WER [in %] taking into account OOV words and speaker adaptation	128
4.5.	Relative WER reduction [in %] yielded by the adaptation to speakers	130
5.1.	Interaction parameters employed	143
5.2.	Perceived quality and user profile parameters employed	147
5.3.	Descriptive statistics of the criteria used	149
5.4.	Type of variables used for the statistical studies	151
5.5.	Significant partial correlations	153
5.6.	Correlations between the criteria used	154
5.7.	Significance variations between the <i>Pearson</i> , <i>Chramer's Tau-b</i> and <i>Spearman's Rho</i> coefficients	155
5.8.	Significance of the relationship between "The user taking the subjective test" and the interaction parameters	156
5.9.	ANOVA table for "task success" and the rest of the interaction parameters regarding the different user groups	157
5.10.	Statistical significance of the most important relationships with "user satisfaction"	159
5.11.	Criteria that were significantly correlated with one initiative type but not with the other	165

List of Figures

1.1.	Modular architecture of spoken dialogue systems	18
1.2.	Evolution of the envisioned research lines in the area of spoken dialogue systems	34
2.1.	Modular architecture of the UAH system	44
2.2.	Latency of the dynamic creation of grammar rules	52
2.3.	Automatic grammar rules update with the GAG tool	54
2.4.	TGC technique vs. dynamic creation of grammar rules measured in terms of user satisfaction	55
2.5.	Perceived interaction speed (left) and user satisfaction (right) using UAH	56
3.1.	Naturalness vs. control in the main emotional corpora generation approaches	69
3.2.	Kappa coefficients used in the experiments	73
3.3.	Proportion of non-neutral annotated utterances	79
3.4.	Percentage of utterances in which annotators agree	80
3.5.	Pair wise disagreement between annotators with the ordered scheme	81
3.6.	Proportion of annotated emotions depending on dialect	82
3.7.	Relative values of agreement by chance and observed agreement for multi- κ	83
3.8.	Kappa <i>maximum</i> , <i>minimum</i> , <i>normal</i> (italics) and observed (bold) values	87
3.9.	Recognition accuracy for <i>angry</i> , <i>bored</i> and <i>doubtful</i> considering non-normalized vs. normalized acoustic features, and no feature selection vs. feature selection	95
3.10.	Recognition accuracy for <i>angry</i> and <i>doubtful</i> <i>OR</i> <i>bored</i> considering non-normalized vs. normalized acoustic features, and no feature selection vs. feature selection	96
3.11.	Example of transitions between system prompts in the UAH system	100

3.12. Emotion recognition accuracy using both acoustic and dialogue context information	103
3.13. Comparison of the recognition accuracies of the methods for automatic emotion recognition	104
3.14. First version of the two-step method for automatic emotion recognition	105
3.15. Impact of the value of dialogue context thresholds in emotion classification success	106
3.16. Comparison of the accuracy of the different methods proposed for automatic emotion recognition	107
3.17. Final version of the two-step method for automatic emotion recognition	108
4.1. Language family groups for Czech, Slovak and Spanish	113
4.2. Scheme of the cross-language adaptation procedure	119
4.3. Outline of the experimental set-up	124
4.4. Effect of the adaptation technique on the performance of the adapted recognizers	128
4.5. Performance of the Slovak (above) and Spanish (below) adapted recognizers	129
5.1. PARADISE architecture model	140
5.2. Example of the computation of the interaction parameters for an UAH dialogue	144
5.3. Users' knowledge about new technologies for information access (1 = Low, 5 = High)	146
5.4. Demographic data for the different user types	148
5.5. Percentage of successful dialogues which are also complete regarding the different user groups	160
5.6. Task success vs. Perceived ease of error correction	162
5.7. Dialogue initiative influence on user confidence	167
5.8. Dialogue duration for each dialogue management initiative	168

We must, with a view to the science which we are seeking, first recount the subjects that should be first discussed (...) for the subsequent free play of thought implies the solution of the previous difficulties, and it is not possible to untie a knot of which one does not know.

Aristotle, *Metaphysics* (Book III)

1

Introduction

Due to their constant improvement in performance and decreasing cost, computers have become an important part of our daily lives. We are surrounded by numerous electronic devices which provide information and functions which we are increasingly interested in accessing any time, anywhere and in our native language. Thus, new interfaces are needed to provide natural, intuitive and efficient ways of communication between humans and computers. Spoken language dialogue systems have emerged as a practical method for providing computers with intelligent communicative capabilities, as speech is the most natural and flexible means of communication among humans.

1.1 Spoken dialogue systems

A spoken dialogue system (SDS) is a software that accepts natural language as an input and produces natural language as an output engaging in a conversation with the user. To successfully manage the interaction with users, spoken dialogue systems usually carry out five main tasks: automatic speech recognition (ASR), natural language understanding (NLU), dialogue management (DM), natural language generation (NLG) and text-to-speech synthesis (TTS). These tasks are usually implemented in different modules. Figure 1.1 shows the typical modular architecture of an SDS.

1.1.1. Speech recognition

Speech recognition is the process of obtaining the text string corresponding to an acoustic input. It is a highly complex task, as there is a great deal of

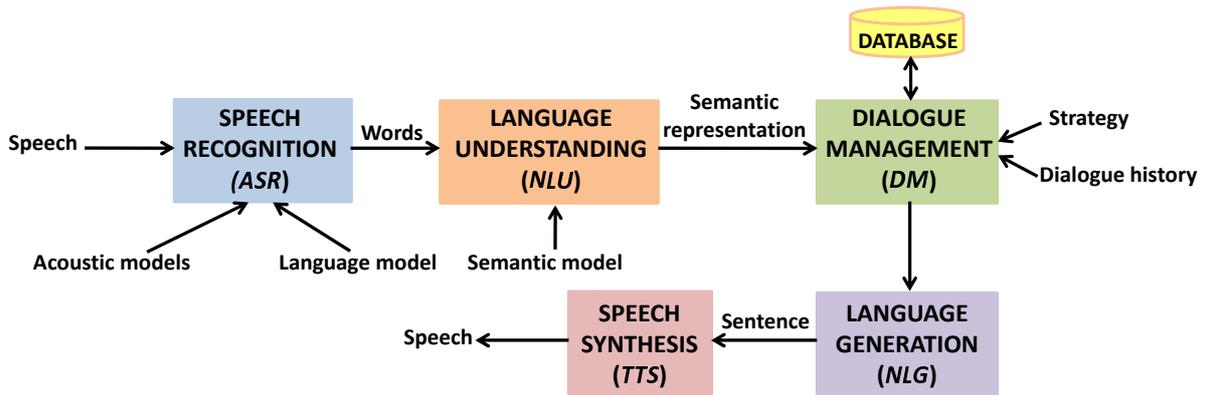


Figure 1.1. Modular architecture of spoken dialogue systems

variation in input characteristics, which can differ according to the linguistics of the utterance, the speaker, the interaction context and the transmission channel. Linguistic variability involves differences in phonetic, syntactic and semantic components that affect the voice signal. Inter-speaker variability refers to the wide differences between speakers regarding speaking style, voice, age, sex or nationality. Furthermore, even the same person does not always pronounce the same words in the same way, as people are affected by physical and psychological determinant factors that are highly variable and usually not predictable. This phenomenon is known as intra-speaker variability. Additionally, differences in the communication channels and/or devices also affect voice signals, due to effects derived from the transmission such as reverberation. Finally, the interaction environment variability is also very important, as the recognizer must be robust to differences in the background noise.

Different applications demand different complexity of the speech recognizer. Cole et al. (1997) identify eight parameters that allow optimal tailoring of the speech recognizer: speech mode, speech style, dependency, vocabulary, language model, perplexity, SNR and transducer. Regarding speech mode, speech recognizers can be classified into isolated-word or continuous-speech recognizers. The former recognize words separated by pauses, while the latter are able to recognize a natural discourse in which the speaker uses his normal speaking rate. Regarding speech style, a discourse can be read or spontaneous; the latter has peculiarities, such as hesitations and repetitions that make it more complex to recognize. Speech recognition can also be speaker-dependent or independent. In the first case the acoustic models

are trained with the voice of only one speaker, for whom optimal success rates are reached; whereas the second is prepared to recognize a wide range of speakers yielding acceptable success rates. Another important parameter is the accepted vocabulary, i.e. the number of words that the recognizer can distinguish. Applications with more than 5,000 accepted words are normally considered as large-vocabulary applications (Jurafsky and Martin, 2000). Finally, to deal with noise and channel variability, speech recognizers employ a noisy channel model, i.e. they treat the recognition problem as if it were necessary to recover a message which has been corrupted after going through a noisy channel. To do so, they use stochastic models which consider the possible original messages and calculate which of them is more likely to be the correct one.

1.1.2. Natural language processing

Once the SDS has recognized what the user uttered, it is necessary to understand what he said. Natural language processing is a method of obtaining the semantics of a text string and generally involves morphological, lexical, syntactical, semantic, discourse and pragmatical knowledge. In the first stage, lexical and morphological knowledge divide the words into their constituents by distinguishing between lexemes and morphemes: lexemes are parts of words that indicate their semantics and morphemes are the different infixes and suffixes that provide different word classes (e.g., *establishment* = *establish* – *ment*). Syntactic analysis yields the hierarchical structure of the sentences. However, in spoken language, phrases are frequently affected by difficulties associated with the so-called disfluency phenomena: filled pauses, repetitions, syntactic incompleteness and repairs (Gibbon et al., 2000). Semantic analysis extracts the meaning of a complex syntactic structure from the meaning of its constituent parts. In the pragmatic and discourse-processing stage, the sentences are interpreted in the context of the whole dialogue, the main complexity of this stage is the resolution of anaphora, and ambiguities derived from phenomena such as irony, sarcasm or double entendre.

There are currently two major approaches to tackling the problem of understanding: rule-based approaches and statistical models learned from data corpus. Rule-based approaches extract semantic information based on a syntactic-semantic analysis of the sentences, using grammars defined for the task, or by means of the detection of keywords with semantic meanings. In the case of statistical methods, the process is based on the definition of language units with semantic content and the learning of models from labelled samples. This kind of analysis uses a probabilistic model to identify concepts, markers and values of the cases, and uses them to represent the relationship between markers of cases and their values to semantically decode the user utterances (Minker, 1998).

1.1.3. Dialogue management

There is no universally agreed-upon definition of the tasks that a dialogue manager has to carry out. Traum and Larsson (2003) state that dialogue management involves four main tasks: i) updating the dialogue context, ii) providing a context for interpretations, iii) coordinating other modules and iv) deciding the information to convey and when to do it. Thus, the dialogue manager has to deal with different sources of information such as the NLU results, database queries results, application domain knowledge, knowledge about the users and the previous dialogue history. Its complexity depends on the task and the dialogue flexibility and initiative. Bernsen et al. (1994) provide a taxonomy which shows that for small and simple tasks single-word dialogue can be convenient with either system or user initiative and limited system feedback. However, for large, well-structured tasks, there is a need for system-directed dialogues with appropriate system feedback, tracking of the dialogue history and simple user models. For larger ill-structured tasks, mixed initiative dialogues are necessary, with dynamic predictions, linguistic and dialogue act, dialogue history and advanced user modelling.

The simplest dialogue management strategy is to model the dialogue as a finite-state machine in which the transitions between the system responses are determined by the user's actions. The users actions are his responses to the system, which are coded in recognition grammars. A significant extension consists of frame-based approaches, which have been developed to overcome the lack of flexibility of dialogue grammars. This is the

approach used by most current commercial systems. Unlike the finite-state approach, frame-based dialogue managers do not have a predefined dialogue path but use a frame structure comprised of one slot per piece of information that the system can gather from the user. In this approach, the system interprets the speech in order to acquire enough information to perform a specific action. Its advantage is that it can capture several data at once and the information can be provided in any order (more than one slot can be filled per dialogue turn and in any order).

For more complex domains, plan-based dialogue management can be used. Its core idea is that humans communicate to achieve goals and during the interaction the mental state of the speakers may change. Thus, plan-based dialogue managers model dialogue as a cooperation between the user and the system to reach common goals, so that each utterance is not considered as a text string, but as a dialogue act in which the user communicates his intentions. From each user utterance the system refines a user model in which it tries to predict his intentions and objectives. This is done recursively to produce the system response to the user utterances until the task is accomplished.

This last approach is related to the so-called “information state” dialogue theory. The information state of a dialogue represents the information needed to uniquely distinguish it from all others. It comprises the accumulated user interventions and previous dialogue actions on which the next system response can be based. The information state is also sometimes known as the conversation store, discourse context or mental state. Following the information state theory, the main tasks of the dialogue manager are to update the information state based on the observed user actions, and based on them, to select the next system action. The Trindi project (TRINDI Consortium, 2001), proposed an architecture and toolkit for building dialogue managers based on an information state approach.

Additionally, when it is necessary to execute and monitor operations in a dynamically changing application domain, an agent-based approach can be employed. The modular agent-based approach to dialogue management makes it possible to combine the benefits of different dialogue control models, such as finite-state based dialogue control and frame-based dialogue management (Chu et al., 2005). Similarly, it can benefit from alternative dialogue management strategies, such as the system-initiative approach and the

mixed-initiative approach (Walker et al., 1998a). It also makes it possible to combine rule-based and machine learning approaches (Turunen et al., 2004).

The application of machine-learning approaches to dialogue management strategy design is a rapidly growing research area. Machine-learning approaches to dialogue management attempt to learn optimal strategies from corpora of real human-computer dialogue data using automated “trial-and-error” methods instead of relying on empirical design principles (Griol, 2007). The Markov-Decision-Process (MDP) model serves in most of these approaches as a formal representation of human-machine dialogue and provides the basis for formulating strategy learning problems (Williams and Young, 2007; Cuayáhuitl et al., 2006; Lemon et al., 2006).

1.1.4. Natural language generation

Natural language generation is the process of obtaining texts in natural language from a non-linguistic representation. It is usually carried out in five steps: content organization, content distribution in sentences, lexicalization, generation of referential expressions and linguistic realization. It is important to obtain legible messages, optimizing the text using referring expressions and linking words and adapting the vocabulary and the complexity of the syntactic structures to the user’s linguistic expertise.

The simplest approach consists of using predefined text messages (e.g. error messages and warnings). Although intuitive, this approach completely lacks from flexibility. The next level of sophistication is template-based generation, in which the same message structure is produced with slight alterations. The template approach is used mainly for multi-sentence generation, particularly in applications whose texts are fairly regular in structure, such as business reports.

Phrase-based systems employ what can be considered as generalized templates at the sentence level (in which case the phrases resemble phrase structure grammar rules), or at the discourse level (in which case they are often called text plans). In such systems, a pattern is first selected to match the top level of the input, and then each part of the pattern is expanded into a more specific one that matches some portion of the input. The cascading process stops when every pattern has been replaced by one or more words.

Finally, feature-based systems represent the maximum level of generalization and flexibility. In feature-based systems, each possible minimal alternative of expression is represented by a single feature; for example, whether the sentence is either positive or negative, if it is a question or an imperative or a statement, or its tense. To arrange the features it is necessary to employ linguistic knowledge. Another alternative is to use corpus-based natural language generation (Oh and Rudnicky, 2000), which stochastically generates system utterances.

1.1.5. Text to speech synthesis

Text-to-speech synthesizers transform a text into an acoustic signal. A text-to-speech system is composed of two parts: a front-end and a back-end. The front-end carries out two major tasks. Firstly, it converts raw text containing symbols such as numbers and abbreviations into their equivalent words. This process is often called text normalization, pre-processing, or tokenization. Secondly, it assigns a phonetic transcription to each word, and divides and marks the text into prosodic units, i.e. phrases, clauses, and sentences. The process of assigning phonetic transcriptions to words is called text-to-phoneme or grapheme-to-phoneme conversion. The output of the front-end is the symbolic representation constituted by the phonetic transcriptions and prosody information.

The back-end (often referred to as the synthesizer) converts the symbolic linguistic representation into sound. On the one hand, speech synthesis can be based on human speech production. This is the case of parametric synthesis which simulates the physiological parameters of the vocal tract, and formant-based synthesis, which models the vibration of vocal chords. In this technique, parameters such as fundamental frequency, voicing, and noise levels are varied over time to create a waveform of artificial speech. Another approach based on physiological models is articulatory synthesis, which refers to computational techniques for synthesizing speech based on models of the human vocal tract and the articulation processes.

On the other hand, concatenative synthesis employs pre-recorded units of human voice. Concatenative synthesis is based on stringing together segments of recorded speech. It generally produces the most natural-sounding synthesized speech; however, differences between natural variations in speech

and the nature of the automated techniques for segmenting the waveforms sometimes result in audible glitches in the output. The quality of the synthesized speech depends on the size of the synthesis unit employed. Unit selection synthesis uses large databases of recorded speech. During database creation, each recorded utterance is segmented into some or all of the following: individual phones, syllables, morphemes, words, phrases, and sentences. Unit selection provides the greatest naturalness, because it applies only a small amount of digital signal processing to the recorded speech. There is a balance between intelligibility and naturalness of the voice output or the automatization of the synthesis procedure. For example, synthesis based on whole words is more intelligible than the phone-based but for each new word it is necessary to obtain a new recording, whereas the phones allow building any new word. In one extreme, domain-specific synthesis concatenates pre-recorded words and phrases to create complete utterances. It is used in applications in which the variety of texts the system will produce is limited to a particular domain, like transit schedule announcements or weather reports. At the other extreme, diphone synthesis uses a minimal speech database containing all the diphones (sound-to-sound transitions) occurring in a language. The number of diphones depends on the phonotactics of the language: for example, Spanish has about 800 diphones and German about 2,500. In diphone synthesis, only one example of each diphone is contained in the speech database.

Finally, HMM-based synthesis is a method in which the frequency spectrum (vocal tract), fundamental frequency (vocal source), and duration (prosody) of speech are modelled simultaneously by HMMs. Speech waveforms are generated from HMMs themselves, based on the maximum likelihood criterion.

1.2 Applications of SDSs

The complexity of the interaction between the user and the dialogue system can vary and some of the previously described components might not be used. For example, for a simple menu, semantic analysis is not necessary. However, for a conversational companion all the modules must be used in order to interpret the user input, take justified decisions on what the system will respond, and finally tailor the answer to user needs and expectations.

There is a large variety of applications in which spoken dialogue systems can be used. One of the most wide-spread is providing information on a specific topic. Some sample applications are:

- DARPA Communicator - Intelligent conversational interfaces to distributed flight and booking information (DARPA, 1992, 1994).
- Voyager - Tourist and travel information for the Greater Boston area (Glass et al., 1995).
- ARISE - Automatic Railway Information Systems for Europe in several languages (den Os et al., 1999).
- AUGUST - Swedish spoken dialogue system using an animated agent to provide information about Stockholm (Gustafson et al., 1999).
- Adapt - Multimodal spoken dialogue system for browsing apartments on the Stockholm real estate market (Gustafson et al., 2000).
- Jupiter - Weather forecast over the phone (Zue et al., 2000).
- Mercury - Flight reservation system and weather forecast (Seneff and Polifroni, 2000).
- CTT-Bank - Speech-controlled telephone banking system (Melin et al., 2001).
- SmartKom - Multimodal dialogue system with several application domains such as cinema booking (Alexandersson and Becker, 2001).
- Let's Go - A spoken dialogue system for the non-native and elderly in the domain of bus information around Pittsburgh (Raux et al., 2005).
- Amities project - Automated Multilingual Interaction with Banking Information and Services (Hardy et al., 2006).
- DisCoH - Spoken Dialogue System for Conference Help (Andeani et al., 2006).
- HIGGINS - Pedestrian city navigation and guidance (Skantze et al., 2006).

- TALK TownInfo - Multimodal dialogue system using reinforcement learning for tourist information scenarios (Lemon et al., 2006).
- Conquest - Spoken dialogue system that provides schedule information during conferences (Bohus et al., 2007).

Spoken dialogue systems have also been used for education and training, particularly in improving phonetic and linguistic skills:

- LARRI - Multimodal spoken dialogue system which provides assistance and guidance to F18 aircraft personnel during maintenance tasks (Bohus and Rudnicky, 2002).
- ITSPOKE - Tutoring spoken dialogue system which engages the students in a spoken dialogue to provide feedback and correct misconceptions (Litman and Silliman, 2004).
- Radiobot-CFF - Spoken dialogue system which can engage in Call For Fire (CFF) radio dialogues to help train soldiers in proper procedures for requesting artillery fire missions (Roque et al., 2006b).
- VOCALIZA - Dialogue application for computer-aided speech therapy in the Spanish language, which helps in the daily work of speech therapists who teach linguistic skills to Spanish speakers with different language pathologies (Vaquero et al., 2006).
- LISTEN - “Literacy Innovation that Speech Technology ENables” is an automated Reading Tutor that displays stories on a computer screen, and listens to children read aloud (Mostow, 2008).

In some cases, spoken interaction can be the only way to access information, as, for example when the screen is too small to display information (e.g. hand-held devices) or when the eyes of the user are busy in other tasks (e.g. driving):

- MUST - Multimodal, multilingual information services for small mobile Terminals (Boves and Os, 2002).
- VICO - Virtual Intelligence CO-driver enabling natural interaction between humans and digital devices and services in the car (Mattasoni et al., 2002).

- Athosmail - A multilingual adaptive spoken dialogue system for e mail message reading on a mobile phone (Jokinen et al., 2004).
- CHAT - Conversational helper for automotive tasks such as query song databases and operate an MP3 player (Weng et al., 2006).
- DICO - A multimodal dialogue system that lets the driver control devices and access Internet services using natural speech (Villing and Larsson, 2006).

Spoken interaction is also useful for remote control of devices and robots, specially in smart environments:

- WITAS - Dialogue interface for multimodal conversations with WITAS, a robotic helicopter (Lemon et al., 2001).
- ODISEA - A spoken dialogue system that allows interaction between users and intelligent environments. The dialogue components are automatically created and permit context-based spoken interaction between the environment and users (Montoro et al., 2004, 2006).
- SENECA - Speech-based user interface in a wide range of entertainment, navigation and communication applications in mobile environments by means of human-machine dialogues (Minker et al., 2004a).
- Clarissa - Fully voice-operated procedure browser, allowing astronauts to make more efficient use of their hands and eyes so as to give full attention to the task while they navigate through the procedure using spoken commands (Rayner et al., 2005).
- MIMUS - Multimodal dialogue system for the control of a smart home. It relies on a flexible architecture that allows for the integration of multiple input and output modalities (Pérez et al., 2006).
- STanford AI Robot (STAIR) - Robotic assistant, capable of conversation, for home and office (Krsmanovic et al., 2006).
- Cogniron - The cognitive robot companion (Menezes et al., 2007).

Finally, one of the most demanding applications for fully natural and understandable dialogues are virtual agents and companions:

- Collagen - Building conversational assistants and collaborative agents (Rich and Sidner, 1998).
- AVATALK - Natural, interactive dialogues with responsive virtual humans (Hubal et al., 2000).
- COMIC - Bathroom design using speech and gesture input/output, in collaboration with an avatar with facial emotions (Catizone et al., 2003).
- NICE - Embodies historical and literary characters capable of natural fun and experientially rich communication with children and adolescents (Corradini et al., 2004).

1.3 Evolution of SDSs

Human beings have always wanted to be able to communicate with artificial companions. There are many examples in cinema and literature. Some of the most ancient examples can be found in Greek and Roman mythology in which heroes could communicate with statues of goddesses or warriors. The first serious attempts at building talking systems were initiated in the 18th and 19th centuries, when the first automata were built to imitate human behaviour. The first of these were clockwork machines in which the masters applied all their skill to building animals or dolls that could produce sounds. In 1770, Baron Von Kempelen developed the first automaton that produced whole words and short phrases, which was subsequently improved by Josef Faber, who built the Euphonia machine in 1857. Euphonia imitated the mechanism of human speech by the use of a bellows which pumped air through a series of plates and chambers that could modulate the sounds by employing a 16-keys keyboard similar to a piano. The machine could speak any word in several European languages. These first machines were mechanical, and it was not until the end of the 19th century when scientists concluded that speech could be produced electrically.

At the beginning of the 20th century, Stewart (1922) built a machine that could generate vocalic sounds electrically. During the 30s, the first

electric systems were built that could produce any type of sound. The first was the VOCODER, a speech analyser and synthesizer developed by Bell Laboratories that could be operated by a keyboard. A skilled human operator could select either a periodic source for sonorant sounds or a noise source for fricative sounds that could be altered by controlling a filter bank. At the same time the first systems appeared with very basic natural language processing capabilities for machine translation applications.

During the 40s, the first computers were developed and some prominent scientists like Allan Turing pointed out their potential for applications demanding “intelligence”. To measure a machine’s capability to demonstrate intelligence, Turing (1950) proposed the so-called “Turing test” in which a human judge engaged in a natural language conversation with the machine. If the judge was not able to reliably tell whether he had talked with a man or a machine, the machine passed the test. This was the starting point that fostered the research initiatives that in the 60s yielded the first conversational agents. For example Weizenbaum’s ELIZA (Weizenbaum, 1966), which was based on keyword spotting and predefined templates. The templates allowed the user input to be transformed into system answers. For example, when the user wrote a sentence such as “I am X”, ELIZA would reply “How long have you been X?” independently of the meaning of ‘X’. Thus, although their behaviour was perceived as human by some naïve users and they might pass the Turing test, in practice the first conversational systems such as ELIZA did not semantically interpret the users’ input. To address this challenge, the research area of computational linguistics appeared in the 70s, grounded on the theoretical work developed in the 50s by Chomsky, Montague and Wood. The first rule-based speech synthesizers appeared about the same time. In the 70s the first continuous speech recognizers also appeared. These were based on decades of research on discrete speech in which verbal stimuli were punctuated by long pauses.

Benefiting from the continual improvements in the areas of speech recognition, natural language processing and speech synthesis, the first research initiatives related to spoken dialogue systems appeared in the 80s. To some extent the origin of this research area is linked to two seminal projects: the DARPA Spoken Language Systems programme in the USA and the Esprit SUNDIAL in Europe. On one hand, the main objective of the DARPA project was the study and development of technologies related to ASR and NLU in the domain of flight reservation by phone, to which they gave the name Air Travel Information Services (ATIS) (DARPA, 1992) (DARPA, 1994). The ATIS dialogue corpus, which is still employed by SDS developers and researchers, was followed by other projects, such as those carried out by AT&T, for example the AMICA project (Pieraccini et al., 1997), in which different stochastic models were applied to a mixed-initiative SDS. ATIS was also the starting point for research in MIT and CMU, where some of the most important systems in academia have been created. The SUNDIAL project was concerned with flight and train timetables in four different European languages. The research carried out in SUNDIAL yielded numerous projects funded by the European Community and mainly concerned with dialogue modelling, such as VERMOBIL (Bos et al., 1999), DISC (Bernsen and Dybkjaer, 1997) and ARISE (den Os et al., 1999). In the ARISE project six different systems were developed simultaneously: two Italian prototypes based on the technologies developed in CSELT (Castagneri et al., 1998) (Baggia et al., 2000), a French prototype developed by LIMSI (Lamel et al., 2000b) and two prototypes in Dutch and French based on the Philips technology. The DARPA ATIS project has been considered by some authors (Bangalore et al., 2006) to have been included in an earlier generation of SDSs than the SUNDIAL, as it was restricted to a closed application domain.

Among the most important research programmes of the 90s with multi-domain capabilities, the DARPA Communicator stands out. This government-funded project aimed to the development of cutting-edge speech technologies, which could employ as an input not only speech but also other modalities. The systems developed in this programme by both US and European partners were able to engage in complex interactions with the users in multiple domains, in which either the user or the system could begin the conversation, change topic or interrupt the other. For example, the CMU researchers developed the Carnegie Mellon Communicator system (Rudnicky

et al., 1999), which provided information on complex itineraries which included booking multiple flights, hotels and car rentals. The system architecture was based on specialized agents which developed modules that worked independently, encapsulating task-dependent information.

The most important research guidelines for the 90s were related to improving the success rates of the different components of the dialogue systems. Firstly, regarding speech recognition, their main concern was robustness. Authors argued about the sudden performance degradation in the systems due to minor changes, such as changing the microphones or the telecommunications channels used, and stated that the state-of-the-art technology used at that time was not capable of providing acceptable solutions. In order to overcome these problems, research was mainly focused on fundamental issues such as robustness (Cole et al., 1995), studying how to model the spectral characteristics (Holmes and Huckvale, 1994; Ostendorf et al., 1996), enhancing coarticulation models (Sun, 1997; Kirchhoff and Bilmes, 1999), modelling speech rate (Pfau and Ruske, 1998; Morgan et al., 1997) and providing speaker-independent recognition methods to cope with speaker differences. Robust speech recognition was acknowledged to be one of the most important aims in the context of the DARPA ATIS domain (Stern et al., 1992). In Europe, the COST Action 249, which took place from 1994 to 2000¹, with a research team from 20 European countries, was concerned with continuous ASR over the telephone, and covered all the previous topics, including selection of acoustic models, phonetic classification methods and adaptation to the characteristics of the telephone links.

In the late 90s, systems had to be enhanced with the ability to cope with differences between landline and cellular phones, as the latter were then becoming popular. The new systems required the ability to handle narrow channel bandwidths and low signal-to-noise ratios. Additionally, the adoption of cellular phones involved dealing with an increasingly richer variety of environments from which the users contacted with the systems and thus had to be robust to handle communication in very noisy backgrounds (Kacic, 1999). For example, one of the main topics studied for the development of the TREC-8 SDR system in LIMSI was noise compensation (Gauvain, 1999).

The main research guidelines in the field of NLU during the 90s were related to work with richer vocabularies, thus moving from isolated word

¹<http://www.elis.ugent.be/cost249/>

recognition applications to spontaneous speech. In order to do so, the authors conducted research on how to deal with out-of-vocabulary words, establishing successful and efficient vocabulary learning as a desirable future achievement. In this case, proposals were also related to studying highly specific methods and algorithms, such as enhancing linguistic analyses or trying to avoid them using more sophisticated word-spotting techniques (Zue and Glass, 2000).

In the 90s, great efforts were made in dialogue management towards the use of less restricted dialogues in which users could take the initiative in the communication. Thus, authors claimed for barge-in to be considered not only from the speech recognition point of view but also from the interaction perspective (Zue and Glass, 2000). Bangalore et al. (2006) describe three generations of spoken dialogue systems regarding, among other characteristics, the dialogue management initiatives and natural language understanding capabilities. Firstly, they describe a first generation in which SDSs used system-directed initiatives and the semantics were directly associated with the detection of keywords. The second generation was comprised of mixed-initiative dialogue systems in which the natural language understanding was carried out using frames. This allowed the users to talk naturally about a single task. Third generation SDSs additionally supported multiple tasks or domains simultaneously and could be enhanced with multi-modal and multi-media capabilities.

One of the major trends during the 90s was related to the definition of standard languages for the development of SDSs (Zue and Glass, 2000). For example, at the end of 1999 the W3C Voice Browser working group presented the first requirement studies for web browsers which settled the basis for the requirements of markup languages for spoken dialogues². The modularization of the systems to obtain more portable and re-usable components was mainly done by obtaining sets of frequent sub-dialogues. It was not until the late 90s when the first architectures for developing plug-and-play components appeared. For example, in 1998 the Galaxy architecture appeared, one of the pioneer works in fostering development of completely independent components for SDS (Seneff et al., 1998).

²<http://www.w3.org/TR/voice-dialog-reqs/>

There was also a generally positive attitude towards the adoption of stochastic approaches to obtain unsupervised methods to augment ASR and NLU capabilities (Glass, 1999). Thus, there was an increasing need for shared linguistic resources and data collections with which to train their algorithms. During the 90s, the first corpora development and system evaluation frameworks appeared yielding big shared resources such as WordNet.

Throughout the history of spoken dialogue systems, some experts have dared to envision what the future research guidelines in the area would be (see Figure 1.2). These objectives have gradually changed towards ever more complex goals. As reflected in the previously mentioned research results, during the 90s the major trends were towards making all system components (ASR, NLU, DM, TTS) more robust (Cole et al., 1995; Kacic, 1999; Zue and Glass, 2000; Mangold, 2001). Since 2003, experts have proposed higher level objectives, such as providing the system with advanced reasoning, problem solving capabilities, adaptiveness, proactiveness, affective intelligence, multimodality and multilinguality (Dale, 2003; Jokinen, 2003; Gao et al., 2005; Haas et al., 2005; Minker et al., 2006b,a). As can be observed, these new objectives refer to the system as a whole, and represent major trends that in practice are achieved through joint work in different areas and components of the dialogue system. For example, for a system to be multilingual, it has to be able to recognize the users' utterances in different languages, interpret their semantics using the corresponding linguistic knowledge, and also generate natural language and synthesize speech in all of them. Thus, in contrast to what happened in the 90s, when each area had different objectives (ASR, NLU, DM, NLG and TTS), current research trends are characterized by large-scale objectives which are shared out between the different researchers in different areas.

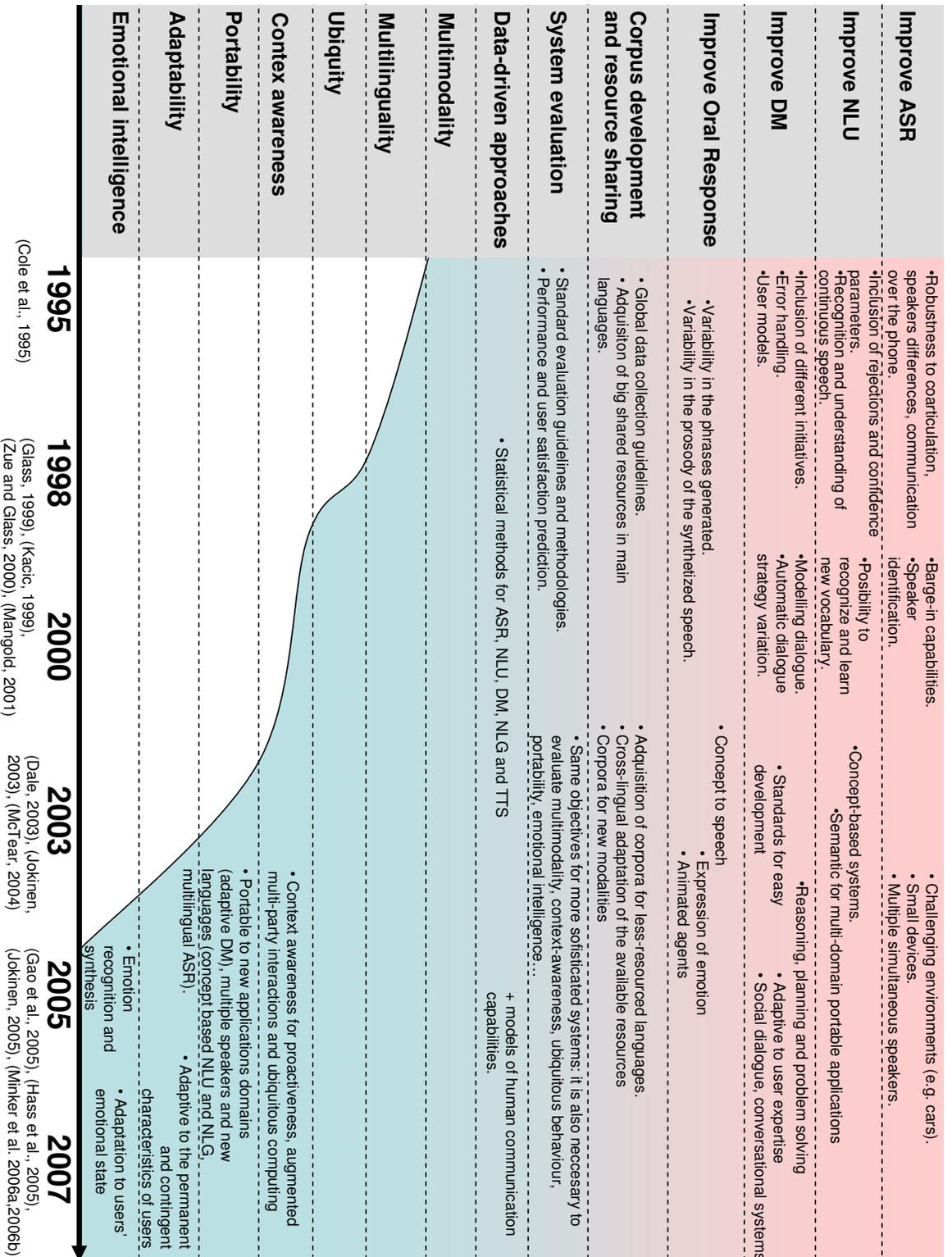


Figure 1.2. Evolution of the envisioned research lines in the area of spoken dialogue systems

Specialists have recently envisioned future dialogue systems as being intelligent, adaptive, proactive, portable and multimodal. All these concepts are not mutually exclusive, as for example the system's intelligence can also be involved in the degree to which it can adapt to new situations, and this adaptiveness can result in better portability for use in different environments.

Proactiveness is necessary for computers to stop being considered a tool and becoming real conversational partners. Proactive systems have the capability of engaging in a conversation with the user even when he has not explicitly requested the system's intervention. This is a key aspect in the development of ubiquitous computing architectures in which the system is embedded in the user's environment, and thus the user is not aware that he is interacting with a computer, but rather he perceives he is interacting with the environment. In such situations, proactiveness enables the system to passively observe the dialogue between human users and capture the relevant conversational context, which the system processes to compute when the conversational situation requires it to take the initiative and get meaningfully involved in the communication. To achieve this goal, it is necessary to provide the systems with problem-solving capabilities and context-awareness. For example, Schneider (2004) presents a proactive shopping assistant integrated in supermarket trolleys. The system observes the shopper's actions and tries to infer his goals. With this information, it proactively offers adapted support tailored to the current context, as for example displaying information about products when the user holds them for a very long time or comparing different products when the user is deciding between two items. Other proactive systems are described in (Baudoin et al., 2005; Kwon et al., 2005).

The interest in developing systems capable of maintaining a conversation as natural and rich as a human conversation, gave rise to research on multimodal interfaces. Unlike traditional keyboard and mouse interfaces or unimodal speech dialogue systems, with multimodal interfaces there is flexibility in the input and output modes, such as speech, gestures or facial expressions. As multimodality permits users to employ different input modalities as well as to obtain responses through different vias, it is specially important for users with special needs, for which the traditional interfaces are not suitable. The first multimodal dialogue systems appeared in the mid-nineties, basically combining speech with pointing maps (Cheyer and Julia, 1995; Oviatt et al., 1997) and speech with pen input (Waibel et al., 1997). Eye movements

and gaze were one of the first modalities being studied, in order to identify which objects the user was referring to in his conversation (Sarukkai and Hunter, 1997). Most of these systems focus on using multimodality in the input or output, but some recent large projects have been directed towards the development of full multimodality. This is the case of the SmartKom project, which provides what is called “full symmetric multimodality” in a mixed-initiative dialogue system (Wahlster, 2006). They define symmetry as the capability of the system not only to understand and represent the user’s modal input but also to create multimodal output. The main contribution of the Smartkom project is that it not only deals with modality integration or synchronization but also covers the dialogue phenomena that are associated with multimodality, such as mutual disambiguation, multimodal deixis and cross-modal reference resolution and generation, multimodal anaphora and ellipsis resolution and generation and multimodal turn-taking.

Adaptivity may also refer to other aspects in speech applications. In speech-based human-computer interaction users have diverse ways of communication. Novice users and experienced users may want the interface to behave completely differently, for example to have system-initiative instead of mixed-initiative. An example of the benefits of adaptivity in the interaction level can be found in (Litman and Pan, 2002). Multilingual applications are another example of adaptive applications. Multilingual recognizers are capable of recognizing simultaneously several languages by sharing acoustic and/or language models. Multilingual acoustic models consist of either a collection of language-dependent acoustic models for each language, or a combination of language-independent acoustic models (Schultz and Kirchhoff, 2006). The development of a speech recognizer is a very arduous and time demanding task. A large amount of data spoken by hundreds of subjects must be recorded and carefully annotated to get a representative set suitable for training the acoustic models. To overcome this problem there has been increasing interest in developing rapid prototyping approaches. Cross-lingual approaches have arisen in order to share language-recognition resources. This is particularly interesting for the development of speech recognizers for the less-resourced languages, which is a very important research area nowadays, and many of the main conferences in speech technologies explicitly devote sessions to studying this topic, as for example the Interspeech 2007 Special Session on Speech and language technology for less-resourced languages

(2007) and the LREC 2008 Workshop on Collaboration: interoperability between people in the creation of language resources for less-resourced languages (2008).

As stated by Jokinen (2003), there are different levels in which the system can adapt to the user. The simplest one is through personal profiles in which the users have static choices to customize the interaction (e.g. whether they want a male or female system's voice), which can be further improved by classifying users into preferences' groups. Systems can also adapt to the users' environment, for example ambience intelligence applications such as the ubiquitous proactive systems described. A more sophisticated approach is to adapt to the user's knowledge and expertise. The main research topics are the adaptation of systems to different user expertise levels (Haseel and Hagen, 2005), user knowledge (Forbes-Riley and Litman, 2004b) and user special needs. This last topic is receiving a lot of attention in terms of how to make systems usable by handicapped people (Heim et al., 2007), children (Batliner et al., 2004) and the elderly (Langner and Black, 2005) and also to adapt to permanent features like users' age, proficiency in the interaction language (Raux et al., 2003) or the user's expertise in using the system (Haseel and Hagen, 2005). Despite their complexity, these characteristics are to some extent rather static, Jokinen identifies another degree of adaptation in which the system not only adapts to the explicit message conveyed during the interaction, but also to the user's intentions and state. Following this guideline, affective computing studies focus on how to recognize and adapt to the user's emotional state during the conversation with the system.

There is an increasing interest in the development of spoken dialogue systems that dynamically adapt their conversational behaviours to the users' affective state. Martinovski and Traum (2003) demonstrated by means of user dialogues with a training system and a telephone-based information system that many breakdowns in man-machine communication could be avoided if the machine was able to recognize the emotional state of the user and responded to it more sensitively. Earlier experiments by Prendinger et al. (2003) showed that an empathetic computer agent can indeed contribute to a more positive perception of the interaction. Polzin and Waibel (2000) present a first approach to adjusting dialogue behaviour to the user's emotional state. For instance, they suggest that more explicit feedback should be given if the user is frustrated. Nevertheless, their approach relies on few

selection rules and is not based on a general framework for affective dialogue. Walker et al. (1997) examined how social factors, such as status, influence the semantic content, the syntactic form and the acoustic realization of conversations. Emotional intelligence includes the ability to recognize the user's emotional state as well as the ability to act on it appropriately. At present, there are a number of projects whose main objective is to endow dialogue systems with emotional intelligence. Some of the latest are MEGA (Camurri et al., 2004), NECA (Gebhard et al., 2004), VICTEC (Hall et al., 2005), NICE (Corradini et al., 2005), HUMAINE (Cowie and Schröder, 2005) and COMPANIONS (Wilks, 2006), to mention just a few.

Portability is currently addressed from very different perspectives, the three main ones being domain, language and technological independence. Ideally, systems should be able to work over different application domains, or at least be easily adaptable between them. Current studies on domain independence centre on how to merge lexical, syntactic and semantic structures from different contexts (Chambers and Allen, 2004) and how to develop dialogue managers that deal with different domains (Mourão et al., 2004; Nguyen and Wobcke, 2006). Regarding language independence, multilingual systems (Schultz and Kirchhoff, 2006) are those which can work with several languages and are thus portable in two main senses: firstly, users can input information in different languages, and secondly they can also receive the response in different languages. This is specially useful in speech-to-speech systems, which can serve as real time interpreters, so that as the first speaker talks on the telephone, the other receives the information translated into another language.

Finally, technological independence deals with the possibility of using dialogue systems with different hardware configurations. Computer processing power will continue to increase, with lower costs for both processor and memory components. The systems that support even the most sophisticated speech applications will move from centralized architectures to distributed configurations and thus must be able to work with different underlying technologies. To achieve this objective, different standard architectures have been studied. The main alternative, as gathered from the efforts of both universities and business companies, is that proposed by the MMI Working Group in W3C³, whose goal is to provide a general and flexible framework ensuring

³<http://www.w3.org/TR/mmi-arch/>

interoperability among modality-specific components from different vendors and underlying technologies.

1.4 Objectives of the Thesis

The main goal of the Thesis was to contribute to fostering the adaptiveness and portability of SDSs, which are the main current trends established by the experts in the area. The work presented in this Thesis consists of the development of models and methods for handling several of the issues described above, and experimentally test their performance with real dialogue systems. Of course, all aspects could not be included in the scope of the Thesis and it was decided to study three of the most challenging topics: emotion recognition, cross-lingual adaptation and field evaluation.

- Emotion recognition. The main objective was to find decisive factors in the recognition of emotions that influence both human and machine-learned recognition. Most state-of-the-art approaches are based on the use of simulated emotions in order to have strict control over the corpora used to train the automatic procedures. The aim of the Thesis was to use natural user interactions with a real dialogue system. This is a particularly challenging objective as in this case emotions are produced more subtly and the proportion of emotional utterances is very low compared with neutral cases. The Thesis centres on studying the recognition of negative emotions that can make the interaction with SDSs fail.
- Cross-lingual adaptation. One of the main claims of the research community has been the need for common resources and for ways of making the most of those available. This is vital for research centred on minority languages or dialects. The objective was to develop a technique that would allow fast adaptation of a speech recognizer to work in another language without the need of building new acoustic models. To maximize portability the goal was to make it employable not only between languages with similar origins, but also with very different languages.
- Field evaluation. To be able to build systems that adapt to the users' needs and expectations it is important to take into account their opin-

ions about previous interactions with the systems. In the literature, most evaluations are carried out with laboratory studies and consider interaction parameters and quality judgments separately. The main objective was to discover significant relationships between both interaction parameters and quality judgements in a field study on non-restricted interactions of the users with a real system.

1.5 Structure of the Thesis

After this concise overview of the main research challenges for the creation of adaptive and portable systems, the contributions of the Thesis are presented on chapters 2 to 5. Each of the chapters is structured as follows: they start with an introduction which explains the objective of the research carried out; then a detailed state-of-the art of the topic is presented and the Thesis contributions are compared with prior work in the dialogue literature. The Thesis relies to a large extent on an empirical approach in which all the proposals are evaluated with real dialogue systems. In each of the chapters there is a section devoted to explaining the experimental set-up, followed by others that present the experimental results and conclusions extracted from them.

Chapter 2 describes the UAH spoken dialogue system. The system was carefully developed using dynamic dialogue management and different dialogue initiatives and confirmation strategies. The objective was to study the effect of each of them on the user's opinions about interactions with the system. In this chapter, the functionality and structure of the UAH system is discussed, giving a detailed description of each of its modules and highlighting its innovative functionalities. The system is accessible to the general public on the telephone and the users' calls were recorded and later annotated. The resulting corpus was used for the research on the field evaluation of SDSs described in chapter 5 where the annotation procedure is roughly described. The UAH corpus was also employed for the emotion recognition research in chapter 3. Thus, UAH was developed using innovative approaches and was also a testbed for the different proposals, i.e. the approaches proposed in the Thesis have been successfully employed in a real SDS.

Chapter 3 presents the research done on emotion recognition and includes a detailed study on the impact of considering context information for the annotation of emotions. The inclusion of the history of user-system interaction and the neutral speaking style of users is proposed. From the human annotation perspective, special attention is paid to how to reliably compute inter-annotator agreement. From the machine-learning perspective, a new method to automatically include both sources of information has been developed, making use of novel techniques for acoustic normalization and dialogue context annotation. The proposals were tested with the corpus extracted from the interactions of approximately 60 users with the UAH system.

Chapter 4 describes the research on cross-lingual adaptation, introducing a methodology that proved to be successful and cost-efficient for porting an existing speech recognition system to other languages is proposed, which was tested with the MyVoice system⁴. The methodology was tested with three different languages, Czech, Spanish and Slovak, to find out whether it was possible to use the proposed method not only with languages of similar origin (in this case Czech and Slovak), but also with languages that belong to very different branches of the Indo-European language family (such as Czech and Spanish). The proposed approach consists of four steps: 1) creating an initial mapping between the phonemes in the original language and the target language, 2) creating a lexicon and automatically mapping the words' pronunciation to the original phoneme set, 3) fine-tuning the models of the phonemes that are unique to each particular language, and 4) collecting data in the new target language and carrying out speaker adaptation. The research described in this chapter was done in collaboration with researchers from the Technical University of Liberec (Czech Republic), during a three-month stay in the Laboratory of Computer Speech Processing⁵ under the supervision of Prof. Jan Nouza.

Chapter 5 shows the work done on the field evaluation of the UAH system. Evaluation of spoken dialogue systems has been traditionally carried out in terms of instrumentally or expert-derived measures (usually called “objective” evaluation), and quality judgments of users who have previously interacted with the system (also called “subjective” evaluation). Different research efforts were made to extract relationships between these evaluation

⁴MyVoice was developed in the Technical University of Liberec (Czech Republic)

⁵<http://itakura.kes.tul.cz/kes/indexe.html>

criteria. In this chapter empirical results obtained from statistical studies are reported. These studies were carried out on interactions of real users with a real SDS, something which is rarely found in the literature.

The Thesis ends with Chapter 6, which presents the overall conclusions, gives a detailed overview of the main contributions, and summarizes the main guidelines for future work. Finally, the Appendix presents the publications referring to the research described in the Thesis.

*Puisque c'est ell que j'ai écoutée se plaindre,
our se vanter, ou même quelquefois se taire.
Puisque c'est ma rose*

Antoine de Saint-Exupery, Le Petit Prince

2

The UAH spoken dialogue system

2.1 Introduction

In this chapter, the Universidad al Habla (UAH - University on the Line) spoken dialogue system is described. Firstly, in Section 2.2 there is a detailed description of the architecture of the system describing its main modules. The design principles followed are described, including efficient database access based on the interaction context, dynamic dialogue managing, tailored help messages and context-based oral responses. Special attention is paid to the creation of grammars with vocabulary which is previously unknown. Section 2.3 proposes a novel approach for the creation of speech recognition grammars from databases. This technique permits a reduction of the time employed during the speech recognition process to obtain a rapid system response to user prompts and also improves user satisfaction, as interaction is smoother.

The development of the UAH system was carried out not only to design and implement different dialogue management strategies in a real dialogue system, but also to employ a real spoken dialogue system as a testbed in which to evaluate the different methodologies proposed in the Thesis. The system has been available to the public on the phone from June 2005, so that the users can interact with UAH to find out information related to the Dept. Languages and Computer Systems. A corpus of 422 user recordings corresponding to a year of UAH use was compiled and annotated, and was used for the approaches presented in the rest of the Thesis.

2.2 Modular architecture

UAH is a spoken dialogue system developed to provide spoken access to academic information of the Dept. Languages and Computer Systems, as well as some other additional information relating to the University of Granada. The first functional version of UAH was made available to the public in June 2005. Figure 2.1 shows the modular architecture of the UAH dialogue system. As can be observed, it is comprised of the five typical modules of current dialogue systems, concerned with automatic speech recognition (ASR), dialogue management (DM), database access (DB Access), data storage (DB) and oral response generation (ORG). In addition, there is a new module called GAG (Grammar Automatic Generation) which permits the automatic creation of the rules of grammars for speech recognition, which will be discussed in Section 2.3.

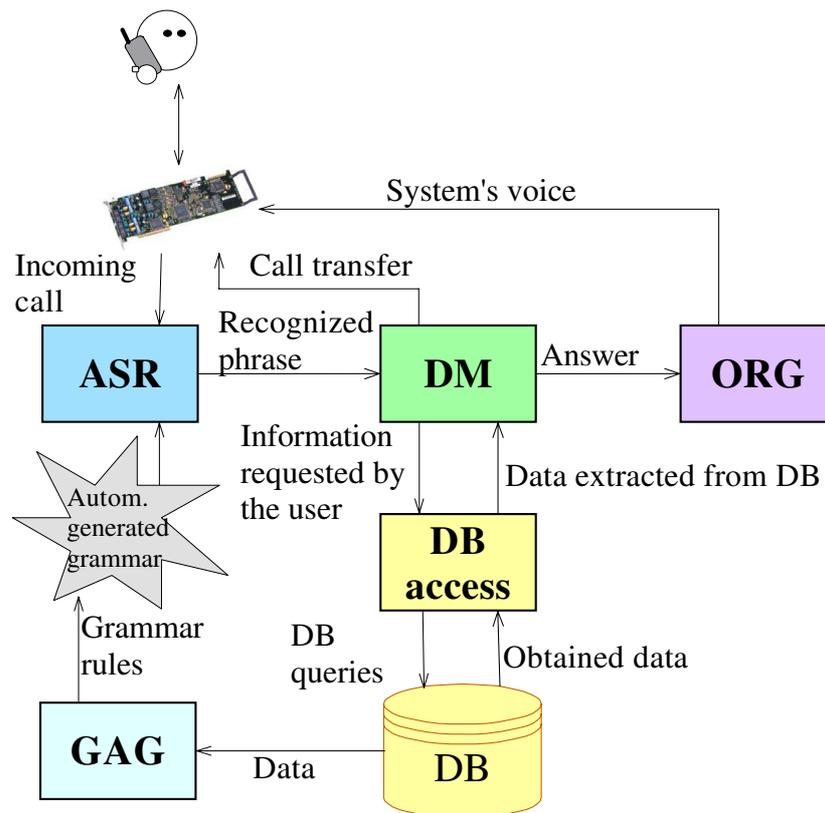


Figure 2.1. Modular architecture of the UAH system

According to the data extracted from the user utterances, the dialogue manager (DM) decides the next system response. The dialogue manager is based on VoiceXML documents that are dynamically created using PHP as the interaction flows. UAH was designed to provide information about teachers, subjects, registration processes and post-graduate studies. In these scenarios different dialogue initiatives were used to obtain measures about how the interaction flexibility affects both the performance of the dialogue and user satisfaction. Information on teachers and subjects had to be accessed using a system-directed initiative, whereas dialogues providing information about registration and post-graduate studies were mixed-initiative.

The UAH system's dialogue manager dynamically adapts the system responses to the context and the dialogue state, for example to decide the confirmation strategy to be used. The system uses explicit confirmations for important actions or situations in which the ASR process is more difficult. For example, before transferring a call to a teacher the system explicitly confirms his/her name (e.g. "Did you say Mr. Smith?"). In the rest of the situations, the system uses explicit confirmations when confidence values are not high, and implicit confirmations otherwise, in order to provide the user with a more natural interaction. The system's oral response generation is carried out firstly in text format instantiating patterns. The phrases are tailored to the interaction context. For example, the system's help messages takes into account the topic the user and the system are talking about at a particular moment. Once the response is obtained in text format, it is transformed into voice using a commercial TTS system.

For purposes of illustration, we give below a sample dialogue translated from Spanish into English (S = system, U = user):

S1> Welcome to the UAH system. How can I help you?
U1> I need information about the registration procedure in the University.
S2> What do you want to know about the registration procedure?
U2> The deadline.
S3> The registration deadline is on the 31st of August. Do you need any other information about the registration procedure?
U4> No, thanks.
S5> Thank you for using the UAH system. Have a nice day.

2.2.1. Automatic speech recognition

The ASR module processes each user utterance provided by an Intel Dialogic D/41JCT-LS telephony card. This card handles the user call and gives the voice signal as a result, which is transformed into a word sequence in text format by the ASR module. The ASR process is carried out using several grammars that represent the valid phrases the users may utter. These grammars are created in different ways depending on the vocabulary and the moment of creation. In total there are four creation methods considering the combination of: i) known or unknown vocabulary at design time, i.e. never changing vocabulary, or vocabulary that can be constantly updated, as for example if it is stored in a database, and ii) static or dynamic grammar creation.

Static grammars with previously known content were created during the design of the dialogue system and do not vary their content with its operation. Their content is previously known and was carefully designed. These grammars are used in the UAH dialogue system for static menus that explain the information that the user can ask about. The UAH system also uses built-in grammars for the recognition of numbers, and Boolean answers like yes/no/true/false. Static grammars with unknown vocabulary will be described in Section 2.3.

Dynamic grammars are used in two ways. On the one hand, they can have a previously known content, which is dynamically generated in execution time only when they are really going to be used. On the other hand, UAH also employs grammars which are created dynamically from the vocabulary extracted from databases. This vocabulary is continuously changing and thus is not known at design time. This approach is used only when grammars have a small vocabulary, because creating grammars from large vocabularies at execution time would imply a delay in the interaction. This type of grammar is used, for example, in disambiguation grammars.

2.2.2. Dialogue management

The dialogue manager (DM) decides which system response to follow according to the data extracted from the user utterances. The complexity of this module depends on various factors. One is complexity of the modelled interaction, while another is the kind of task to be carried out by means of the

dialogue. The DM complexity depends as well on the flexibility desired for the dialogue and the type of initiative implemented (user directed, system directed or mixed). In the case of the UAH system, the application domain is information extraction in the University environment. The dialogue initiative is system-directed in the so-called basic election dialogue states, while it is mixed in the detail election states. In the basic election dialogue states the user decides, without going into detail, the kind of information s/he wants to know (e.g. information about teachers, subjects, timetables, etc.). In this case, the dialogue initiative is system-directed in order to ensure the user obtains the required information easily from the diverse kinds of information the system provides. On the other hand, in the detail election dialogue states, the user identifies the specific information s/he wants to obtain inside the previously selected area (e.g. a teacher's telephone number). The use of mixed initiative in these states provides the user with greater expression flexibility once the area has been determined (Narayanan et al., 2000). The system can take the initiative even in a detail election state if it detects a misunderstanding. In this case it provides the user with additional means for introducing the information. For example, if the user makes more than two mistakes in giving a teacher's full name, the system takes the initiative and gives prompts for the data one by one, i.e. first name and then surnames¹.

Although there is no consensus about the tasks a dialogue manager should perform, one of the most widely accepted proposals is presented by Traum and Larsson (2003). They propose that the dialogue manager must update the dialogue context to obtain the correct semantic interpretations from the user utterances. Furthermore, the dialogue manager must carry out tasks within a specific domain and decide which information to provide to the user and when and how to express it.

The UAH dialogue manager dynamically adapts the system responses to the context and the dialogue state, changing some phrases to improve the naturalness of the interaction. For example, the system's help messages take into account the topic the user and the system are talking about at a particular moment.

¹In Spain people have two surnames: father's and mother's

The context is used as well to decide the confirmation strategy to use. The system uses explicit confirmations for important actions or situations in which the ASR process is more difficult. For example, before transferring a call to a teacher the system confirms explicitly his/her name (e.g. “Did you say teacher Ms. Zoraida?”). In the rest of the situations the system uses implicit confirmations, in order to provide the user with a more natural interaction (Bernsen et al., 1994).

2.2.3. Database access

In the literature it is possible to find multiple references to the importance of separating the access and query of databases from the rest of the tasks carried out in a dialogue system. For example, in the GEMINI project (Hamerich et al., 2004), an assistant was built to connect to the database, so that users could create dialogue systems semi-automatically, regardless of the features of the database employed. A dichotomy between dialogue management and information access is achieved by means of creating a database access module. The dialogue manager supplies to this module the information that the user wants to know. The access module then constructs the query and extracts the data from the database.

Once the information is extracted, the access module executes a PHP program that validates the data. Besides, it checks that there are no repeated data in the result before sending it to the dialogue manager. Finally, the dialogue manager decides how to communicate the data to the user. However, there are situations in which the access module receives not only the kind of information to be extracted (e.g. telephone number of teachers called “García”), but also additional restrictions, such as gender. For example, if the user asks for Mrs. García’s telephone number, then only data of female teachers with this surname are extracted.

The UAH system makes two types of database queries: explicit and implicit. In the first case, the query is carried out by the user initiative. For example, if the user asks for the telephone number of the teacher “Juan García”, then two queries are executed. The first one checks the number of records in the teacher’s database that correspond to the teacher’s name provided. If the number is zero, the system informs the user no teachers were found with this name. If the number is greater than one, the system requests

the user to provide additional information to select the correct record. Once the name is determined, another query takes place to extract the telephone number. The first query is implicit because it is executed by the system's initiative, without an explicit user request. On the other hand, the second is explicit because it corresponds to the user request (in this example the user asked for a telephone number).

Furthermore, the complexity of the queries varies with the flexibility of the current dialogue state. In fact, queries vary from a simple "select FIELD from TABLE" to a very complex data selection based on partial or complete string matching. For example, when a user utters a teacher's name, the information provided can be a combination of name and surnames in up to seven different possibilities (e.g. name, name and first surname, first and second surname, etc.), including incomplete names (e.g. a user could say "María" instead of "María José").

The system extracts information querying a database which contains public information about the University of Granada (e.g. fax number and email of teachers, but not their personal data). The database has been designed to store data typically included in the departments' web pages. The UAH system could work with a database that stored many different kinds of information, creating views, if necessary, to keep economic and personal data private. The database used in the system's implementation is relational and has 22 tables that have been obtained from an entity-relation diagram, followed by a fusion and normalization process.

2.2.4. Oral response generation

The system's oral response generation is carried out firstly in text format using patterns which have been classified into several categories: teacher, PhD, subjects, registration, additional information, confirmation, greetings and help. Once the response is obtained in text format, it is transformed into voice using a commercial TTS system.

Inside each response category there is a specific pattern for each data the system could give to the user. These patterns are composed of several information segments that can be selected or ignored dynamically depending on the information to be provided. For example, the pattern used to inform about the location of a teacher's office is composed of three main information

segments: number of the office, name of Faculty where the office is, and floor inside the Faculty. If one of these data is not available at the current dialogue state, it is not included in the resultant text. For example, if the floor is not available the system response could be “office number 3 in Computer Science Faculty”.

The patterns’ morphological structure can also change according to the information provided, as it must take into account gender (e.g. “Ms. Mary Jones”, “Mr. John Williams”) and number (e.g. “The subject Parallel Programming has two credits”, “The subject Software Engineering has one credit”). The system also uses a dynamic adaptation of linking words (e.g. “There are two teachers called Jones: Mary Jones and Michael Jones”, “There are three teachers called Jones: Mary Jones, Michael Jones and William Jones”).

Finally, the last stage in the creation of the system response in text form is the adaptation of several special words and symbols, as well as the inclusion of tags the TTS module uses to generate special pronunciations. The special words’ adaptation translates the information stored in the database into a more suitable form to be read aloud by the system, whereas the tag inclusion permits a better understanding of the information synthesized by the system. For example, it is better to spell a web page’s URL than reading it as if it were a long word. Also, it is preferable to read aloud a date in the format month-day-year than reading it.

2.3 Automatic grammar generation

Most of the commercial development environments for setting up spoken dialogue systems provide tools for rule-based grammar creation and testing. In these tools the grammar creation is done statically; that is, grammars are created in design time, before the system is put into service. Although this is the most straightforward way of creating grammars, static creation can cause inconsistencies between the database contents and the vocabulary included in the grammars when the vocabulary is not previously known or does not remain unchanged.

To overcome this drawback, there are several techniques for dynamic grammar creation. The first constructs the grammars dynamically in execution time (Truillet et al., 2004). This technique offers flexibility as the

grammars are always updated with the last changes in the database. Nevertheless, this method may imply a very high computational load, which introduces a delay time that, in systems with very large databases, can be excessive; thus causing the system to be considered “slow” by users.

A second technique creates the grammars at the beginning of each interaction, before the ASR process takes place. These grammars are always updated regardless of changes in the database (Schalkwyck and Story, 2003). This method does not imply an increase in the ASR time. There is therefore less increase in waiting time from the user point of view, but there is still a delay in the system start-up that increases with the database size.

In both cases, the suitability of the technique depends on the size of the vocabulary. As has been shown previously (Nielsen, 1994), there are three basic execution times to be considered. Firstly, the limit for the user to consider the interaction as real-time is 0.1 seconds. Secondly, from 0.1 to 1.0 seconds, which is the interval in which the user notices the delay but his flow of thought is uninterrupted. And thirdly, from 1.0 to 10.0 seconds, when the user still pays attention to the system. However for delays longer than 10 seconds it is necessary to give some feedback to the user with a system prompt and/or music (Cerrato, 2002).

The execution time needed to construct a grammar rule from a specific column in a database was measured regarding the number of words gathered. As shown in Figure 2.2, the 0.1 limit is reached with 10,000 words, the 1 second limit is reached with 100,000, and the maximum limit (10 seconds) is reached with vocabularies around 300,000 words. For example, for a vocabulary of 1 million words there is a 30 seconds delay. Thus, for large vocabularies (more than 300,000 words) the overload introduced is noticed by the user and a feedback technique is needed. However, feedback techniques usually cause a bad impression on users, who usually find them very irritating (Mäkelä et al., 2001). Furthermore, a minimal system response delay (defined as the time elapsed between the user finishing talking and the system response (Möller, 2005)) is vital to minimize interaction costs. This, coupled with the maximization of task success, is essential to enhance user satisfaction and is usually employed as a key parameter for spoken dialogue systems evaluation, as proposed in the PARADISE framework (Walker et al., 2000a).

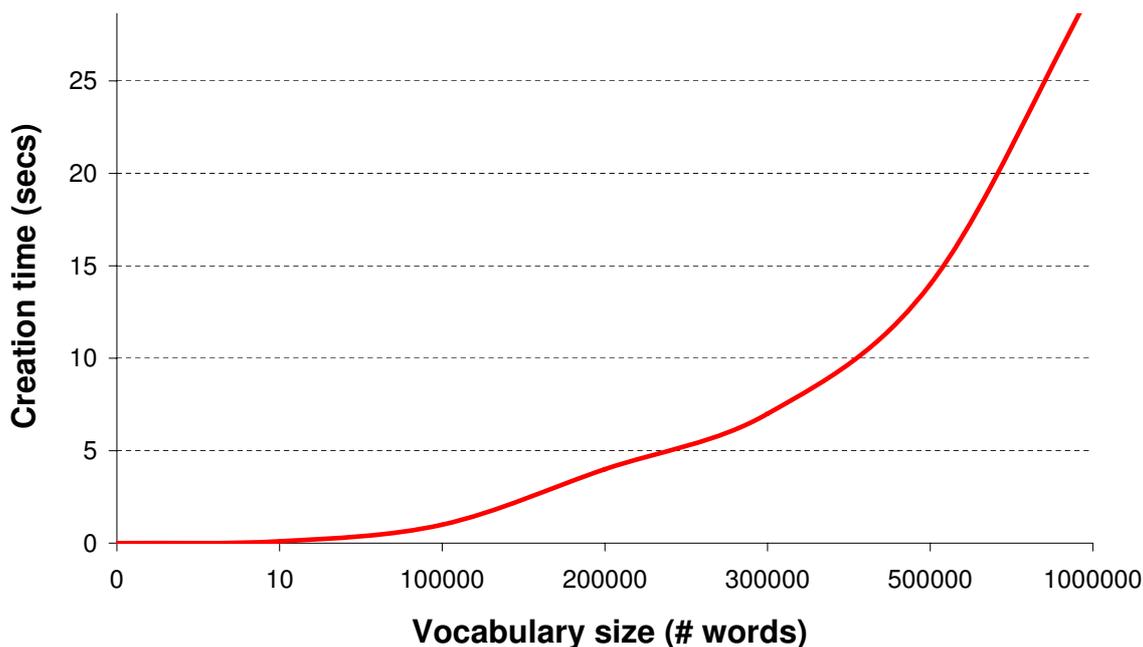


Figure 2.2. *Latency of the dynamic creation of grammar rules*

To avoid the problems discussed above, the TGC technique (Triggered Grammar rules Creation) can be employed. TGC uses automatically-generated static grammar rules which are updated using a triggering mechanism as the database changes. This way there is no creation delay in execution time and the user feels he is interacting with a fast system.

The main advantage of the TGC technique in comparison with others we can find in the literature (e.g. McTear (2004)) is that it is general and thus suitable for generating grammar rules to be used by any dialogue system that extracts vocabulary from a database.

The creation of grammar rules is carried out following three steps: i) information is extracted from the database, ii) the grammar rules are created from the extracted data formatting it into JSGF (Java Speech Grammar Format) or ABNF (Augmented Backus-Naur Form); in this way, grammar rules are generated before the ASR is carried out; iii) grammar rules are updated when the database changes. To ensure grammars are permanently updated when the database changes, the proposed technique employs a mechanism based on database triggers. These are fired when changes occur in the database fields from which the vocabulary of the grammar rules is extracted.

The W3C Voice Browser working group specifies in (Brown, 1999) that rules must be re-definable in execution time. To do so, they point out several mechanisms, as for example the division of the rules space into static and dynamic arenas. Using the proposed technique, the sentence structure is created statically containing references to dynamic rules, which can be stored in the same file or in an external one. The TGC technique updates these rules with new data extracted from the database without affecting the syntactic structure of the phrases.

The TGC technique was implemented in the so-called Grammar Automatic Generation (GAG) tool using PHP, HTML, JavaScript and PostgreSQL. This tool is used as an additional module in the UAH system and has an easy-to-use web interface to let the system designer choose the database fields to be used to extract the vocabulary. The process of creating rules using the tool is divided into three steps. Firstly, the GAG tool prompts the user to provide the database name, host, user name and password to access the database where the information is stored. It also prompts for the name of the grammar rule and the type of grammar to be created (JSGF or ABNF). Secondly, the selection of fields from the database tables is visualized in a drop-down menu. Fields appear in the menu in the same order as in the corresponding table, but the system designer can select them in any other to create a grammar rule. When a field is selected, a number appears automatically next to the field indicating the selection order. For example, in an academic application the system designer could select a teacher's name and surname, obtaining the grammar rule `<teacher> = ("Claire Smith" | ... | "Jack McNeal")`. Alternatively, he could select for example the teacher's surname and name, in which case the obtained grammar rule would be: `<teacher> = "Smith Claire" | ... | "McNeal Jack"`. Thirdly, after the field selection the designer must enter the file name to store the grammar rule. If he marks the "Add to file" option in the interface, the rule is created as a part of an existing grammar. Otherwise, a new grammar with just one rule is created in the specified file.

The GAG tool implements a triggering mechanism to keep the grammar rules updated with the last changes made in the database. To do so, at the end of each grammar creation the system designer chooses whether to update automatically the vocabulary in the rule with database changes. If he indicates the rules must be automatically updated, triggers are dynami-

cally created to be activated when the values of the corresponding columns are updated, deleted or inserted. For example, if he creates a grammar rule for teachers' names and surnames, the trigger includes "obligatory" as a new word in the rule if this word is in the field "Type" of the "Subject" table. It also deletes the word "obligatory" from the rule if it is erased from the table, and changes "obligatory" to "optional" if the field is updated with the latter word.

As triggers provide the old and new values of the table field, only the pertinent changes are made in the grammar rules instead of generating them again right from the start when any change takes place, which introduces a delay equivalent to creating them dynamically. The whole process is illustrated in Figure 2.3, where a distributed architecture is depicted in which the ASR takes place in a computer with a telephony card, while the dialogue manager and the GAG tool operate in an application server that uses the data provided by a server in which the database is hosted.

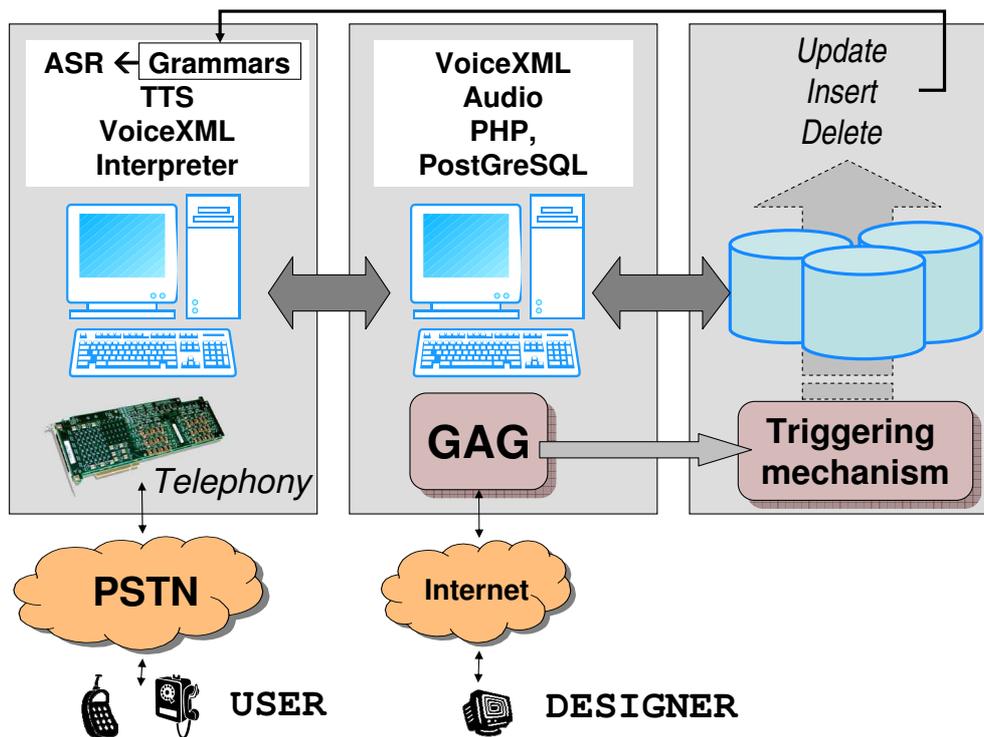


Figure 2.3. Automatic grammar rules update with the GAG tool

Two different approaches to addressing the creation of ASR grammar rules from databases were tested: dynamic creation and creation using the GAG module. The experiments were carried out employing different vocabulary sizes (as shown in Figure 2.2). Figure 2.4 shows that user satisfaction with dynamic grammar creation decreases with larger vocabularies, which indicates that user preference for the GAG tool outperforms preference for dynamic creation in most cases. There is a correspondence between user satisfaction and the temporary constraints previously commented.

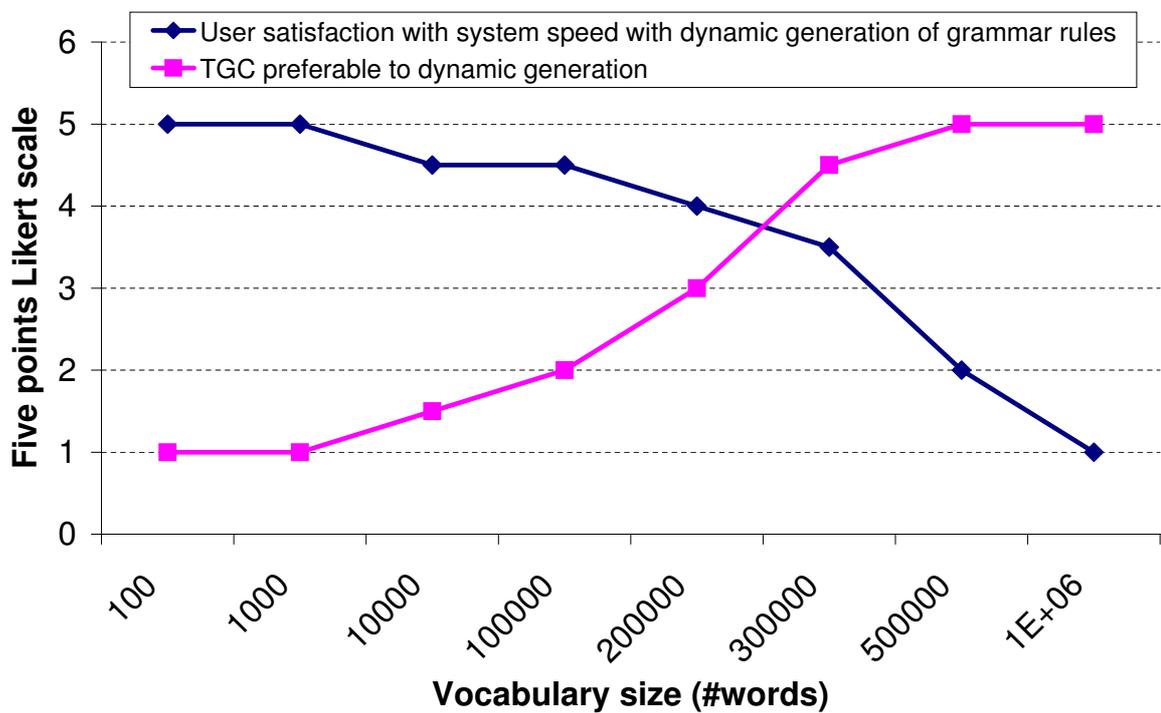


Figure 2.4. *TGC technique vs. dynamic creation of grammar rules measured in terms of user satisfaction*

Thus, in vocabularies with less than 10,000 words (generation time less than 0.1 seconds) users are not aware of the system delay with the dynamic technique, which makes both techniques be equally rated by users. It is also observable that in this figure there is a point at which both lines cross. This corresponds to the value around 300,000 words for which a vocabulary can be considered large enough to make the proposed technique clearly preferable to dynamic creation, given the delay introduced by the latter and its consequent negative effect on users' satisfaction.

Furthermore, regarding the use of the GAG tool, it is shown that the users' perception varies depending on the type of interaction initiative used by the dialogue system. Although 71% of the users find the interaction speed adequate or fast, 20% of them consider it slow (only 3% thought it very slow). This occurs because of the system-directed initiative, which presents several options to the user in order to obtain the kind of information he wishes. As shown in Figure 2.5, overall user satisfaction with the interaction (rated on a five-point Likert scale) is in 78% of cases greater than or equal to 3, which shows the good results achieved by the system employing the proposed technique.

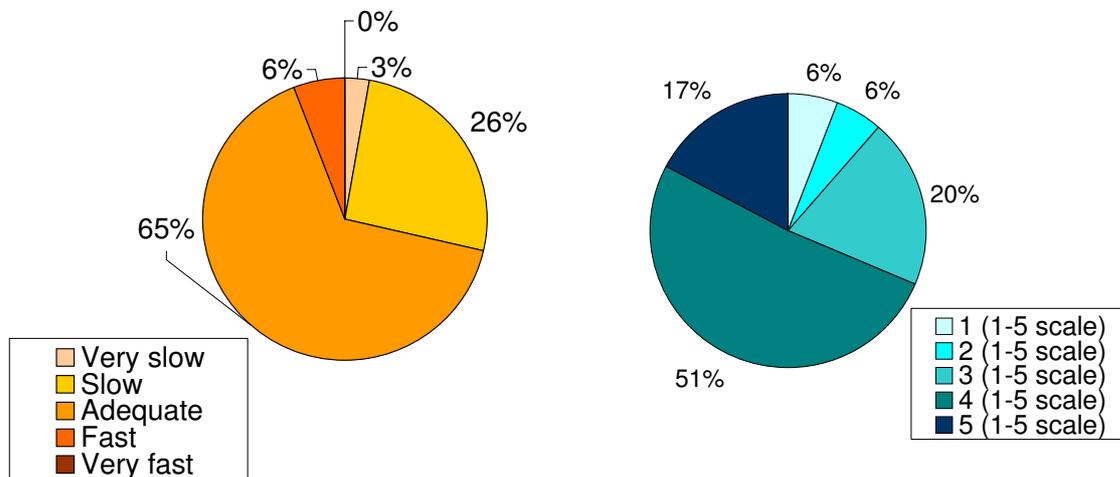


Figure 2.5. *Perceived interaction speed (left) and user satisfaction (right) using UAH*

It was found that, even when the objective evaluation measures indicate that the system delay is negligible, users sometimes are not of the same opinion. Thus, users may consider the interaction slow if the initiative is system-directed, even when the system response time is small. The reason for this is that they feel that the interaction could be faster if they had the opportunity to directly utter their queries.

2.4 The UAH speech corpus

The corpus used for the experiments described in this Thesis is comprised of 85 dialogues by 60 different users interacting with UAH. It contains 422 user turns, with an average of 5 user turns per dialogue. The recorded material has a duration of 150 minutes. It was semi-automatically annotated with de-facto standard evaluation criteria and the users' opinions were obtained with questionnaires as described in Chapter 5. Each user utterance was subsequently tagged with an emotional state by nine non-expert annotators as described in Chapter 3. The size of the corpus is similar to that of other real emotional speech corpora such as those used by Forbes-Riley and Litman (2004a) (10 dialogues, 453 turns) or Morrison et al. (2007) (391 user turns).

2.5 Conclusions

UAH is a spoken dialogue system developed in VoiceXML standard language. It uses dynamic dialogue management and introduces several novel techniques that make the interaction smoother. It was developed to cover different interaction initiatives and confirmation strategies in order to be able to compare their suitability. The introduction of the GAG module is noteworthy. This provides a new way of creating rules for speech recognition grammars without introducing delays and keeps the vocabulary updated. The UAH system has been functional since 2005 when it was made available to the public. Since then, the interactions are being recorded and the main interaction parameters are automatically processed to compute evaluation criteria. The dialogues corresponding to a year of usage of the UAH system were also annotated to introduce new parameters and the opinions of the users about the system have been surveyed. With this information a speech corpus was created which formed the basis for the research described in the next chapters.

*Innegable, señor.
Es indisimulable.
¿Está usted aburrido?
Me parece que está usted aburrido.
Dígame, ¿adónde va tan aburrido?*

Rafael Alberti, El aburrimiento

3

Recognition of non-acted emotions

3.1 Introduction

One of the main research objectives of dialogue systems is to achieve human-like communication between people and machines. This eliminates the need for keyboard and mouse in favour of more intuitive ways of interaction, such as natural language, thus leading to a new paradigm in which technologies can be accessed by non-expert users or handicapped people.

However, multimodal human-computer interaction is still not comparable to human dialogue. One of the reasons for this is that human interaction involves exchanging not only explicit content, but also implicit information about the affective state of the interlocutor. Systems that make use of such information are described as incorporating “affective computing” or “emotion intelligence”, which covers the areas of emotion recognition, interpretation, management and generation.

Due to its benefits and huge variety of applications, affective computing has become an outstanding research topic in the field of HCI, and numerous important international and interdisciplinary related projects have appeared. Some of the latest are MEGA (Camurri et al., 2004), NECA (Gebhard et al., 2004), VICTEC (Hall et al., 2005), NICE (Corradini et al., 2005), HUMAINE (Cowie and Schröder, 2005) and COMPANIONS (Wilks, 2006), to mention just a few.

Accurate annotation is a first step towards optimized detection and management of emotions, which is a very important task in order to avoid

significant problems in communication, such as misunderstandings and user dissatisfaction, that lead to very low task completion rates. Despite its benefits, the annotation of emotions in spoken dialogue systems encounters restrictions as a result of certain important problems. Firstly, as shown in Section 3.3.2, the percentage of neutral vs. emotive speech is usually very unbalanced (Forbes-Riley and Litman, 2004a; Morrison et al., 2007). Secondly, all information must be gathered through the oral modality and in some systems where the dialogue is less flexible, the length of the user utterances can be insufficient to enable other knowledge sources like linguistic information to be employed.

To solve these problems, we propose to use contextual information for the annotation of user emotions in spoken dialogue systems. The main interest is to recognize negative emotions as some studies, for example (Riccardi and Hakkani-Tür, 2005), have shown that once the user is in a negative emotional state, it is difficult to guide him out. Furthermore, these bad experiences can also discourage users from employing the system again. Concretely, three negative emotions are taken into account. The first is *doubtful*, which is useful to identify when the user is uncertain about what to do next. A user is in this emotional state when he has doubts about what to say in that turn. The second and third emotions are *angry* and *bored*, two negative emotional states that must be recognized before the user gets too frustrated because of system malfunctions. In the activation-evaluation space (Russell, 1980; Scherer, 2005), *angry* corresponds to an active negative emotion, whereas *bored* and *doubtful* to passive negative emotions.

Different approaches are presented in this chapter in order to include contextual information in both human annotation (as discussed in Section 3.3) and machine learned classification (Section 3.4). In human annotation, non-expert annotators were provided with contextual information by giving them the utterances to be annotated along with the dialogues where these were produced. In this way, annotators had information about the user speaking style and the moment of the dialogue at which each sentence was uttered. For machine-learned classification, a novel method in two steps is introduced, which enhances negative emotion classification with automatically generated context information. The first step calculates users' neutral speaking style, which is employed to classify emotions into *angry* and *doubtful* OR *bored*; whereas the second step introduces dialogue context and

allows the distinction between *bored* and *doubtful*. One of the main advantages of the proposed method is that it does not require including additional manually annotated data. Hence, it permits a straightforward automatic integration of context information within emotion recognizers for spoken dialogue systems.

To evaluate the benefits of the proposals, different experiments have been performed over a corpus of real emotions extracted from the interaction of around 60 different users with the UAH spoken dialogue system (Chapter 2). The objective is to demonstrate that the proposed contextual information influences human as well as machine recognition, and that better results can be obtained when context is included using the proposed methods, if compared with recognition based on traditional acoustic features or the baseline classification methods.

The chapter is structured as follows: in Section 3.2 there is an overview of the related work done in the area and the points in which the Thesis makes its main contributions. Section 3.3 presents the human annotation procedure and discusses the corpus facts and the results in terms of emotions annotated and agreement between annotators. In Section 3.4 there is a description of the automatic classification of emotions, and a discussion of the experimental results obtained for it. Section 3.4.5 describes the previous approaches studied to take into account both context sources until the optimal approach was encountered. Finally, Section 3.5 presents a summary of the benefits of the proposed methods and the conclusions extracted from them.

3.2 Related work

Unfortunately there is no consensus about what an emotion is. Psychological and biological studies in this matter have been carried out for centuries and it is not strange to find references to Darwin or Descartes in some recent papers about the topic. In an effort to clarify concepts and underlying the direction of main research lines in the area, Cowie (2000) distinguishes two senses in which the word emotion can be interpreted: the first one is as discrete states (p.e. fear, happiness, anger) which are usually referred to as “full-blown” emotions in the literature; and the second one is as an attribute of certain states, which the author names as “emotional states”. In the HUMAINE project (Humaine emotion-research.net, 2007) they also make this

distinction using other nomenclature: episodic and pervasive emotions. Main research efforts are carried out towards the study of pervasive emotions or “emotional states”; whereas the main objective in the study of “full-blown” emotions is to find a restricted set of categories or emotions. There exists the extended theory that “full-blown” emotions can take only a few forms easily distinguishable from each other. However, making such a simplification is not always suitable as in some applications it can be interesting to study blended, simulated and/or conflictive emotions.

Regardless of the definition of emotion, several ways to represent them can be found in the literature. They can be represented using a discrete set or as points in a continuous space. In the continuous case, emotions are represented by coordinates in a space with a small number of dimensions. The typical approach is the bidimensional activation-evaluation space (Cowie et al., 2001). In the horizontal axe, evaluation deals with the “valence” of emotions, that is, positive or negative evaluations of people, things or events. In the vertical one, activation measures the user disposition to take some action rather than none. “Full-blown” emotions form a circular pattern in the activation-evaluation space which made other authors propose a representation based on angles and distance to the centre. Some tools, for example FEELTRACE (Cowie et al., 2000), have been implemented to give a visual representation of the dynamic progress of emotions inside this circle. Additionally, 3D models can be used to distinguish between emotions that are very near each other in the circle (e.g. fear and anger), the new dimensions used are usually perceived control or inclination to engage. Emotions can also be represented in a structural way, which treats them from a cognitive perspective that describes how users deal with the situation that caused the emotion. Furthermore, following the Ortony and G. L. Clore (1988) theory, emotions can be classified according to the emotion-eliciting situations which can be related to events, actions of agents or aspects of objects. OCC theory is usually employed for emotion synthesis (Zong et al., 2000; de Melo and Paiva, 2005).

Emotionally intelligent systems can put the emphasis on the emotions’ cause or effect. In the first case the focus is in the reasons of the apparition of some emotion, which can be external or internal to the user; the second describes the effects of these characteristics in the listener (Cowie, 2000). The research in these areas is generally recognition driven in the cause-type

case and synthesis driven in the effect-type case. The research described in this chapter is focused on the emotion recognition case.

Emotion recognition can be carried out with invasive and non invasive methods. Invasive methods are based on physiological measures like breathing rate or conductivity of skin (Picard, 1997). One of the most widespread methods consists in measuring the galvanic skin response (GSR) as there is a relationship between the arousal of emotions and changes in GSR (Lee et al., 2005). Some other methods are EMG, which measures facial muscles (Mahlke, 2006), hear rate or more recently the usage of brain images (Critchley et al., 2005). Non invasive methods are usually based on audio and video. On the one hand, audio emotion recognition can be carried out from the acoustic information or from linguistic information. Speech is deeply affected by emotions: acoustic, contour, tone, voice quality and articulation change with different emotions, a comprehensive study of those changes is presented in (Cowie et al., 2001). Language information deals with linguistic changes depending on the emotional state of the user. For this purpose the technique of word emotional salience has gained remarkable attention. This measure represents the frequency of apparition of a word in a given emotional state or category and it is calculated from a corpus of user-system interactions (Lee et al., 2005). On the other hand, video recognition usually pays attention to facial expression, body posture and movements of the hands; a summary of all these features can be found in (Picard and Daily, 2005). Other authors emphasize that emotions are influenced by cultural and social settings and defend an “interactional approach” (Boehner et al., 2007) to be considered along with physiological, audio or video measures.

Emotion recognition has been used in Human Computer Interaction (HCI) systems for several purposes. In some application domains it is necessary to recognize the affective state of the user to adapt the systems to it or even change it. For example, in emergency services (Bickmore and Giorgino, 2004) or intelligent tutors (Ai et al., 2006), it is necessary to know the users’ emotional state to calm them down, or to encourage them in learning activities. However, there are also some applications in which emotion management is not a central aspect, but contributes to the better functioning of the system as a whole. In these systems emotion management can be used to resolve stages of the dialogue that cause negative emotional states, as well as to avoid them and foster positive ones in future interactions. For exam-

ple, Burkhardt et al. (2005) use an anger detector to avoid user frustration during the interaction with their voice portal. Furthermore, emotions are of interest not just for their own sake, but also because they affect the explicit message conveyed during the interaction: they change peoples' voices, facial expressions, gestures, speed of speech, etc. This is usually called "emotional colouring" and can be of great importance for the interpretation of user input. For example, Wahlster (2006) use emotional colouring in the context of the SmartKom system to detect sarcasm and thus tackle false positive sentences.

As explained before, emotion recognition is a key aspect to obtain human-like interaction. This is why it has received a lot of attention for the dialogue systems research community. From applications in with the changes in the users' emotional state are uniquely indicators on when the system is not fulfilling users' expectations, to complicated systems in which emotions are a key-stone, like psychological aid systems; emotion recognition is gaining increasing attention from the research community. This is reflected in the number of international interdisciplinary projects that have treated the topic, some of the latest are:

- SAFIRA - Supporting Affective Interactions for Real-time Applications (The Safira Project - DFKI Page, 2002). 24 months from 2000-05-02 (Completed). Its purpose was the enrichment of applications with an affective dimension to support affective behaviour and control in real-time multi-agent systems interacting with users.
- MEGA - Multisensory Expressive Gesture Applications (MEGA Project, 2001). 36 months from 2000-11-01 (Completed). Its purpose was the modelling and real-time analysis, synthesis, and networked communication of expressive and emotional content in non-verbal interaction (e.g. music or dance) by multi-sensory interfaces, from a multimodal perspective.
- MAGICSTER - Embodied Believable Agents (MagiCster Project Pages, 2007). 39 months from 2000-12-01 (Completed). Its purpose was the design and evaluation of a believable conversational interface agent, which makes use of gaze, facial expression, gesture and body posture as well as speech in a synchronised fashion.

- NECA - A Net Environment for embodied emotional Conversational Agents (NECA project, 2005). 30 months from 2001-10-01 (Completed). Its purpose was the creation of multi-user and multi-agent virtual spaces populated by affective conversational agents able to express themselves through synchronised emotional speech and non-verbal expression.
- ERMIS - Emotionally Rich Man-Machine Interaction Systems (EUROPA - CORDIS: Community Research and Development Information Service, 2006). 36 months from 2002-01-01 (Completed). Its purpose was the development of a prototype system for human computer interaction that can interpret users' attitude or emotional state, in terms of their speech and/or their facial gestures.
- PF-STAR - Preparing future multisensorial interaction research (PF-STAR home page, 2004). 24 months from 2002-10-01 (Completed). Its purpose was the contribution to the field of multisensorial and multilingual communication by providing technological baselines, comparative evaluations, and assessment of prospects of core technologies specially in the topic of technologies for speech-to-speech translation, the detection and expressions of emotional states, and core speech technologies for children.
- VICTEC - Virtual ICT with Empathic Characters (VICTEC in Lynne Hall web page, 2005). 35 months from 2002-03-01 (Completed). Its purpose was the development of a toolkit that supports the creation of believable synthetic characters in a virtual environment who establish credible and empathic relations with children.
- NICE - Natural Interactive Communication for Edutainment (NICE project - Main page, 2007). 36 months from 2002-03-01 (Completed). Its purpose was to foster universal natural interactive access, in particular for children and adolescents, by developing natural, fun and experientially rich communication between humans and embodied historical and literary characters.

- CHIL - Computers in the Human Interaction Loop (CHIL - Computers In the Human Interaction Loop, 2007). 36 months from 2003-12-18 (Completed). Its purpose was the creation of environments in which computers serve humans that focus on interacting with other humans instead of having to attend to and being preoccupied with the machines themselves.
- HUMAINE - Research on Emotions and Human-Machine Interaction (Humaine emotion-research.net, 2007). Since 2003-12-18 (In execution). Its purpose is to lay the foundations for European development of systems that can register, model and influence human emotional and emotion-related states coordinating efforts to come to a shared understanding of the issues involved.
- AMI - Augmented Multi-party Interaction (Augmented Multiparty Interaction Project, 2007). Since 2003-12-18 (In execution). Its purpose is the creation of new multimodal technologies to support human interaction in the context of smart meeting rooms and remote meeting assistants.
- INTREPID - A Virtual Reality Intelligent Multi-sensor Wearable System for Phobias' Treatment (Intrepid Project, 2004). 24 months from 2004-01-01 (Completed). Its purpose was the development of a multi-sensor context-aware wearable system for the treatment of phobias.
- AUBADE - A wearable EMG Augmentation system for robust behavioural understanding (AUBADE, 2005). 34 months from 2004-01-01 (Completed). Its purpose was the development of a wearable platform, to ubiquitously monitor and recognise the emotional state of its users in real time, using signals obtained from their face.
- COSY - Cognitive Systems for Cognitive Assistants (CoSy Home, 2007). 48 months from 2006-09-01 (In execution). Its purpose is the construction of physically instantiated systems that can perceive, understand and interact with their environment, and evolve in order to achieve human-like performance in activities requiring specific knowledge about the context.

- CALLAS - Conveying Effectiveness in Leading-Edge Living Adaptive Systems (Callas - Conveying Affectiveness in Leading-Edge Living Adaptive Systems, 2007). 42 months from 2006-11-01 (In execution). Its purpose is the definition and development of a multimodal architecture including emotional aspects, to support experiments and target new media applications essentially in an ambient intelligence paradigm.
- COMPANIONS - Intelligent, Persistent, Personalised Multimodal Interfaces to the Internet (Companions, 2007). 48 months from 2006-11-01 (In execution). Its purpose is the creation of companions: personalized, conversational interface to the Internet that knows its owner, on a range of platforms, indoor and nomadic, based on integrated high-quality research in multi-modal human-computer interfaces, intelligent agents, and human language technology.

In the area of emotion recognition the great majority of studies¹ focus on studying the appropriateness of different machine learning classifiers (Shafran and Mohri, 2005), such as K-nearest neighbours (Lee and Narayanan, 2005), Hidden Markov Models (Ververidis and Kotropoulos, 2006; Pitterman and Pitterman, 2006), Support Vector Machines (Morrison et al., 2007), Neural Networks (Morrison et al., 2007) or Boosting Algorithms (Liscombe et al., 2005; Forbes-Riley and Litman, 2004a). In addition, important research has been directed towards finding the best features to be used for classification. These features can be categorized at different levels. The lowest level deals with physiological features, which are usually measured with intrusive methods. Some examples are galvanic skin response (Lee et al., 2005), facial muscle movements (Mahlke, 2006) or brain images (Critchley et al., 2005). Acoustic and linguistic levels are more widespread and features like articulation changes (Cowie et al., 2001), statistical measures of acoustic features (Ververidis and Kotropoulos, 2006) or word emotional salience (Lee and Narayanan, 2005) are frequently found in the literature. In addition, visual features like facial expression, body posture and movements of hands have recently been adopted, especially in multimodal systems (Picard and Daily, 2005; Zeng et al., 2006). More recently, some authors like Boehner et al. (2007) have proposed cultural information as an additional source of information for detecting emotional states.

¹For further information, see Ververidis and Kotropoulos (2006).

However, less attention is being paid to the training process of the algorithms in which automatic emotion classification is based, and for which emotional annotated corpora are needed. A good annotation scheme is essential as it affects the rest of the stages in the learning process. Besides, manual annotation of corpora is very difficult, time-consuming and expensive, and thus must be carefully designed. Authors that study emotional corpora are mainly interested in how it is gathered, especially comparing acted vs. real emotions acquisition (Morrison et al., 2007), but less work has been done in how the annotation of such a corpus must be achieved. Among others, Devillers et al. (2005) have proposed guidelines to design and develop successful annotation schemes in terms of labels, segmentation rules and validation processes. Gut and Bayerl (2004) have also worked on reliability measures of human annotations, whereas Craggs and Wood (2003) have proposed several layers of emotion annotation.

This Thesis goes a step further and studies how to add contextual information to the corpus annotation process, and suggest the inclusion of two new context sources: users' neutral speaking style and dialogue history. The former provides information about how users talk when they are not conveying any emotion, which can lead to a better recognition of users' non-neutral emotional states (Section 3.4.2). The latter involves using information about the current dialogue state in terms of dialogue length and number of confirmations and repetitions (as will be discussed in Section 3.4.3), which gives a reliable indication of the users' emotional state at each moment. For example, the user is likely to be angry if he has to repeat the same piece of information in numerous consecutive turns in the dialogue.

In the literature there are three main approaches for collecting emotional speech corpora: recording spontaneous emotional speech, recording induced emotions, and using actors to simulate the emotions. As shown in Figure 3.1, in these approaches there is a compromise between naturalness of the emotions and control over the collected data: the more control over the generated data, the less spontaneity and naturalness of the expressed emotion, and vice versa. Therefore, spontaneous emotional speech, which reflects completely natural emotional speech production in the application domain of the emotion recognizer, is the most realistic approach. However, a lot of effort is necessary for the annotation of the corpus, as it requires an interpretation of which emotion is being expressed in each recording. Sometimes,

the corpus is recorded from human-to-human interaction in the application context (Forbes-Riley and Litman, 2004a). In these cases, the result is also natural but it is not directly applicable to the case in which humans interact with a machine. At the other extreme, acted emotional speech is easier to manipulate and avoids the need for annotation, as emotions conveyed in each recording are known beforehand. The results obtained with acted speech are highly dependent on the skills of the actors, therefore the best results are obtained with actors with good drama preparation. When non-expert actors are used, another phase is necessary to discard the recordings that fail to reproduce the required emotion appropriately. In a middle point are the induced emotions, which can be more natural, like the ones elicited when playing computer games (Johnstone, 1996), or easier to manipulate like the ones induced by making people read texts that relate to specific emotions (Stibbard, 2000).

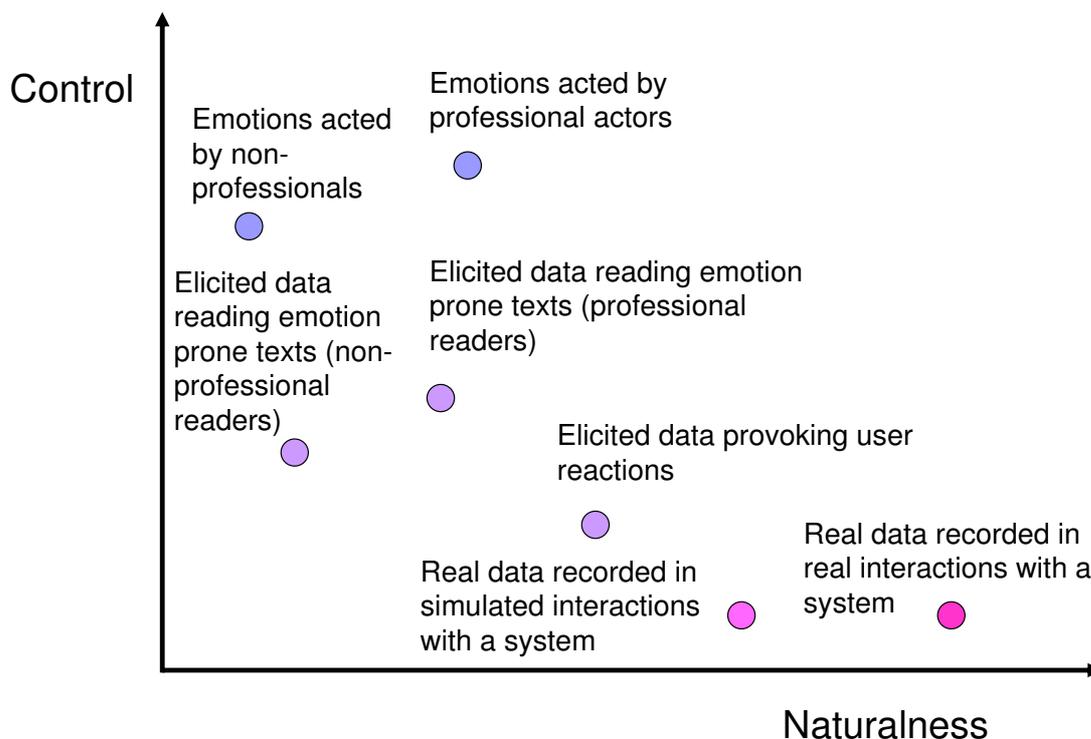


Figure 3.1. *Naturalness vs. control in the main emotional corpora generation approaches*

As some authors have indicated, e.g. Douglas-Cowie et al. (2003), the relationship between acted data and spontaneous emotional speech is not exactly known. But, as stated by Johnstone (1996), even professionally acted speech loses realism as there are some effects that cannot be controlled consciously. Thus, different studies have shown that it is not appropriate to use acted data to recognize naturally occurring emotions (Vogt and André, 2005; Wilting et al., 2006).

As the objective is to build an emotion recognizer for the UAH dialogue system (Chapter 2), and it would have to work with natural emotions occurring in real time, an utterance corpus collected from real users interacting with the system was used in the experiments. This is an important contribution to the state of the art as real non-elicited emotions are difficult to find in Spanish corpora. For example, out of the 70 corpora studied by Douglas-Cowie et al. (2003) and Ververidis and Kotropoulos (2006), only three are in Spanish: González (1999), Montero et al. (1999) and Iriondo et al. (2000). As shown in Table 3.1, two of these are used for emotion synthesis instead of recognition. The table also sets out Spanish corpora employed for emotion feature studies (Adell et al., 2005) and general purpose studies (Hozjan et al., 2002). None of these corpora were collected from real user interactions and the maximum number of actors used was 8, whereas the UAH corpus was collected from 60 users and real emotional speech.

Reference	Actors/users	Purpose	Kind
González (1999)	-	Recognition	Elicited
Montero et al. (1999)	1 actor	Synthesis	Simulated
Iriondo et al. (2000)	8 actors	Synthesis	Simulated
Hozjan et al. (2002)	2 actors	Study, synthesis and recognition	Simulated
Adell et al. (2005)	1 actress, 1 professional reader, 1 member of Spanish Parliament	Emotion features study	2 simulated, 1 natural
UAH corpus	60 UAH system users	Recognition	Natural

Table 3.1. Summary of Spanish emotional speech corpora

3.3 Human annotation of the UAH corpus

The annotation of emotions is a highly subjective task, given that for the same utterance, different annotators may perceive different emotions. The most reliable way to obtain rigorous annotations is to recruit specialized annotators, for example psychologists who are trained to recognize human emotions. Unfortunately, in most cases expert annotators are difficult to find and thus the annotation must be done by non-experts. In this case, all annotators were non-expert as they had not received any specific training on emotion recognition.

To get the best possible annotation employing non-expert annotators, the labelling process must be rigorously designed. Vidrascu and Devillers (2005) suggest several phases to decide the list of labels and annotation scheme, segmentation rules, number of annotators, validation procedures, and consistency study.

The first step is to decide the labels to be used for annotation. Our main interest is to study negative emotional states of the users, mainly to detect frustration because of system malfunctions. Thus, classification is made between the three major negative emotions encountered in the UAH corpus, namely *angry*, *bored* and *doubtful*. For the human annotation of the corpus a fourth category has been used: *neutral*, which represents a non-negative emotional state (i.e. positive emotions such as happiness are also treated as *neutral*). The neutral category was used only for the human annotation of the corpus. The rest of the experiments will focus exclusively on the distinction between the negative emotions considered.

A decision was made to use an odd, high number of annotators - nine, which is more than is typically reported in previous studies, e.g. Forbes-Riley and Litman (2004a) and Lee and Narayanan (2005). Regarding the “segment length”, in this study this is the whole utterance because it was not useful to employ smaller segmentation units (i.e. words). The reason is that our goal was to analyse the emotion as a whole response to a system prompt, without considering the possible emotional changes within an utterance.

In the proposed annotation procedure the corpus was annotated twice by every annotator, firstly in an ordered style and secondly in an unordered style. In the first mode the annotators had information about the dialogue context and the users’ speaking style. In the second case the annotators

did not have this information, so their annotations were based only on the acoustic information of the current utterance.

The final emotion assigned to each utterance in the ordered and unordered schemes was the one annotated by a majority of annotators in each of them. Gold standard emotions for the whole corpus were then computed from the results of each of the schemes. In situations where there was no majority of an emotion above the others (e.g. 4 *neutral*, 4 *bored* and 1 *doubtful*), priority was given to the non-neutral ones (in the last example *bored*). If this conflict was between two non-neutral emotions (e.g. 4 *doubtful*, 4 *bored* and 1 *neutral*), the results were compared between both annotation schemes to choose the emotion annotated by majority among the 18 annotations (the 9 of the ordered and the 9 of the unordered scheme).

3.3.1. Calculation of the agreement between annotators

Several Kappa coefficients were used to study the degree of inter-annotator agreement for both annotation styles (for the ordered and unordered case). Kappa coefficients are based on the idea of rating the proportion of pairs of annotators in agreement (P_o) with the expected proportion of pairs of annotators that agree by chance (P_c). Thus obtaining a proportion of the agreement actually achieved beyond chance ($P_o - P_c$) with all possible agreements that are not by chance ($1 - P_c$):

$$\kappa = \frac{P_o - P_c}{1 - P_c} \quad (3.1)$$

For the Thesis four different Kappa coefficients were used (Figure 3.2) with which two main issues have been studied: i) the impact of annotator bias, that is, given a fixed number of agreements, the effect that the distribution of disagreements between categories has in the Kappa value; and ii) the level of importance of all possible disagreements in our task, i.e. disagreement between emotions which are easily distinguishable should have a more negative impact in the Kappa coefficient than disagreements in very different categories.

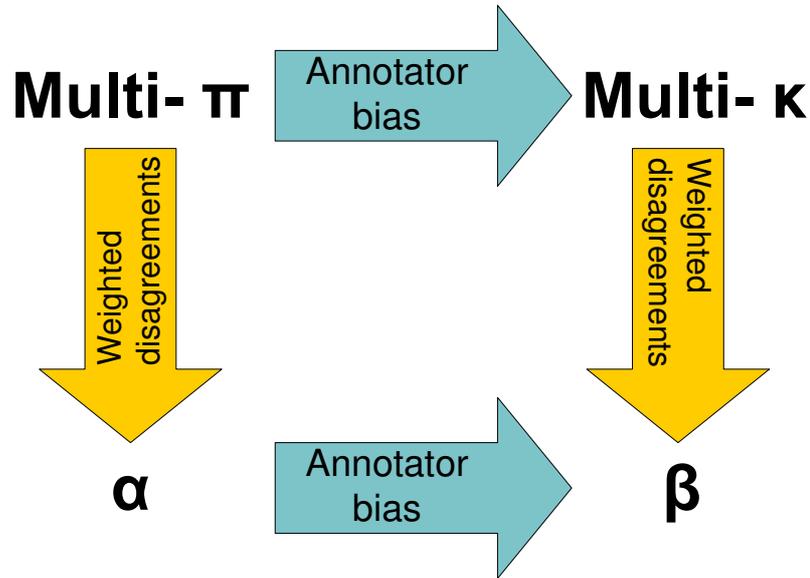


Figure 3.2. *Kappa coefficients used in the experiments*

The simplest Kappa coefficient used was proposed by Fleiss (1971), as a generalization for multiple annotators of the two-coders Scott’s π (Scott, 1955). There has been much confusion in the literature about Fleiss’ Kappa, as many authors have reported it as a generalization of Cohen’s κ instead (Cohen, 1960). This is further discussed by Artstein and Poesio (2005), who made a considerable effort to clarify the definitions of the different Kappa coefficients. In order to avoid inconsistencies their notation is followed for all the Kappa coefficients employed in the chapter. In particular, Fleiss’ Kappa has been noted as multi- π .

The calculation of multi- π is based on Equation 3.1, where the observed agreement (P_o) is computed as the number of cases in which two different annotators agreed to annotate a particular utterance with the same emotion:

$$P_o = \frac{1}{UA(A-1)} \sum_{u=1}^U \sum_{e=1}^E n_{ue}(n_{ue} - 1) \quad (3.2)$$

In Equation 3.2, U is the number of utterances to be annotated, A the number of annotators, E the number of emotions, and n_{ue} the number of times the utterance ‘u’ was annotated with emotion ‘e’.

Fleiss assumes that all annotators share the same probability distribution. This means that the probability that an annotator classifies an utterance ‘u’ with a particular emotion ‘e’, can be computed as the overall probability of annotating ‘u’ as ‘e’. This global probability was computed as the total number of assignments to emotion ‘e’ made by all annotators (n_e in Equation 3.3) divided by the total number of assignments ($U \cdot A$). Chance agreement (Equation 3.3) was then computed as the probability that any pair of coders annotated the same utterance with the same emotion, which was assumed to be the joint probability of each of them making such assignment independently, as annotators judged all utterances independently from each others.

$$P_c^\pi = \sum_{e=1}^E \left(\frac{1}{UA} n_e \right)^2 \quad (3.3)$$

The calculation of multi- π assumes that each annotator follows the same overall distribution of utterances into emotions. However, such a simplification may not be plausible in all domains due to the effect of the so-called *annotator bias* in the Kappa value. In our experiments, the annotator bias can be defined as the extent to which annotators disagree on the proportion of emotions, given a particular number of agreements. With the rest of the parameters fixed, the Kappa value increases as the bias value gets higher, that is, when disagreement proportions are not equal for all emotions and there is a high skew among them. This is the so-called *Kappa second paradox*. Different studies of its impact can be found in the literature, e.g. Feinstein and Cicchetti (1990), Cicchetti and Feinstein (1990), Lantz and Nebenzahl (1996), and Artstein and Poesio (2005).

To study whether inclusion of the different annotating behaviours could improve the Kappa values, Davies and Fleiss (1982) Kappa was calculated, which has been noted as multi- κ , following the study of Artstein and Poesio (2005). As happens with multi- π , the calculation of multi- κ also relies on Equation 3.1, and has the same observed agreement (Equation 3.2). However, for the chance agreement, it includes a separate distribution for each annotator. Thus, in this case the probability that an annotator ‘a’ classifies an utterance ‘u’ with emotion ‘e’ is computed with the observed number of utterances assigned to emotion ‘e’ by that annotator (n_{ae}), divided by the total number of utterances (U). The probability that two annotators agree

in annotating an utterance ‘u’ with emotion ‘e’ is again the joint probability of each annotator doing the annotation independently:

$$P_c^\kappa = \frac{1}{\binom{A}{2}} \sum_{e=1}^E \sum_{j=1}^{A-1} \sum_{k=j+1}^A \frac{n_{a_j e}}{U} \frac{n_{a_k e}}{U} \quad (3.4)$$

Despite of including differences between annotators, multi- κ gives all disagreements the same importance. In practice, all disagreements are not equally probable and do not have the same impact on the quality of the annotation results. For example, in our experiments, a disagreement between *neutral* and *angry* is stronger than between *neutral* and *doubtful*, because the first two categories are more easily distinguishable.

To take all this information into account weighted Kappa coefficients have been used (Cohen, 1968; Fleiss and Cohen, 1973), which put the emphasis on disagreements instead of agreements². Their calculation is based on Equation 3.5 (equivalent to Equation 3.1):

$$\kappa_w = 1 - \frac{\overline{P}_o}{\overline{P}_c} \quad (3.5)$$

where \overline{P}_o indicates observed disagreement, and \overline{P}_c disagreement by chance. For all the coefficients used, the observed disagreement has been calculated as the number of times each utterance ‘u’ was annotated with two different emotions e_j and e_k by every pair of annotators, weighted by the distance between the emotions:

$$\overline{P}_o = \frac{1}{UA(A-1)} \sum_{u=1}^U \sum_{j=1}^{E-1} \sum_{k=j+1}^E n_{ue_j} n_{ue_k} \text{distance}(e_j, e_k) \quad (3.6)$$

Consequently, the computation of the weighted coefficients implies employing distance metrics between the four emotions used for annotation (*neutral*, *angry*, *bored* and *doubtful*). To do so, the discrete list of emotions have been arranged within a continuous space, using the bidimensional activation-evaluation space (Russell, 1980). Emotions form a circular pattern

²An alternative calculation based on agreements can be found in (Sim and Wright, 2005)

in this space. This is why other authors proposed a representation based on angles and distance to the centre. Taking advantage of this circular disposition, angular distances between the emotions studied have been used for the calculation of the weighted Kappa coefficients. Instead of establishing our own placement of the emotions in the space, some already established angular dispositions were employed to avoid introducing measurement errors. With this purpose, the list of 40 emotions with their respective angles proposed by Plutchik (1980) was employed, which has been widely accepted and used by the scientific community. In this list, *bored* (136.0°) and *angry* (212.0°) were explicitly contemplated, but this was not the case for *doubtful*. The most similar emotions found were “uncertain”, “bewildered” and “confused”, which only differentiated in 2° in the circle. “Uncertain” (139.3°) was chosen because it was the one that better reflected the emotion wanted to be annotated. However, other authors like Scherer (2005) have explicitly considered *doubtful* as an emotional state. Plutchik (1980) did not reflect neutral in his list as it really is not an emotion but the absence of emotion. Instead, he used a state called “accepting” as the starting point of the circle (0°), which was used as *neutral* in our experiments.

The distance between the four emotions was calculated in degrees with the angle that each of them formed in the circle. The smallest angle between the emotions being considered (x or 360-x) was always chosen. This way, the distance between every two angles was always between 0 and 180 degrees. For the calculation of the Kappa coefficients, distances were converted into weights with values between 0 and 1. A 0 weight (which corresponds to 0° distance in the proposed approach) implies annotating the same emotion, and thus having no disagreement. On the contrary, weight=1 (180° distance) corresponds to completely opposite annotations and thus maximum disagreement. The resulting distances and weights are listed in Table 3.2.

Angle/ Weight	Neutral	Angry	Bored	Doubtful
Neutral	0.00° / 0.00	148.00° / 0.82	136.00° / 0.75	139.30° / 0.77
Angry	148.00° / 0.82	0.00° / 0.00	76.00° / 0.42	72.70° / 0.40
Bored	136.00° / 0.75	76.00° / 0.42	0.00° / 0.00	3.30° / 0.02
Doubtful	139.30° / 0.77	72.70° / 0.40	3.30° / 0.02	0° / 0.00

Table 3.2. Distance between the emotions considered

There is not a consensus in the scientific community about the properties of the distance measures. However, Artstein and Poesio (2005) propose some constraints: the distance between a category and itself should be minimal and the distance between two categories should not depend on the order (i.e. distance from A to B should be equal to distance from B to A). As can be observed by the symmetry of the table, our distance measures and weights follow these restrictions:

- The angle an emotion forms with itself is 0°

$$\forall e \in E, distance(e, e) = 0$$

- The angle between emotion A and emotion B is the same in both directions (as it was established to choose the minimal angle):

$$\forall e_A, e_B \in E, distance(e_A, e_B) = distance(e_B, e_A)$$

As can be observed in Table 3.2, the highest distances were between non-neutrals and neutral. Thus, when calculating weighted Kappa coefficients, disagreements in which an annotator judged an utterance as neutral and the other as non-neutral were given more importance than for example an *angry* vs. *bored* disagreement.

Two weighted Kappa coefficients were calculated: Krippendorff's α (Krippendorff, 2003) and Artstein and Poesio's β . Both of them shared the same observed disagreement calculation (Equation 3.5), for α disagreement by chance was:

$$\overline{P}_c^\alpha = \frac{1}{UA(UA - 1)} \sum_{j=1}^{E-1} \sum_{k=j+1}^E n_{e_j} n_{e_k} distance(e_j, e_k) \quad (3.7)$$

As can be observed in Equation 3.7, this coefficient does not consider annotator bias. This was solved for the β coefficient, with which also the observed behaviour of each annotator was measured:

$$\bar{P}_c^\beta = \sum_{j=1}^{E-1} \sum_{k=j+1}^E \left[\frac{1}{U^2 \binom{A}{2}} \sum_{m=1}^{A-1} \sum_{n=m+1}^A n_{a_m e_j} n_{a_n e_k} \text{distance}(e_j, e_k) \right] \quad (3.8)$$

The results for each described coefficient are listed in Table 3.3 and discussed in the next section.

Coefficient	Unordered	Ordered
multi- π	0.3256	0.3241
multi- κ	0.3355	0.3256
α	0.3382	0.3220
β	0.3393	0.3237

Table 3.3. Values of the Kappa coefficients for unordered and ordered annotation schemes

3.3.2. Discussion of human annotation results

As previously commented, one of the difficulties of emotion recognition in spoken dialogue systems is that in most application domains the corpora obtained are very unbalanced, because there is usually a higher proportion of neutral than emotional utterances (Forbes-Riley and Litman, 2004a; Morrison et al., 2007). This is in accordance with our experimental results since, on average among the nine annotators, more than 85.00% of utterances were annotated as *neutral*. It was also observed that this proportion is affected in 3.40% of the cases by the annotation style. Concretely, for the ordered annotation, 87.28% were tagged as *neutral*, whereas for the unordered annotation the corpus was even more unbalanced: 90.68% of the utterances were annotated as *neutral*. Figure 3.3 shows the proportion of non-neutral emotions tagged by the 9 annotators. As can be observed, the ordered annotation style yielded a greater percentage for the *bored* category: 39.00% more than in the

unordered style. The figure also shows that the *angry* category is substantially affected by the annotation style (i.e. ordered vs. unordered): 70.58% more *angry* annotations were found in the ordered annotation style. On the contrary, the *doubtful* category is virtually independent of the annotation style: only 2.75% more doubts were found in the unordered annotation.

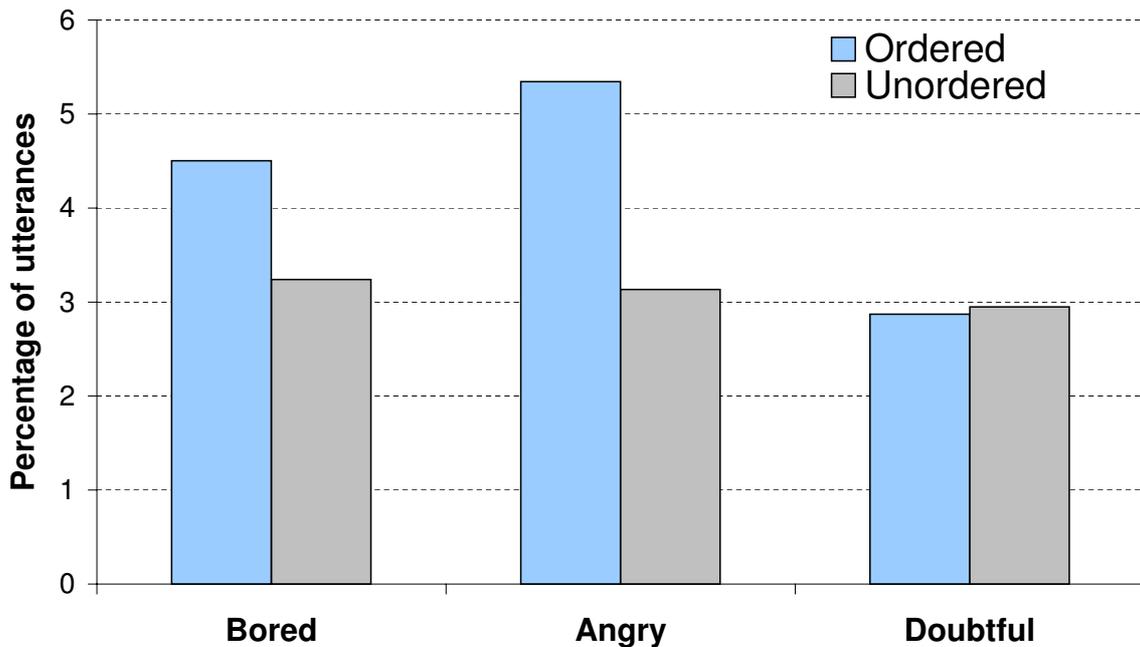


Figure 3.3. Proportion of non-neutral annotated utterances

A plausible reason for these results is that the incorporation of context in the ordered case influences the annotators in assigning the utterances belonging to the same dialogues into the same emotional categories. This way, there are no very noticeable transitions between consecutive utterances. For example, if anger is detected in one utterance then the next one is probably also annotated as *angry*. Besides, the context allows the annotators to have information about user’s speaking style and the interaction history. In contrast, in the unordered case the annotators only have information about the current utterance. Hence, sometimes they cannot tell whether the user is either angry or he normally speaks loudly and fast.

Thus, it is an important fact to be taken into account when annotation is carried out by non-expert annotators, which is the most common, cheapest and least time consuming method. In addition, when listening to the corpus in order, the annotators had information about the position of the current

user turn within the whole dialogue, which also gives a reliable clue to the user's state. For example, a user is more likely to get bored after a long dialogue, or to become angry after many confirmation requests.

As can be observed in Table 3.3, the values of the different Kappa coefficients also vary slightly depending on the annotation scheme used. In the unordered case, both taking into account annotator bias (multi- κ vs. multi- π , and β vs. α), and weighting disagreements (β and α vs. multi- κ) improves the agreement values. However, in the ordered case only taking into account annotator bias enhances the agreement values, whereas weighting the disagreements reduces Kappa. This is a consequence of the increment of non-neutral annotations already discussed. Taking into account that the great majority of agreements occur when annotators tag the same utterance as neutral (as can be observed in Figure 3.4), an increment in the number of emotions annotated as non-neutral provokes more discrepancies among the annotators and thus reduces the Kappa value.

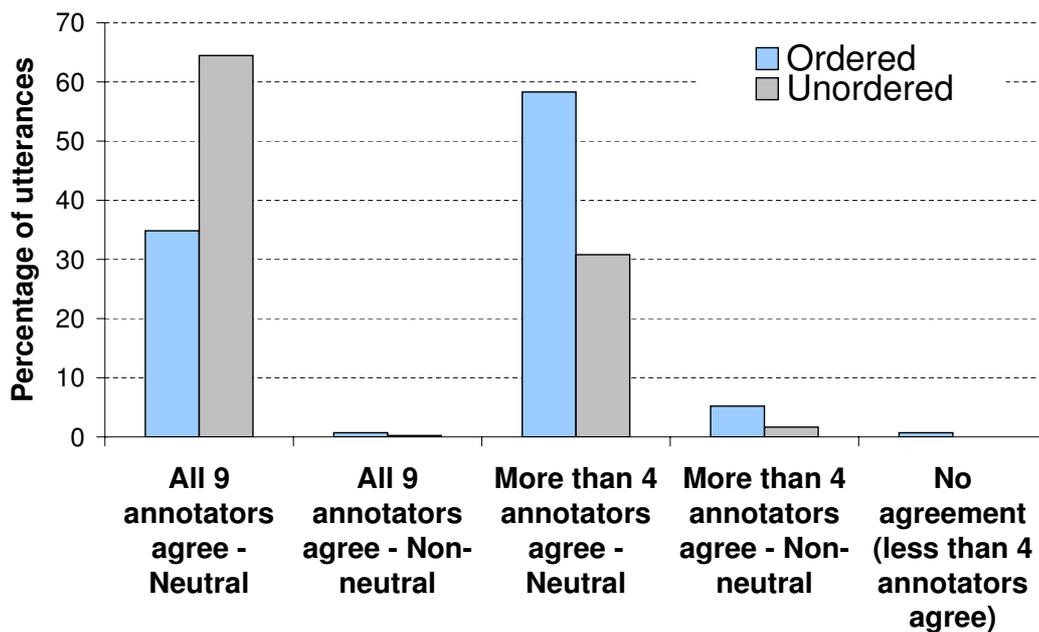


Figure 3.4. Percentage of utterances in which annotators agree

Furthermore, as can be observed in Figure 3.5 most of the disagreements occur between neutral and non-neutral categories, which are the emotions with higher distances according to our weighting scheme (Table 3.2), thus provoking weighted Kappas to be lower in the case of the ordered scheme.

To study the effect of annotator bias, pair wise agreement between all annotators was measured. As can be observed in Figure 3.5 there were no annotators who had a significantly poor agreement with the rest. However, when the annotation results were examined, it was found that there were remarkable differences between those annotators who were used to the Andalusian dialect³ (in which the utterances were pronounced) and those who were not so accustomed. As previously explained, the corpus was recorded from user interactions with the UAH system. The users were mainly students and professors at the University of Granada, which is in south eastern Spain. The way these users express themselves is influenced by the Eastern Andalusian dialect (Gerfen, 2002; O’Neill, 2005), which although similar to Spanish Castilian has several differences such as a faster rhythm and a lower expiratory strength. In our group of annotators, 6 were used to the Andalusian dialect (annotators 1, 2, 3, 4, 6 and 9 in Figure 3.5) and 3 were not (annotators 5, 7 and 8 in the figure).

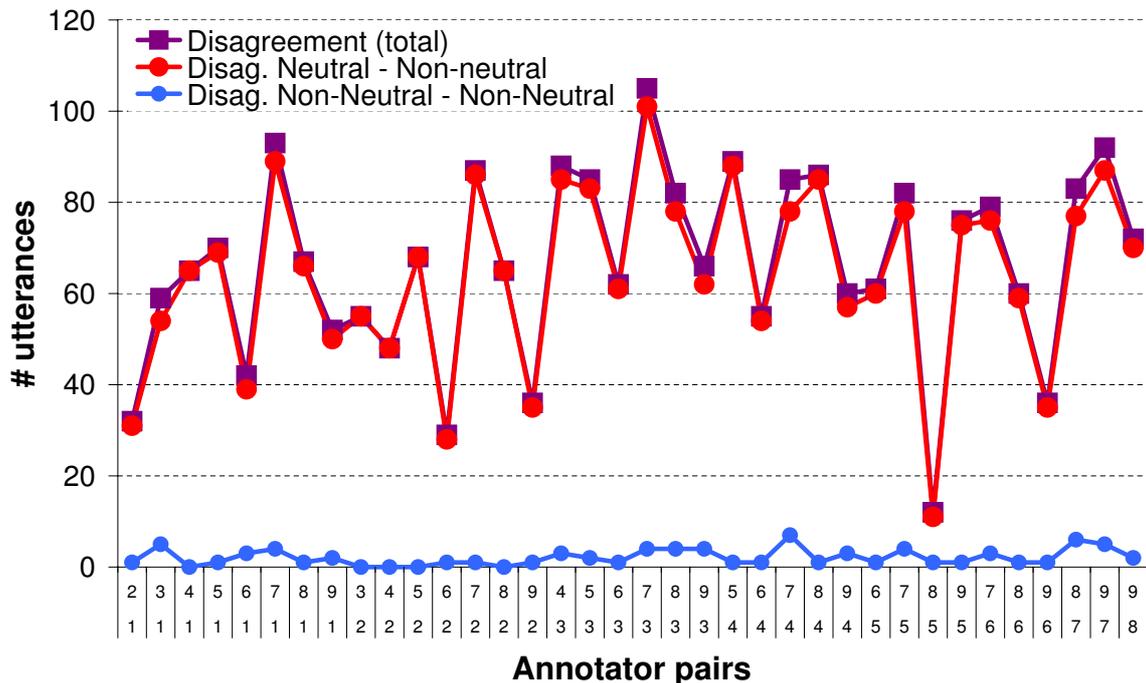


Figure 3.5. Pair wise disagreement between annotators with the ordered scheme

³Spanish spoken in Southern Spain.

Figure 3.6 shows, for the total number of annotations made in each category, which percentage corresponds to each type of annotators. As can be observed, in all the cases but one (especially in those obtained employing the ordered scheme), the annotators not used to the Andalusian dialect marked around 50% of the emotions encountered for the emotional category. This is caused by the confusion of characteristics of the dialect with emotional cues, for example confusing the Andalusian fast rhythm with an indication of anger.

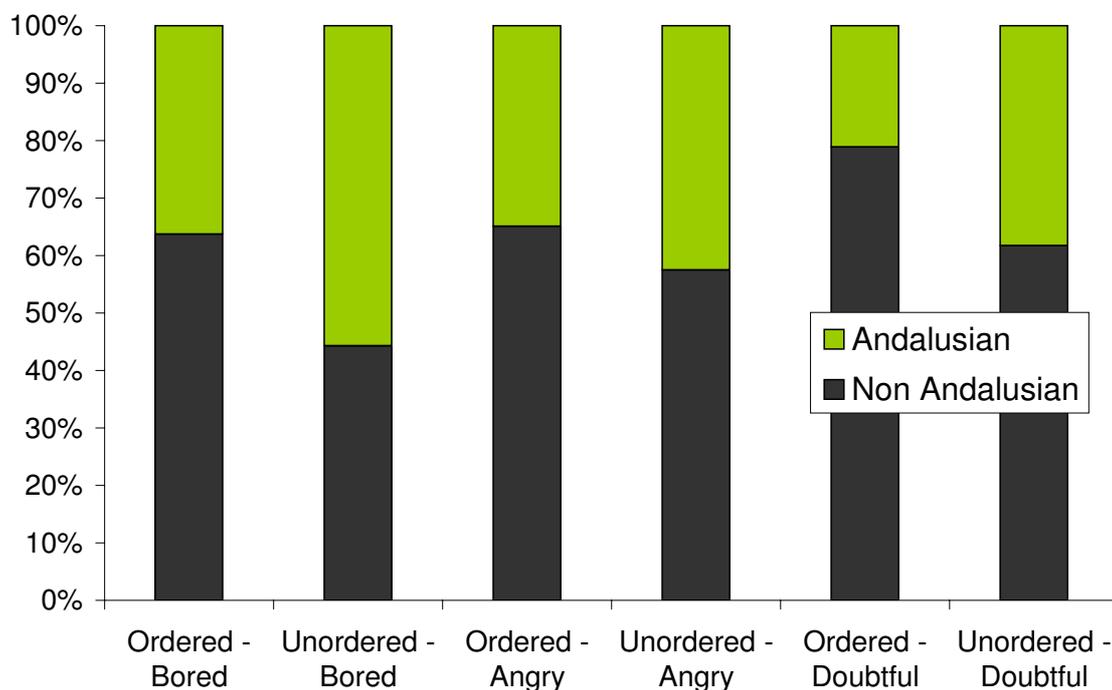


Figure 3.6. *Proportion of annotated emotions depending on dialect*

The effect on the annotation schemes for both kinds of annotators was studied and the results obtained are shown in Table 3.4. As can be observed, the annotators used to the Andalusian dialect obtained Kappa values for both annotation schemes which were more similar (ranging between 0.3234 to 0.3621). For these annotators, the Kappa values were smaller for the ordered scheme because there were fewer utterances annotated as neutral.

	Andalusian annotators		Non-andalusian annotators	
	Unordered	Ordered	Unordered	Ordered
multi- π	0.3608	0.3234	0.3734	0.5593
multi- κ	0.3621	0.3275	0.3746	0.5598
α	0.3595	0.3248	0.3644	0.5691
β	0.3607	0.3265	0.3703	0.5697

Table 3.4. *Kappa values for the different annotator types*

On the contrary, annotators not used to the Andalusian dialect had very different Kappa values depending on the annotation scheme used: in the ordered case values ranged from 0.5593 to 0.5697 whereas in the unordered these were between 0.3639 and 0.3746. This is due to a big decrement of the chance agreement. As shown in Figure 3.7, the observed agreement was more or less constant, whereas the chance agreement decreased in the ordered scheme.

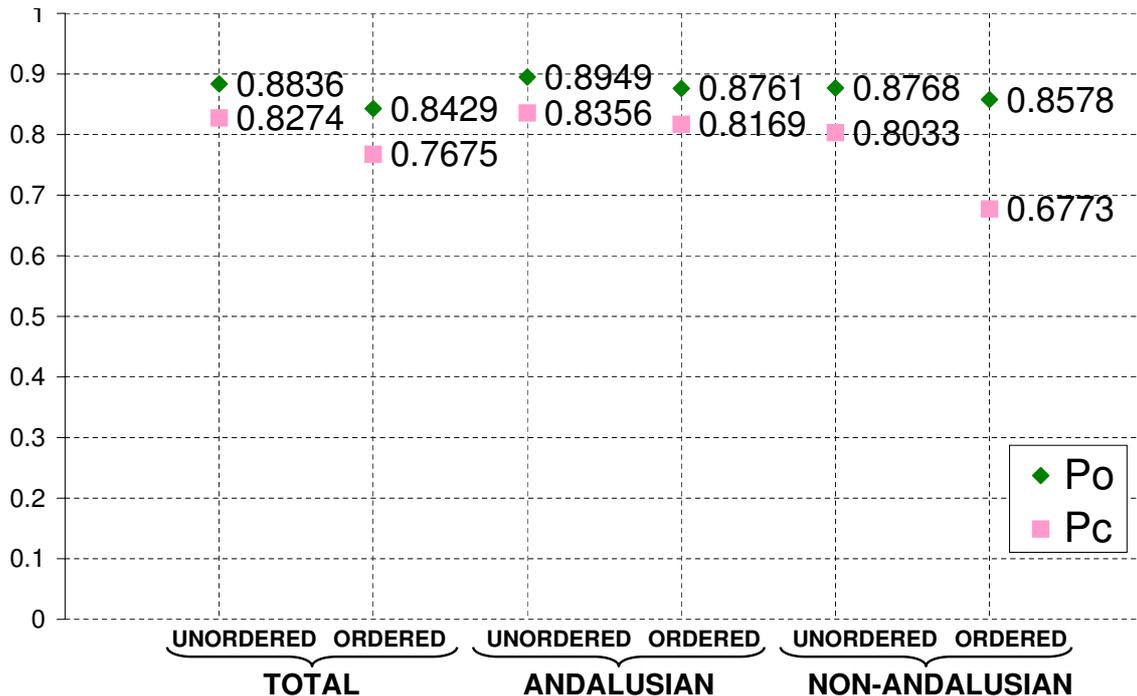


Figure 3.7. *Relative values of agreement by chance and observed agreement for multi- κ*

The most likely reason for this is the decrement in the number of neutrals annotated by annotators not used to Andalusian. This happens for both annotation schemes, but the number of neutrals annotated is higher in the unordered one, and that is why results are more similar to those obtained by Andalusian annotators with the unordered annotation scheme. Even though the number of non-neutral annotations increased proportionally with the decrement of neutrals, the unbalancement of the corpus made the probability of agreeing by chance in the neutral emotion more important in the computation of the overall agreement by chance. For example, in the case of multi- κ , agreement by chance (P_c) was calculated as the sum of agreeing by chance in each emotion ($P_c = P_c^{neutral} + P_c^{bored} + P_c^{angry} + P_c^{doubtful}$). The values for agreeing by chance when annotators not used to Andalusian used the ordered scheme were:

- $P_c^{neutral} = 0.6645$,
- $P_c^{bored} = 0.0052$,
- $P_c^{angry} = 0.0069$ and
- $P_c^{doubtful} = 0.0008$.

For the rest of annotators these values were:

- $P_c^{neutral} = 0.8137$,
- $P_c^{bored} = 0.0010$,
- $P_c^{angry} = 0.0014$ and
- $P_c^{doubtful} = 0.0008$.

Thus, $P_c^{neutral}$ was the determining factor in obtaining the global P_c .

The situation in which although having an almost identical number of agreements, the distribution of these across the different annotation categories deeply affects Kappa, is typically known as the *first Kappa paradox*. This phenomenon establishes that other things being equal, Kappa increases with more symmetrical distributions of agreement. That is, if the prevalence of a category compared with the others is very high, then the agreement by

chance (P_c) is also high and the Kappa is considerably decremented (Feinstein and Cicchetti, 1990; Cicchetti and Feinstein, 1990).

As already reported by other authors, e.g. Feinstein and Cicchetti (1990), the first Kappa paradox can drastically affect Kappa values and thus must be considered in its interpretation. There is not an unique and generally accepted interpretation of the Kappa values. One of the most widely used is the one presented by Landis and Koch (1977), which makes a correspondence between intervals for Kappa values and interpretations of agreement. Following this approach, our experimental results indicate fair agreement for both annotation schemes and with the four different Kappa coefficients. Alternatively, Krippendorff (2003) established 0.65 as a threshold for acceptability of agreement results. Hence, considering this value the 0.3393 highest Kappa obtained would not be acceptable. However, most authors seem to agree in that using a fixed benchmark of Kappa intervals does not provide enough information to make a justified interpretation of acceptability of the agreement results. In order to provide a more complete framework, some authors like Dunn (1989), propose to place Kappa into perspective by reporting *maximum*, *minimum* and *normal* values of Kappa which can be calculated from the observed agreement (P_o) as follows (Lantz and Nebenzahl, 1996):

$$kappa_{max} = \frac{P_o^2}{(1 - P_o)^2 + 1} \quad (3.9)$$

$$kappa_{min} = \frac{P_o - 1}{P_o + 1} \quad (3.10)$$

$$kappa_{nor} = 2P_o - 1 \quad (3.11)$$

The obtained Kappa values (Table 3.3) are shown in Figure 3.8 with their $kappa_{max}$, $kappa_{min}$ and $kappa_{nor}$ values, where *normal* values are marked in italics and the actual values obtained in bold.

As can be observed in the figure, for the same observed agreement, the possible values of Kappa can deeply vary from $kappa_{min}$ to $kappa_{max}$ depending on the balancement of the corpus. $Kappa_{max}$ is obtained when maximally skewing disagreements while maintaining balanced agreements, whereas $kappa_{min}$ is obtained when agreements are skewed and disagreements balanced. $Kappa_{nor}$ does not correspond to an ideal value of Kappa, but rather to symmetrical distributions of both agreements and disagree-

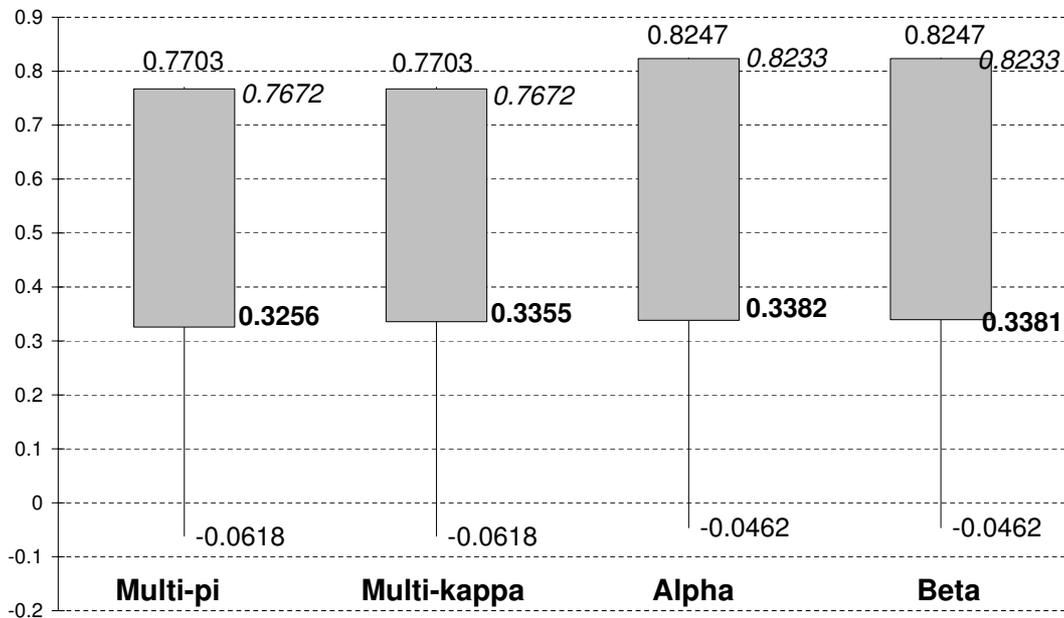
ments. It can be observed in the figure how the displacement between actual and normal values is smaller in the ordered scheme (Figure 3.8(b)). Thus, this scheme does not only allow recognizing more non-neutral emotions, but also obtaining Kappa values which, although smaller than in the unordered scheme in absolute value, are much closer to the *normal* and *maximum* agreement values attainable and further from the *minimum*.

As stated in (Lantz and Nebenzahl, 1996), departures from $kappa_{nor}$ value indicate asymmetry in agreements or disagreements depending in if they are closer to the minimum or maximum value respectively. In Figure 3.8, the shift between the observed and the normal Kappa values is represented with a box. The results corroborate that presenting Kappa values is more informative when they are put into context, as it is obtained a valuable indicator of possible unbalancements that has to be considered to reach appropriate conclusions about reliability of the annotations. For example, in our case there were significant departures from $kappa_{nor}$ in all cases, which corroborates that there was a big asymmetry in the categories. This is due to the prevalence phenomenon previously discussed (first Kappa paradox).

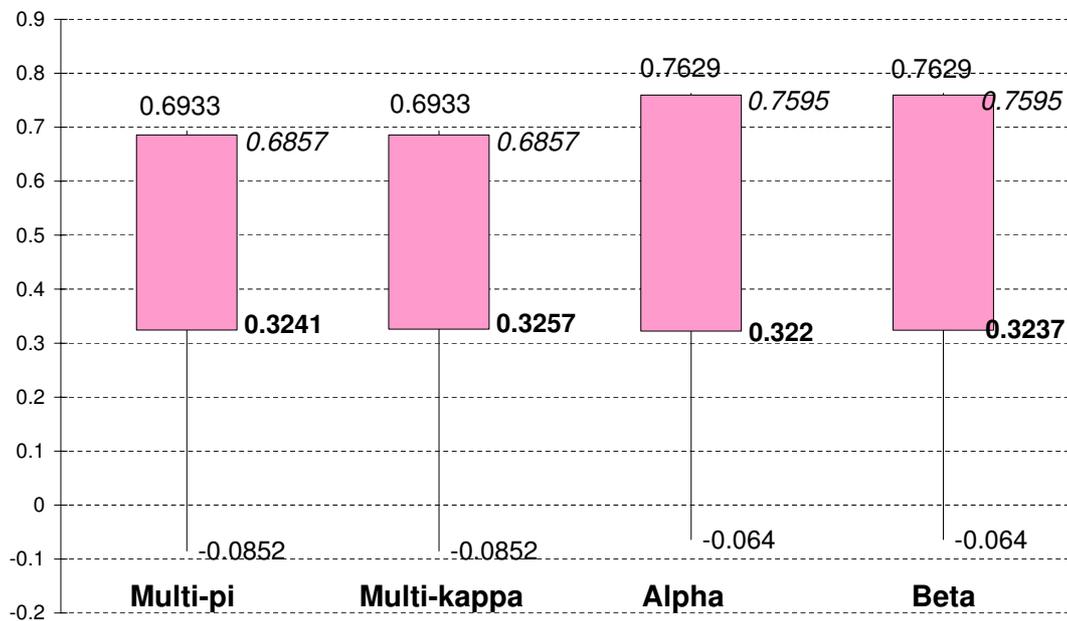
As discussed before, prevalence appeared as an unavoidable consequence of the natural unbalancement of non-acted emotional corpora, where the neutral category is clearly predominant. Thus, approaches based uniquely on already established values of acceptability (Landis and Koch, 1977; Krippendorff, 2003) are not suitable for our application domain. Some authors have already reported additional measures to complement the information provided with the Kappa coefficients. For example, Forbes-Riley and Litman (2004a) report on both observed agreement and Kappa, whereas Lee and Narayanan (2005) report on Kappa along with an hypothesis test.

Although reported Kappa values in emotion recognition employing unbalanced corpora are usually low, e.g. from 0.32 to 0.42 in (Shafran et al., 2003) and below 0.48 in Ang et al. (2002) and Lee and Narayanan (2005), there is not a deep discussion about the problematic of Kappa values in the area, not even in papers explicitly devoted to challenges in emotion annotation, e.g. Devillers et al. (2005). Furthermore, even when other agreement measures are reported along with Kappa, e.g. Forbes-Riley and Litman (2004a) and Lee and Narayanan (2005), there is only one Kappa coefficient calculated (usually multi- π) and no discussion about why there is such a big difference between the Kappa values and the other measures reported. Thus,

the study presented may be one of the first reports about different Kappa values and the issues related to their use and interpretation in annotation of real emotions.



(a) Unordered scheme



(b) Ordered scheme

Figure 3.8. *Kappa maximum, minimum, normal (italics) and observed (bold) values*

Finally, to obtain a more approximate idea about the real level of agreement reached by the nine annotators, the values of observed agreement are reported in Table 3.5, which has been used along with Kappa by other authors in different areas of study (Ang et al., 2002; Forbes-Riley and Litman, 2004a). As can be observed in the table, in all the cases the observed agreement was above 0.85. This measure does not contemplate the effect of prevalence (see Figure 3.7), and thus values were not higher for the annotators not used to the Andalusian dialect in the ordered case.

Annotation scheme	User type	Observed agreement	Weighted observed agreement
Unordered	Total	0.8836	0.9117
	Andalusian	0.8950	0.9197
	Non-andalusian	0.8767	0.9050
Ordered	Total	0.8429	0.8800
	Andalusian	0.8761	0.9049
	Non-andalusian	0.8578	0.895

Table 3.5. Observed agreement values

From all the previous results it is possible to conclude that employing the ordered scheme allowed the annotation of more non-neutral emotions. Unfortunately, this translates in lower Kappa coefficients as most of the agreements occur for neutrals. These low Kappas indicate that multiple annotators should be used for annotating natural emotions to obtain reliable emotional corpora. One possible way to overcome the problem of high chance agreements, consists in maximizing the observed agreement. For example, Litman and Forbes-Riley (2006) propose the usage of “consensus labelling”, i.e. to reach a consensus between annotators until a 100% observed agreement is obtained.

In our case, the computed Kappa values were useful to compare the two annotation schemes. As shown in Figure 3.8, although the Kappa value and observed agreement percentages are lower in the ordered case, it was found that it can be useful to obtain results which are closer to the maximum achievable. It can also be deduced from our study that evaluating the reliability of an emotion annotation process where agreements are so highly skewed, can lead to very low Kappas (Table 3.3) that are far from the high

agreement values observed (Table 3.5). Hence, it was necessary to include other sources of information like observed agreement and minimal, maximal and normal values along with the values obtained for the different Kappa coefficients in order to make meaningful interpretations. Besides, as shown in Table 3.5, giving a weight to the different disagreement types can considerably increment the observed agreement between annotators. A method to compute distances between such disagreements has been presented.

3.4 Automatic classification of the UAH corpus

As shown by the experimental results described in Section 3.3, contextual information is very important for human annotators. Therefore, in this section it is examined whether discrimination between emotions in machine-learned classification is also affected by this factor. The specific interest in this Thesis relies on distinguishing between emotions and thus only non-neutral emotions are input for the learning algorithms. The reason for doing this is not to reduce the effect of the corpus unbalancement, but to carry out a deeper study on the differences between the negative emotions considered. Thus, our main future work guideline will be to add a neutral vs. non-neutral emotion recognizer to build an emotion recognizer that copes with this natural skewness (see Chapter 6). The experiments in this section can be classified into two types: speech-related and dialogue-related. For the first group machine learning has been applied to distinguish the three emotions of interest (*angry*, *bored* and *doubtful*) and have measured the benefits of using the novel approach proposed for acoustic normalization to improve classification. For the second group knowledge about the context of the interaction has been considered in addition.

For comparison purposes, the first approach used is a baseline that always annotates user utterances with the same emotion regardless of the input. In the UAH corpus the most frequent emotion category is *angry*, therefore the baseline annotated each utterance with this label. The second algorithm used is a Multilayer Perceptron (MLP) (Rumelhart et al., 1986; Bishop, 2006). A topology with a hidden layer with $\frac{\text{number of features} + \text{number of emotions}}{2}$ nodes was used. The learning rate, which determines the speed of the search con-

vergence, was set at 0.3 to prevent it being too large (in which case it might miss minimums or oscillate abruptly) or too small (thus provoking slow convergence). To prevent the MLP from over fitting, the passes through the data (epochs) were restricted to 500. In addition, a validation threshold of 0.2 was set to determine the consecutive times that the validation set error could deteriorate before the training was stopped. To improve the performance, a momentum of 0.2 was introduced for the learning of the weights that configured the MLP. To train the MLPs and carry out the experimentation the Weka toolkit (Witten and Frank, 2005) was employed. It is an open source collection of machine learning algorithms for data pre-processing, classification, regression, clustering and visualization.

For training and testing our classifier a 5-fold cross-validation technique has been used. Therefore, the experiments consisted of five trials where the corpus was randomly split into five approximately equal subsets (20% of the corpus each). Every trial used each of the partitions in turn for testing, and the remainder (80% of the corpus) for training, so that after the 5 trials every instance had been used exactly once for testing. Additionally, a tuning partition (20%) was extracted from each training partition in order to make the feature selection. Thus, the evaluation was carried out using two phases. In the first one, the learning algorithms were trained with the 60% of training utterances and evaluated with the 20% employed for tuning. In the second phase, the complete training partition was used for training the MLP, and the test part (20%) for evaluation. For comparison purposes, this second step was carried out, on the one hand, employing all the features of the 80% training utterances, whilst on the other, employing only the features selected in the first step.

Finally, for all the experiments described in this section, the significance of the results was checked using the corrected paired *t*-tester available in the analysis tool of the Weka 3.5.4 Experimenter⁴ (Witten and Frank, 2005), with 0.05 significance.

⁴*t*-statistic is calculated in Weka as:

$$t = \frac{\bar{d}}{\sqrt{\left(\frac{1}{k} + \frac{n_2}{n_1}\right)\sigma_d^2}}$$

In our case, with 5-fold cross-validation repeated 5 times: $k = 25$, $\frac{n_2}{n_1} = \frac{0.2}{0.8}$ and σ_d^2 is the variance on 25 differences.

3.4.1. Automatic classification based on standard acoustic features

For these experiments 60 features were used, which incorporate those typically used in previous studies (Devillers et al., 2005; Lee and Narayanan, 2005; Morrison et al., 2007). These are utterance-level statistics corresponding to the four groups set out in Table 3.6.

Category	Features
Fundamental frequency (F0)	Min, max, range, mean, median, standard deviation, slope, correlation coef., regression error, value at first voiced segment, value at last voiced segment
F1, F2, B1, B2	Min, max, range, mean, median value at first voiced segment, value at last voiced segment
Energy	Min, max, range, mean, median, standard deviation, slope, correlation coef., regression error, value at first voiced segment, value at last voiced segment
Rhythm	Rate, voiced duration, unvoiced duration, value at first voiced, number of unvoiced segments

Table 3.6. *Acoustic features used for classification*

The first group is comprised of pitch features. Pitch depends on the tension of the vocal folds and the sub glottal air pressure (Ververidis and Kotropoulos, 2006), and can be used to obtain information about emotions in speech. As noted by Hansen (1996), mean pitch values may be employed as significant indicators for emotional speech when compared with neutral conditions. All the pitch features in the voiced portion of speech have been computed. Specifically, the focus was on the minimum value, maximum value, mean, median, standard deviation, value in the first voiced segment, value in the last voiced segment, correlation coefficient, slope, and error of the linear regression that describes the line that fits the pitch contour. All the duration parameters (e.g. slope) were normalized by the utterance duration to obtain comparable results for all the utterances in the corpus. To extract the pitch the modified autocorrelation algorithm (Boersma, 1993) was used.

The second group is comprised of features related to the first two formant frequencies (F1 and F2) and their bandwidths (B1 and B2). Only the first two formants were used, because it has been empirically demonstrated

that adding information about a third frequency does not introduce any informative features, neither in real nor in acted emotional corpora (Morrison et al., 2007). These frequencies are a representation of the vocal tract resonances. Speakers change the configuration of the vocal tract to distinguish the phonemes that they wish to utter, thus resulting in shifts of formant frequencies. Different speaking styles produce variations of the typical positions of formants. In the particular case of emotional speech, the vocal tract is modified by the emotional state. As pointed out by Hansen (1996), in stressed or depressed states speakers do not articulate voiced sounds with the same effort as in neutral emotional states. The features that have been used for categories F1, F2, B1 and B2 are minimum value, maximum value, range, mean, median, standard deviation and value in the first and last voiced segments of each utterance.

Energy is considered in the third group of features. As stated by Ververidis and Kotropoulos (2006), this feature can be exploited for emotion recognition because it is related to the arousal level of emotions. The variation of energy of words or utterances can be used as a significant indicator for various speech styles, as the vocal effort and ratio (duration) of voiced/unvoiced parts of speech change. For example, Hansen (1996) demonstrated that loud and angry emotions significantly increase intensity, i.e. energy. For these features, only non-zero values of energy have been used, similarly as for pitch, obtaining minimum value, maximum value, mean, median, standard deviation, value in the first voiced segment, value in the last voiced segment, correlation, slope, and error of the energy linear regression.

The fourth group is comprised of rhythm features. These are based on the duration of voiced and unvoiced segments. A segment is considered to be unvoiced if its fundamental frequency is zero. The reason for this is that F0 equals the fundamental frequency of the glottal pulses, which are only generated in the presence of speech. Rhythm and duration features can be good emotion indicators, as shown by previous studies. For example, Boersma (1993) noted that the duration variance decreases for most domains under fast stress conditions.

Five rhythm features were calculated: speech rate, duration of voiced segments, duration of unvoiced segments, duration of longest voiced segment and number of unvoiced segments. All these features were normalized by the overall duration of the utterance. The speech rate has been computed as

the number of syllables normalized by the utterance duration. To compute the utterance duration the number of frames used has been multiplied by the frame step. Using the 60 acoustic features described above and the 5-fold cross-validation strategy, an emotion recognition rate of 35.42% was obtained for the MLP, whereas for the majority-class baseline it was 51.67%. The significance studies using a t-test with 0.05 significance showed that this difference is significant.

Not all the features employed for classification are necessarily very informative. Unnecessary features make the learning process slower and increase the dimensionality of the problem. Therefore, a feature selection process (Guyon and Elisseeff, 2003) has been carried out employing three methods. Firstly, a forward selection algorithm like that used by Lee and Narayanan (2005), which selects the B1 value at the last voiced segment and the maximum energy. Secondly, a genetic search method that starts with no attributes and uses a population size of 20, 20 generations, 0.6 crossover probability and 0.033 mutation probability. The selected features for this method were the following: F1 maximum, F1 median, B1 minimum, B1 range, B1 median, B1 in the last voiced segment, B2 minimum, B2 maximum, B2 median, energy maximum, energy range, and energy in the last voiced segment. Thirdly, the attributes were ranked using information gain as a ranking filter. The results employing this method were: energy maximum ranked with 0.50, B1 in the last voiced segment with 0.46, and other features were evaluated with 0. Taking into account the three approaches, the optimal subset was composed of B1 in the last voiced segment and energy maximum. When classifying with the selected features only, no improvements were obtained in the experiment, as the percentage of correctly classified utterances was 49.00%, which is worse than obtained with the baseline. However, this difference was not significant in the t-test, which indicates that the results for both the MLP and the baseline are equivalent after feature selection.

3.4.2. Automatic classification based on normalized acoustic features

To reproduce the user's speaking style information that the annotators had when they labelled the corpus in the ordered case, a new approach is proposed in which acoustic features are normalized around the neutral voice of the user.

For example, let us say that user ‘A’ always speaks very fast and loudly, and user ‘B’ always speaks in a very relaxed way. Then, some acoustic features may be the same for ‘A’ neutral as for ‘B’ angry, which would make the automatic classification fail for one of the users. This is what happened to the annotators who were not used to the Andalusian dialect (Section 3.3.2), as they were confused by the fast rhythm and loud speech of the speakers who recorded the corpus.

In order to carry out the proposed normalization the user’s neutral voice features are obtained in each dialogue, and subtracted from the feature values obtained in the rest of the user’s utterances. To calculate the neutral voice, it could have been used the average value of all utterances of that user in the corpus labelled as *neutral* by the annotators. However, our intention for future work is to integrate our emotion recognizer in the UAH system, so that it can be adaptive to user’s emotions. It is impossible to carry out this computation in execution time as this would require to have all the user turns in the dialogue in advance. Therefore, the first utterance of the user was considered to be neutral, assuming that he is initially in a non-negative emotional state. Besides, the interest of the Thesis is only on emotions caused by the interaction with the system, assuming that the user is in a neutral emotional state when he starts the interaction with the dialogue system. This assumption is possible in domains not directly related to highly affective situations, such as bookings or information extraction, which are the typical application domains of spoken dialogue systems. For these application domains, dialogues in which the user is already in a negative emotional state are negligible.

The accuracy obtained with the Multilayer Perceptron using the normalized features was 53.17%. Thus, introducing acoustic context enables the MLP to improve over the results obtained by the baseline, but the improvement was not significant. Employing the features selected in Section 3.4.1 (B1 in the last voiced segment and energy maximum) 69.33% correctly classified utterances were obtained, which showed to be a significant improvement following the t-test. In the non-normalized case the feature selection did not yield any improvement. Thus, using normalized acoustic features yielded an improvement of 17.66% (69.33% recognition rate) over the baseline, which was also the best case in the non-normalized classification (Figure 3.9).

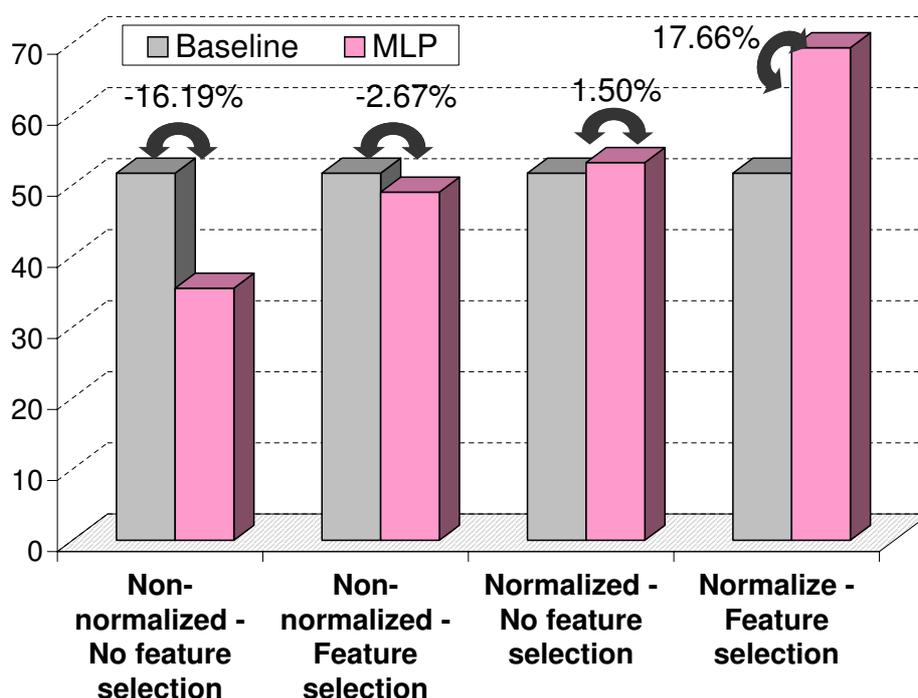


Figure 3.9. Recognition accuracy for *angry*, *bored* and *doubtful* considering non-normalized vs. normalized acoustic features, and no feature selection vs. feature selection

Thus, the normalization of traditional acoustic features yields a noticeable increase in the percentage of correctly recognized emotions with respect to the baseline. This is a very important result as, due to the natural skewness of non-acted emotional corpora, high accuracies can be obtained when directly assigning the most frequent category to all the prompts. In our case, the baseline yielded an accuracy higher than 50.00%. Only with normalization the MLP could obtain better results than the baseline, which were improved by a 17.66% when using acoustic features selection.

A study of the confusion matrices in both annotation schemes showed that the *doubtful* category was often confused with the *angry* or *bored* categories, with confusion percentages above 20% in most cases. A similar result was obtained for human annotation given that the ordered scheme did not improve the annotation of the *doubtful* emotion (as can be observed in Figure 3.3). These results show that contextual information affects automatic speech recognition using these classification methods, similarly as it affects human annotation.

Thus, in order to distinguish between *doubtful* and the other negative emotions, additional sources of contextual information must be added. Our proposal is to automatically recognize the three emotions using a two-step method. In the first one, acoustic information and contextual information about the user's neutral voice are used to distinguish between *angry* and *doubtfulORbored*. In the second step, dialogue context is used to discern between *doubtful* and *bored*. In the first step, the previously described normalization procedure was used to recognize *angry* vs. *doubtfulORbored*.

To optimize the results, another feature selection was carried out, in which the optimal features are those that best discriminate between *angry* and *doubtfulORbored*. Using the same feature selection algorithms, a subset comprised of three features was obtained. These were F2 median, energy maximum and energy mean. The results obtained are shown in Figure 3.10. All of them proved to be significantly better than the baseline using the t-test, except for the first case (non-normalized and no feature selection), where the results were the same order for the MLP and the baseline. The best result for *angry* and *doubtfulORbored* was achieved with feature selection in the ordered scheme, where an 80.00% accuracy was obtained.

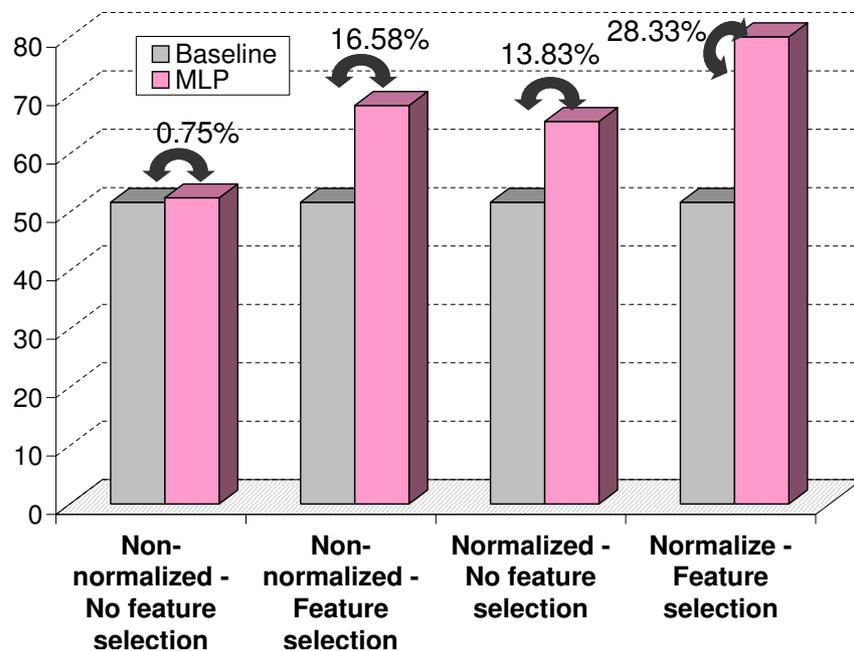


Figure 3.10. Recognition accuracy for *angry* and *doubtfulORbored* considering non-normalized vs. normalized acoustic features, and no feature selection vs. feature selection

With these experiments it has been shown that acoustic features normalized with neutral style of the users are preferable to the non-normalized ones, as these yield 17.66% improvement (69.33% vs. 51.67% success rate) when recognizing between the three negative emotions, as shown in Figure 3.9. Moreover, if acoustic information is used in the first recognition step which distinguishes between *angry* and *doubtfulORbored*, accuracy of 80.00% can be achieved, which represents an improvement of 28.33% over the baseline, and of 11.75% over the case when no context information about the user’s neutral voice is used (Figure 3.10). In the next section, a second step is described so that dialogue context can be added to distinguish between *doubtful* and *bored*.

3.4.3. Automatic classification based on dialogue context

As discussed in Section 3.3, dialogue context was provided to human annotators by giving them the ordered sequence of utterances in each dialogue. To represent context information for automatic recognition two labels have been employed: *depth* and *width*. The first of these indicates the total number of dialogue turns, whereas the second denotes the number of additional user turns necessary to obtain a particular piece of information (e.g. a person’s surname). Other authors have already studied the use of discourse structure in similar ways for other areas. For example, Rotaru and Litman (2006) studied how specific locations in the discourse structure are more prone to speech recognition errors in dialogue systems. To do so, they quantified the position of the user turns employing two similar features: “depth” and “transition”. However, their approach is based on intentional models of dialogue, which consider it divided in different goals or intentions that have to be satisfied to complete a task. Thus, they define “depth” of a user’s turn as the number of different intention subdialogues (or subgoals) in the previous history. This way in their approach several user turns can have the same “depth”. Similarly, their “transition” feature captures the position in the discourse structure relative to the previous turn, describing it using different labels for the type of transition between the last and the current turn (e.g. if the new turn introduces a new intention or continues with the same goal).

Although also based on vertical and horizontal positions of user turns in the dialogue structure, the proposed approach employs numerical representations, which allows the automatic calculation of dialogue context from the dialogue history, without introducing new manual annotations like Rotaru and Litman's "transition" tags. Thus, a completely new calculation of the dialogue context is proposed in which the different values of the vertical component (*depth*) are not shared by different turns. Only our *width* label is concerned with the subdialogue in which the user is engaged. However, our interest is only on subdialogue types that could cause a negative emotion, which have been considered to be those in which the user can not provide any new information and has to repeat, correct or confirm data.

The computation of the dialogue context in the proposed approach is as follows:

- *Depth* is initialized to 1, given that 0 would mean that the user hangs up the telephone before he says anything. *Depth* is incremented by 1 for each new user turn and each time the interaction goes backwards (e.g. to the main menu).
- *Width* is initialized to 0, which indicates that the user does not need to repeat any information to the system initially. It is incremented by 1 for each user turn generated to confirm, repeat data or ask the system for help. It is also incremented by 1 if the user has to disambiguate his own input, even if providing new information. Then *width* is reestablished to 0 when the user introduces a new piece of information.

For illustration purposes, a sample dialogue from the UAH corpus is shown below (translated from Spanish to English), where S = system, U = user:

S1> Welcome to the UAH system.
U1> I need information about professors.
S2> Please say the name of the professor.
U2> López.
S3> I have found two professors named López: Pedro López and Sara López. Who do you mean?
U3> Sara López.
S4> What information do you need about Prof. Sara López?
U4> I want her telephone number.

S5> Sorry, I did not understand. What information do you need?

U5> Telephone.

S6> The telephone number of Prof. Sara López is 958123456. Do you need any other information about professors?

U6> No thanks.

By applying the annotation scheme to this sample dialogue the following values have been obtained for *depth* and *width*:

U1> Depth=1, Width=0

U2> Depth=2, Width=0

U3> Depth=3, Width=1

U4> Depth=4, Width=0

U5> Depth=5, Width=1

U6> Depth=6, Width=0

It can be observed in this example that the user needed to employ two turns (U2 and U3) to make the system understand the professors' name. In turn U5 he rephrased what he said in turn U4, which solved the system misunderstanding. This is the reason why *width* is 1 for these two user turns.

This scheme is implemented automatically in the system using the dialogue history, which is stored in log files. The *depth* and *width* values of a user turn are computed checking the type of the previous system prompt. For example, as shown in Figure 3.11, *width* would only be 0 after a system prompt of type "subject name" if the current prompt type were "subject information". If the current prompt type were "subject disambiguation", *width* would be incremented by 1 because an additional user turn would be needed to provide the desired subject to the system.

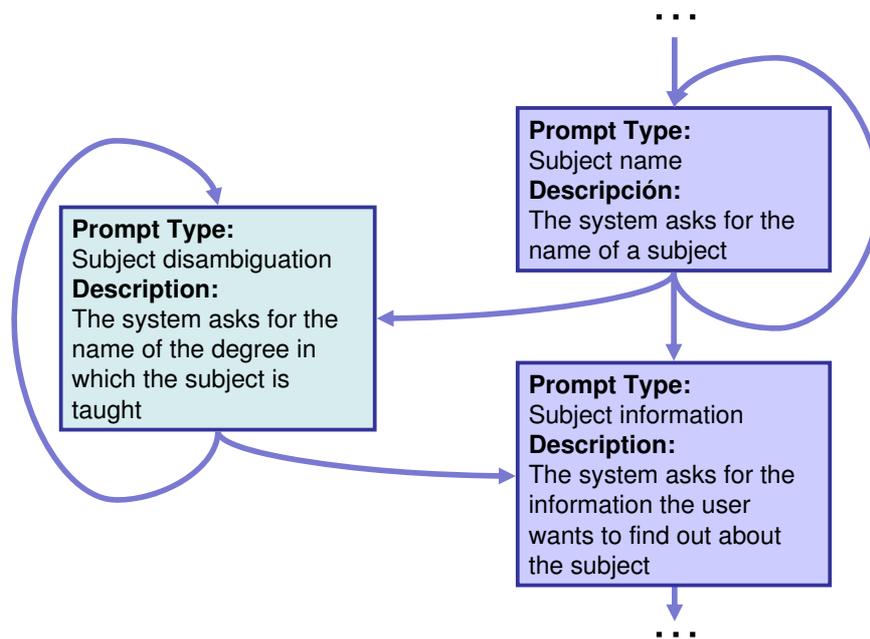


Figure 3.11. Example of transitions between system prompts in the UAH system

An exhaustive study of the UAH corpus showed that different users react with different emotions to the same dialogue state, in a less predictable way than initially expected. Employing one threshold for *depth* and another for *width* to distinguish independently between emotions was found to be inefficient as emotions are influenced by a mixture of the two. Furthermore, the study of the UAH corpus revealed that it is not sufficient to compute *width* considering only the previous turn or current subdialogue, but it is necessary to take into account the whole dialogue history. This differs from the results reported by Rotaru and Litman (2006), which annotate their horizontal variable (“transition”) only with information about the previous system prompt. For example, in the following dialogue:

(...)
 S1> Please say the name of the professor.
 U1> Martín.
 S2> Sorry, I did not understand. Please repeat the name.
 U2> Luis Martín.
 S3> Did you say Luis Marín?

U3> No, Luis Martín.

S4> What information do you need about Prof. Martín?

U4> His email address.

(...)

width would be 0 in U1, 1 in U2, 2 in U3 and 0 again in U4 because the dialogue starts to deal with a new piece of information. A high *width* value in U2 indicates a higher probability of the user being angry because of the misunderstandings and repetitions needed to make the system understand the name of the professor referred to. However, in turn U4 the user may still be angry but it has *width*=0.

This is why it has been defined another measure called *accumulated width*. Whereas *width* is a measure of the extra turns necessary to obtain a particular piece of information, *accumulated width* denotes the total number of extra turns employed in the whole dialogue up to the current utterance. *Accumulated width* is initialized to 0 and it is increased by 1 each time *width* is incremented. Thus, in the previous example, in U3 *width* = 2, which indicates that it was necessary to repeat or confirm the information of the professor's name twice. Note that in turn U4 *width* = 0 again because the system is gathering different data, namely the type of information about the professor that the user wants. Hence, *accumulated width* is more representative than *width* because it takes into account all the problematic points in the previous dialogue. For example, in U4 *accumulated width* = 2, which lets us know that there were 2 problematic turns before the current prompt.

The algorithm employed to classify the emotions based on the dialogue context information is as follows:

```

if Any of the 2 previous turns has been tagged as angry then
    angry
else if ( $D \leq 4$ ) AND ( $A \leq 1$ ) then
    doubtful
else if ( $\frac{A}{D} \geq 0.5$ ) OR (( $D > 4$ ) AND ( $\frac{A}{D} < 0.2$ )) then
    bored
else
    angry
end if

```

Where ‘D’ denotes *depth* and ‘A’ the *accumulated width*. In the proposed approach, the user utterances are considered as *doubtful* when the dialogues are short and have no more than one error, as in the first stages of the dialogue is more probable that the users are unsure about how to interact with the system. An utterance is recognized as *bored* when most of the dialogue has been employed in repeating or confirming information to the system. The user can also be *bored* when the number of errors is low but the dialogue has been long. Finally, an utterance is recognized as *angry* when the user was considered to be angry in at least one of the two previous turns in the dialogue (as described at the beginning of Section 3.3.2 with human annotation), or the utterance is not in any of the previous situations (i.e. the percentage of the full dialogue length comprised by the confirmations and/or repetitions is between 20% and 50%).

When considering a baseline that always classifies utterances with the most frequent emotion, which in our case is *angry* (same baseline as in Section 3.4.2), 51.67% accuracy is obtained in distinguishing between the three emotions. This rate is improved by 13.61% employing the proposed algorithm, which attains an accuracy of 65.28%.

3.4.4. Automatic classification based on normalized acoustic features and dialogue context (two-step method)

A method in two steps is proposed, which integrates both contextual sources: the users’ neutral speaking style (Section 3.4.2) and the dialogue context (Section 3.4.3). The acoustic features normalized with the users’ neutral speaking style are used to discriminate whether each utterance is *angry* or *doubtfulORbored*. Then, if an utterance is classified as *doubtfulORbored*, dialogue context information is used to distinguish between *doubtful* and *bored*. Additionally, dialogue context is used to classify utterances as *angry* if they were misrecognized in the first step. The results obtained by the two-step method are shown in Figure 3.12, and proved to be significant following the t-test.

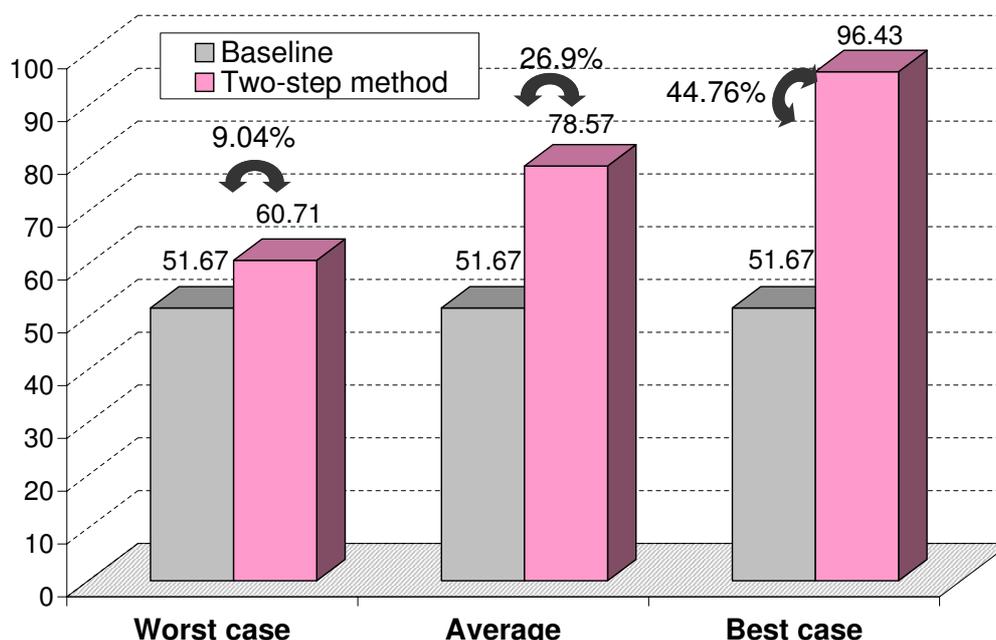


Figure 3.12. Emotion recognition accuracy using both acoustic and dialogue context information

Obviously, the result of the two-step method depends on the result of the first one, given that the *angry* vs. *doubtfulORbored* step can fail in the distinction of the two categories and the second step may have to categorize as *doubtful* or *bored* an utterance that belongs to the *angry* category. In the worst case, the first step can fail to recognize all the *angry* utterances, so that all the utterances are recognized as *doubtfulORbored* and passed to the second step. In this case, the recognition accuracy is 60.71%, as a mechanism to detect possible angry utterances has been incorporated to the second step (see Section 3.4.3). In an ideal best case, the first step would have 100% accuracy, and thus would correctly classify all the utterances as *angry* or *doubtfulORbored*. Thus, the second step would only have to classify the *doubtful* and *bored* utterances. The recognition rate in this case is 96.46%. However, as it was discussed in Section 3.4.2, with the UAH corpus the first step obtains a maximum 80.00% accuracy, which means that 20.00% of the *angry* utterances may be misrecognized. Employing the two-step method the recognition rate was again 96.43%. Thus, the misrecognized *angry* utterances could be correctly classified in the second step, obtaining a recognition rate for our best case in practice, which is identical to the ideal best case.

On average between the worst and best case, the two-step method obtains a 78.57% accuracy (as observed in Figure 3.13), which outperforms the baseline by 26.90%. The improvement over the baseline is 44.76% in the best case, i.e. when the first step does not fail. The average improvement over the recognition based only on neutral acoustic context is 9.24% (27.10% in the best case). If the recognition is based on dialogue context only, the average improvement is 14.29% (32.15% in the best case), and if it is based on the traditional approaches considering non-normalized acoustic features, the average improvement is 29.57% (47.43% in the best case).

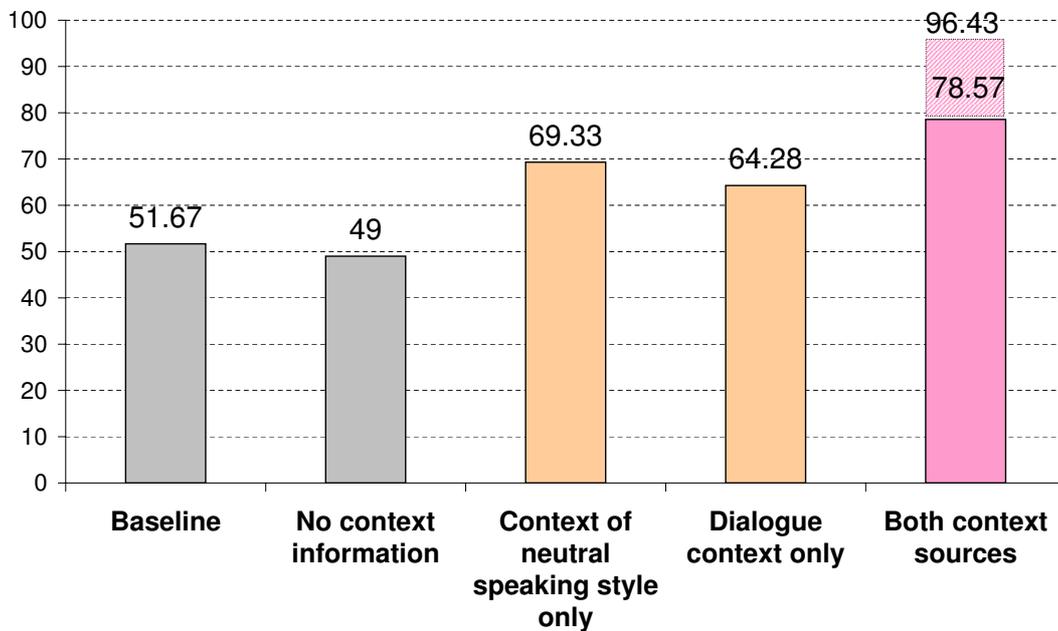


Figure 3.13. Comparison of the recognition accuracies of the methods for automatic emotion recognition

Thus, using only one context source (neutral voice or dialogue context), improves over both the baseline and the traditional approach where no context information is used. Besides, combining the two context sources in the proposed two-step method considerably outperforms the baseline, the traditional approach based on acoustic features without additional context sources, and the approach considering only one context source either the neutral voice of the user or the dialogue context.

3.4.5. Previous version of the two-step method

Before finding the optimal approach, a detailed study was carried out of the different possibilities to recognize *bored*, *angry* and *doubtful* emotions in a two-step method using the proposed context sources. The main objectives were to maximize: i) the significance of the obtained results, ii) the difference between the baseline and the proposed methods in each step, and iii) the use of contextual knowledge in the whole process.

A previous version of the two-step method distinguished firstly between *doubtful* and *angryORbored* using dialogue context, and secondly between *angry* and *bored* using the neutral speaking style of the user, as shown in Figure 3.14.

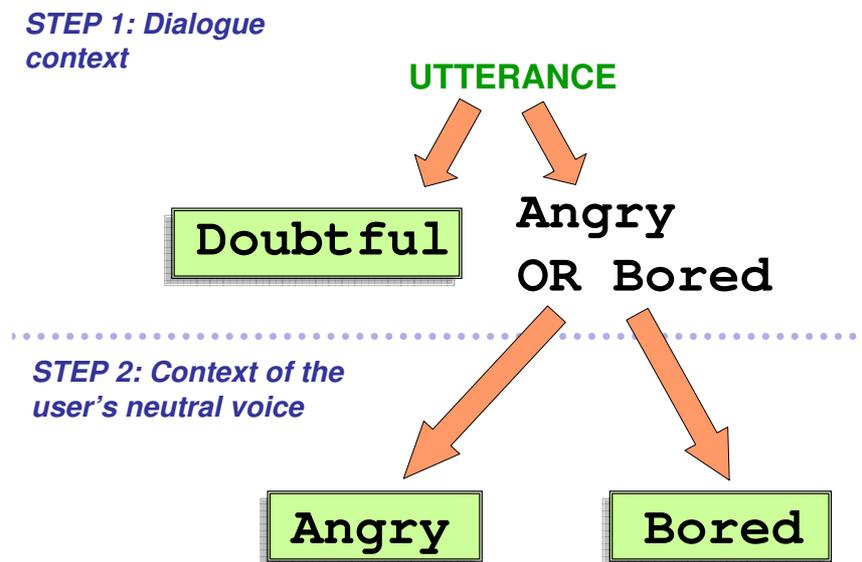


Figure 3.14. First version of the two-step method for automatic emotion recognition

In this first version of the two-step method, a static threshold T was used for dialogue context, which was computed as follows:

$$T = D + A$$

where 'D' denoted *depth* and 'A' the *accumulated width*. In this approach, a value of T greater than or equal to the threshold indicated *angryORbored*, whereas a smaller value indicated *doubtful*. Several values for the threshold were studied, which classification results are shown in Figure 3.15.

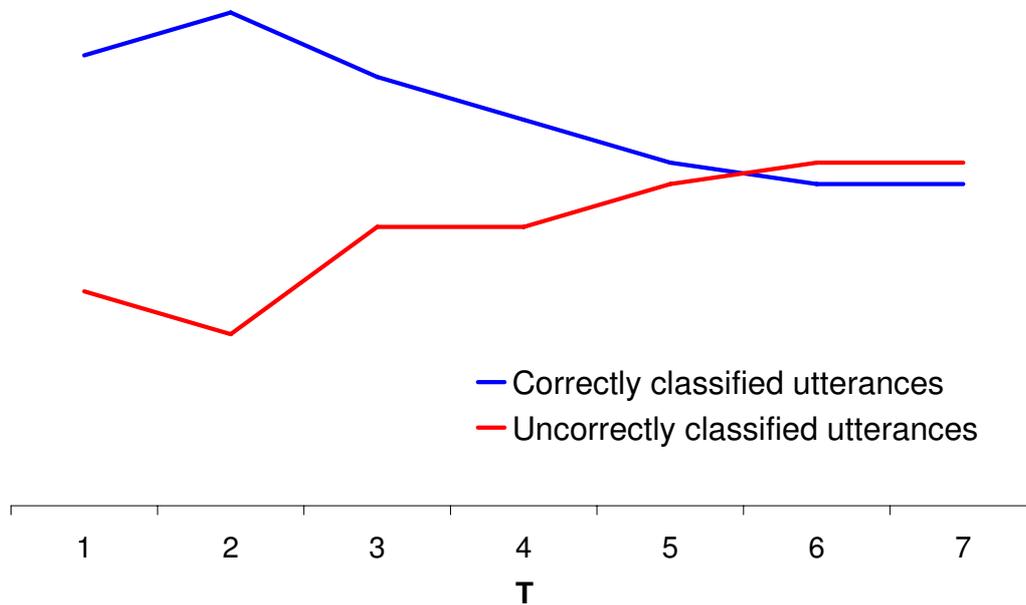


Figure 3.15. Impact of the value of dialogue context thresholds in emotion classification success

As can be observed, for thresholds greater than five the percentage of correctly classified utterances is smaller than the number of incorrectly classified. For lower threshold values there were more correctly than incorrectly classified utterances. However, although very low T values like $T=2$ yielded a higher difference between correctly and incorrectly classified instances, they were not optimal because the corpus was unbalanced (there were more utterances labelled as *angryORbored* than as *doubtful*) and thus the best results were obtained when almost all values were classified as *angryORbored*. In fact, the confusion matrices obtained showed that for values of T smaller than four, *doubtful* utterances were mostly incorrectly classified. As a result, $T = 4$ was employed as the optimal value for the threshold. Thus, the classification approach consisted in assigning *doubtful* to the user turns with $T < 4$ and *angryORbored* when $T \geq 4$; which yielded 70% classification accuracy.

Once an utterance was classified as *angryORbored*, the normalized acoustic features enabled the distinction between *bored* and *angry* following the same procedure as described for the final version of the two-step method in Section 3.4.4. The classification rate with the acoustic features was 85.71%

for the distinction between *angry* and *bored*. Thus, a maximum 60% classification rate could be attained for the three emotions (*angry*, *bored* and *doubtful*), assuming that the first step was completely successful (70% rate). This was 24.52% better than the case in which no context information was used, but it was worse than using only one of the context information sources separately, as shown in Figure 3.16

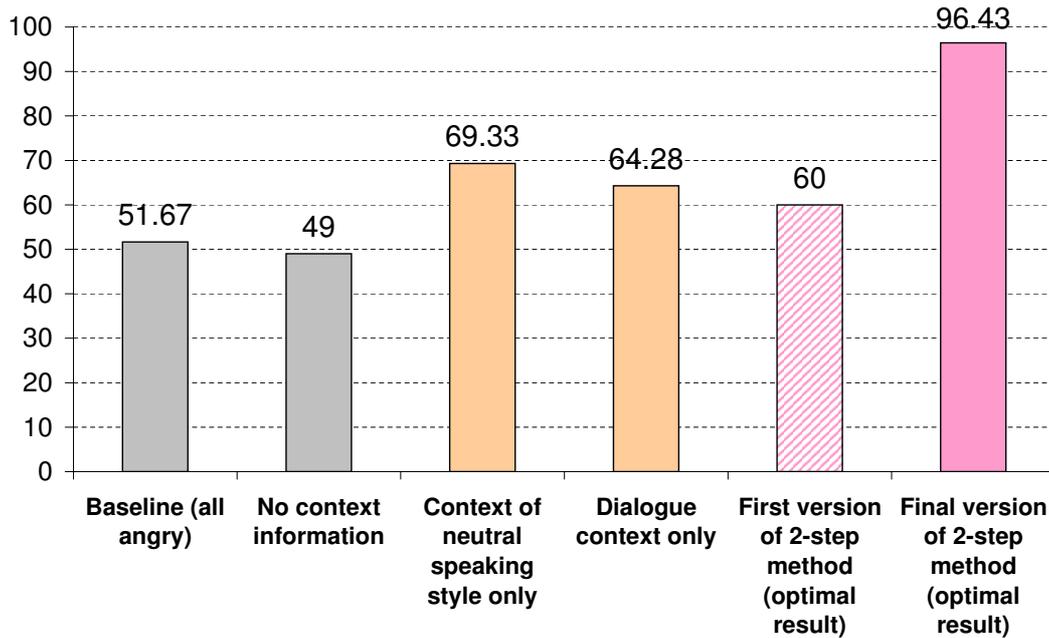


Figure 3.16. Comparison of the accuracy of the different methods proposed for automatic emotion recognition

After a deep study of the characteristics of the three emotions, the method was improved and the optimal version was obtained, which is the one that has been described in Section 3.4, and is shown in Figure 3.17.

In this scheme, employing acoustic information along with the context of the user's neutral voice, the utterances were classified as *angry* and *doubtfulORbored*. In a second step, those classified as *doubtfulORbored* were marked as *bored* or *doubtful* using the dialogue context. The results are totally comparable between the final two-step method, the baseline, and each context source employed separately. This was not possible in the first version of the method where the approach used to take into account dialogue context could not distinguish between angry and bored, and thus it could not be used in isolation to distinguish between the three emotions.

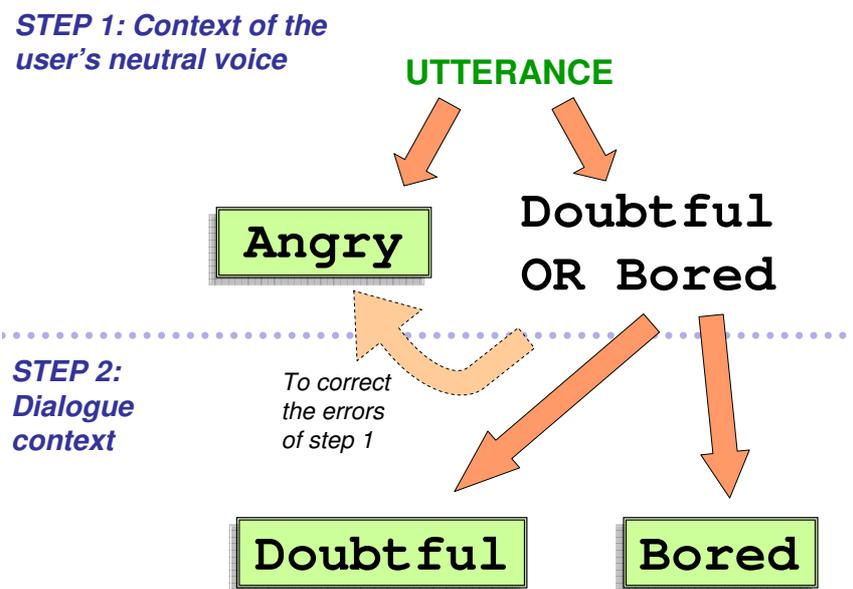


Figure 3.17. Final version of the two-step method for automatic emotion recognition

3.5 Conclusions

In this chapter several experiments have been carried out to study the annotation of human emotions in a corpus collected from real interactions with the UAH dialogue system. The experiments considered both manual annotation by 9 non-expert human annotators, as well as automatic classification employing MLPs and different feature selection techniques. It was found that human annotators marked 3.40% more non-neutral emotions when they had contextual knowledge. A plausible reason for this result is that context information makes it possible to identify non-trivial emotional speech (e.g. detecting emotions expressed more subtly). On the contrary, when the traditional non-normalized acoustic features were used, only very easily distinguishable emotions were annotated. Additionally, we have discussed the problems that the nature of non-acted emotional corpora impose in evaluating reliability of human annotations. Although deeply affected by the so-called paradoxes of Kappa coefficients, it has been studied how the inclusion of context information during annotation helps to obtain values closer to the maximum agreement rates obtainable when compared with not using any additional information.

For machine-learned classification methods, the experimental results show that, due to the natural unbalancement of the corpus, it is difficult to improve the baseline. This makes traditional recognition based on acoustic features yield results very similar to the majority-class baseline. However, as with human annotators, the emotion classification process is substantially improved when adding information about the user's neutral voice and the dialogue history. Just introducing the user's neutral acoustic context gives an improvement of 17.66%. Similarly, employing dialogue context information improved the baseline results in 12.67%. In this chapter it has been described a method in two steps to integrate both sources of contextual information. In this method, the normalized acoustic features are useful to distinguish between *angry* and *doubtfulORBored* categories with a 80.00% success rate. Once an utterance is classified as *doubtfulORBored*, the dialogue context enable us to distinguish between *doubtful* and *bored*. When the first step attains maximum accuracy, the two-step method obtains 96.43% accuracy. In the average case, the proposed method obtains 78.57% accuracy, which is 29.57% better than not using contextual information, 47.43% better in the best case (when the first step reaches its maximum performance).

In addition, the proposed methods can be employed during the running of a dialogue system as the contextual information sources can be obtained automatically and at execution time. To do so, a normalization process of the different acoustic measures with the users' neutral speaking style has been proposed; as well as a representation of the dialogue structure based on two parameters that can be numerically calculated from the information generated by the dialogue manager at run time.

*Kolika jazyků znáš,
tolikrát jsi člověkem*
Czech proverb

4

Cross-lingual adaptation of speech recognizers

4.1 Introduction

Cross-lingual adaptation makes it possible to employ corpora and resources already available in a language for the recognition of a different one. This allows fast and low-cost implementation of speech recognizers, although in detriment of the accuracy of the recognition result. However, this decrease in performance can be in many cases considered negligible when balanced against the cost of obtaining the resources required for building a recognizer for the new language. This approach is especially useful for minority languages or dialects in which the number of shared resources available is very limited or even not existent.

The hypothesis that we wanted to demonstrate in this chapter of the Thesis was that a fully functional system based on the Czech language can be easily and rapidly adapted for the interaction in another language without the need of building a new speech recognizer or getting involved in an arduous linguistic study, as for example morphological diacritization in (Kirchhoff and Vergyri, 2005). Thus, a new approach to reach this objective is proposed and experimental results that measure its appropriateness are presented with both a language that is similar to Czech (Slovak) and a language from a very different origin (Spanish).

Firstly, with the adaptation to the Slovak language we wanted to prove the suitability of the proposed technique to rapidly adapt an existing speech recognition system to work with a phonetically similar language. Besides, it

was also our aim to show that it is possible to take the most of the costly process of obtaining all the resources necessary to build speech recognition systems for a minority language such as Czech (spoken by approximately 12 million people), by employing it with a less-resourced language like Slovak (spoken approximately by 6 million people), and obtain high accuracy rates.

Secondly, the main contribution is the adaptation to Spanish, with which the objective was to find out whether this straightforward cross-lingual adaptation could also yield good results with languages that belong to different families and thus are phonetically less similar. As can be observed in Figure 4.1¹, Czech belongs to the family of Slavic languages like Russian, concretely to the Czech-Slovak together with Slovak. On the other hand, Spanish is an Italic language like Italian or French; concretely it belongs to the West-Iberian group as Portuguese. Thus, one of the challenges of the Thesis was to obtain a satisfactory mapping for such different languages (Czech and Spanish); especially when previous researches have obtained poor results in cross-language tasks between Slavic and Italic languages. For example, this is the case for Zgank et al. (2004), who studied Slovenian and Spanish and, based on their experimental results, recommended addressing only languages that are very similar to ensure maximal overlap of phonemes. In the literature it is frequent the usage of languages with the same roots, for example Italian and Spanish (Bonaventura et al., 1997) (Italic language family) or English and Afrikaans (Nieuwoudt and Botha, 2002) (Germanic language family).

The chapter is structured as follows. Section 4.2 presents related work in the area of cross-linguality. It describes our proposal and compares it with the state-of-the-art methods. Section 4.3 describes the previously created Czech speech recognizer which was used to recognize Slovak and Spanish, and the MyVoice system. In Section 4.4, the cross-lingual adaptation is explained for every language used in the experiments. Section 4.5 describes the experiments carried out to test the performance of the proposed technique, whereas Section 4.6 discusses the results obtained, and Section 4.7 describes the conclusions reached.

¹The continuous arrows indicate “belongs to family”, e.g. Spanish belongs to the West-Iberian family, which belongs to the Ibero-Romance one. The dotted arrows give examples of languages in other sub-families (e.g. Russian belongs to the Slavic family, in a subfamily different from the West-Slavic).

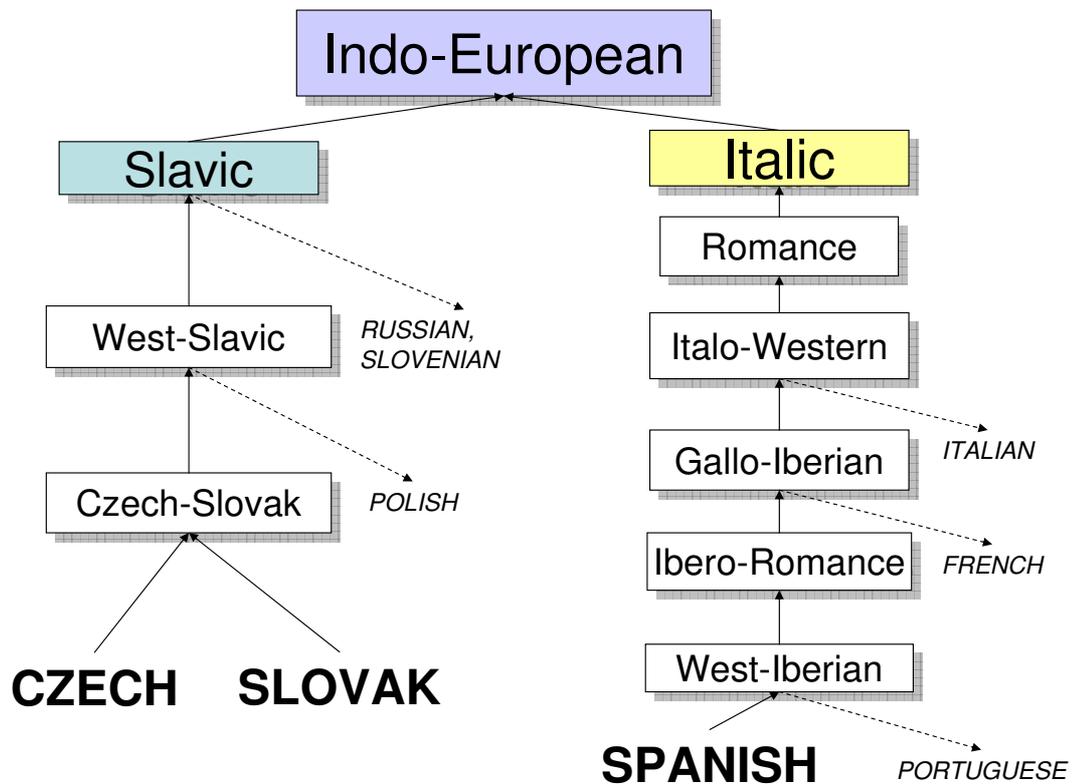


Figure 4.1. Language family groups for Czech, Slovak and Spanish

4.2 Related work

The term “cross-linguality” is used in many application domains in the field of computational linguistics, mainly to describe systems that can work employing several languages, and can be implemented at different levels. In the lexical level there are the natural language processing applications such as text retrieval, in which cross-lingual systems can search and rank documents written in a language different from the one in which the query is made (Fluhr et al., 1999; Martín-Valdivia et al., 2005). Cross-linguality is also applied to related areas such as question answering and text summarization (Radev et al., 2001; Ligozat et al., 2006). However, in these areas it is always necessary a translation between the involved languages, which must be carried out at some level of the text and query processing. Similarly, in

the semantic domain some intermediate representations are used to describe concepts and map them to the specific lexicalization of each language. For example, recent research is being done on how to map concepts to pictures to carry out multilingual web searches (Mihalcea and Leong, 2006).

When natural language is not written but spoken, cross-linguality can also be applied at the acoustic level. Traditionally, the most wide-spread systems that deal with cross-lingual acoustics are multilingual speech recognition systems (Schultz and Kirchhoff, 2006). These have been employed as components of speech-based interactive systems, as for example multilingual dialogue systems (López-Cózar and Araki, 2005), with which users can interact in different languages. Another example are speech-to-speech translation systems (Nakamura et al., 2004), which serve as real-time interpreters. Employing these systems, a user speaks to the telephone and the interlocutor receives the utterances translated to another language.

Acoustic cross-linguality has also been addressed from the point of view of resource sharing. The development of a speech recognizer is a very arduous and time demanding task. A large amount of data spoken by hundreds of subjects must be recorded and carefully annotated to get a representative set suitable for training an acoustic model. For example, the system used in our experiments for the speech recognition of Czech, which is described in Section 4.3, relies on phoneme HMMs that have been trained on approximately 50 hours of phonetically annotated speech provided by 700 speakers. For other languages the resources are often even much larger. Thus, collecting and annotating the necessary data generally requires many years of human effort. This is the reason why any possibility to share acoustic data between languages is very welcome in the scientific community. Similarly, a big effort is needed to create the linguistic part of a speech recognition system. For example, in the case of Czech it was necessary to collect a text corpus of around 4 GB, comprised of 456 million tokens with 1.9 different words and word-forms, to enable us to make a representative lexical analysis. It was shown that to obtain approximately 99% coverage of this language, at least 500,000 words had to be included in the general-purpose lexicon. This happens because of the inflective nature of Czech, which implies a big variety of word forms. However, the size of the needed vocabulary varies between languages. For example, English requires only 20,000 words to obtain the same order of coverage (Németh and Zainkó, 2003).

Frequently, these databases are gathered ad hoc by each developing team and are not freely available for the scientific community. This implies that in most cases, building a new speech recognizer requires obtaining all the necessary acoustic and linguistic resources right from the start. Thus, cross-lingual methods are an alternative to the complete creation of a new recognizer. Finally, sharing acoustics across languages has not only been used to create a speech recognizer from the resources of an already available one, but also to improve the performance of speech recognizers using models from another language. For example, some previous studies have shown that the recognition of Afrikaans can be improved by using additional English speech data (Nieuwoudt and Botha, 1999).

The usage of acoustic information across languages can be addressed following different strategies, which can be divided into two approaches: multilingual applications that can handle multiple languages simultaneously, and language adaptation where an existing recognizer is adapted to a new target language.

In the first approach, multilingual recognizers are capable of recognizing simultaneously several languages by sharing acoustic and/or language models. Multilingual acoustic models consist of either a collection of language-dependent acoustic models for each language, or a combination of language-independent acoustic models (Schultz and Kirchhoff, 2006). The main idea of the latter is to combine phonemes of several languages into one single acoustic model. To determine which phonemes of the different languages must be combined in the same category, some multilingual phoneme databases have been used, for example GlobalPhone (Schultz and Waibel, 1998). This technique relies on abstractions over the concept of *phoneme* in higher-order units such as meta-phonemes or archi-phonemes (Cahill and Tiberius, 2002), and assuming that phonemes in different languages can be grouped together similarly as allophones are considered inside the concept of phoneme.

The second approach carries out an adaptation of the “mono-language” speech recognizers to other languages. One possible method to do this is to create a mapping between phonemes. Alternative methods like word mapping (Bayeh et al., 2004), have been proposed. However, although using this method can sometimes lead to better results, it is more expensive and less practical than using phonemes. This happens because word mapping is less

prone to reusability, as it requires a complete translation of all possible words along with their different inflections. On the contrary, phonemes constitute a smaller set that can be used to automatically build any other higher-order construction such as words.

The basic idea of phoneme mapping is to establish a correspondence between phonetic units in the origin and target languages. Thus, the result depends on the phonetic similarity between both languages. This mapping can be done either automatically or by experts. The automatic procedure employs data-driven measures, which are frequently extracted from phoneme confusion matrices. The expert-driven approach is based on human knowledge about the languages being processed. Usually, International Phonetic Alphabet (IPA) ² symbol tables are used for the different languages to determine the equivalent phonemes (Schultz and Kirchoff, 2006), as IPA defines a unique representation of phonemes which can be used to compute equivalences between languages.

The mapping approach has been used because it was not in our scope to build a multilingual recognizer, but to use the Czech one with Slovak or Spanish utterances. Thus, the meta-phoneme concept was not effective for our purposes. To carry out the mapping, the expert-driven approach was chosen. The reasons for this are two. Firstly, although automatic mapping has the advantage of not needing human intervention and thus obtains more objective results, it requires considerable speech material for computing the similarities between languages. Even though this is not as much material as needed for building a full new recognizer for the target language, it makes the adaptation process be more costly. Secondly, results depend on a close match between the acoustics of both languages, on the used distance measure (Kumar et al., 2005), and on the recording conditions.

4.3 The MyVoice system and Czech speech recognizer

The Czech speech recognizer used in the experiments has been developed during more than a decade in Technical University of Liberec (Nouza et al., 2005). Its acoustic models are based on three-state left-to-right HMMs of

²IPA official web page: <http://www.arts.gla.ac.uk/IPA/>

context-independent speech units which are supplemented by several models of noise, with output distributions using at least 64 Gaussians. According to the application conditions, these models can be either speaker-independent (SI), gender-dependent (GD), or speaker-adapted (SA).

The recognizer's decoding module uses a lexicon of alphabetically ordered words, each represented by its text and phonetic form. This recognizer has been successfully employed for the development of the MyVoice and MyDictate systems (Cerva and Nouza, 2007). The former allows persons with non-functional hands to work with a PC in a hands-free manner by using several hundreds of voice commands. To do so, MyVoice interprets spoken commands into one or more basic actions; for example, pressing, holding and releasing key or combination of keys in a keyboard, moving the mouse cursor and clicking mouse buttons, starting executable programs, and printing sequences of characters. The latter is the first dictation program developed for the Czech language. It works with a very large vocabulary comprised of the 540,000 most frequent Czech words and it is primarily meant as a powerful aid for motor-handicapped users.

For the experimentation presented in the chapter, the MyVoice system has been employed and translated to Spanish and Slovak. MyVoice is structured in several command groups, each dealing with a specific task. For example, the group that controls the mouse is different from the one that deals with the keyboard, but they can be accessed easily from each other by voice commands. The size of these groups varies between 5 and 137 commands, where the largest group contains mainly the names of the alphabet letters, and of the keys on a PC keyboard, which makes recognition very difficult as the acoustic difference between them is very subtle. However, as a specific vocabulary is defined for each task, better recognition results are achieved when the commands are grouped. Besides, the grouping facilitates the interaction, as the user is aware of the valid words than can be uttered at any time. The MyVoice software is currently employed by 60 handicapped users in the Czech Republic, whose reports show that the word error rate (WER) typically lies between 1% and 2%, if the user does not have any speech disorder.

The development of MyVoice was motivated by the fact that there were no commercial tools of that type for Czech handicapped users. This is also the case for many other languages for which speech technology has not

been developed yet, and for which deploying such systems would require a big investment that would obtain little benefits due to the reduced target population. Therefore, we started to investigate the possibility to port software like MyVoice to other languages. Firstly focusing on the Slovak language, which is very similar to Czech. Secondly, making a more complex attempt to apply the same porting strategy to Spanish, which is a language that is acoustically and linguistically more different.

4.4 Cross-lingual adaptation

For the cross-language adaptation Slovak and Spanish texts were used along with an automatically generated Czech phonetic representation, which is discussed in Sections 4.4.1 and 4.4.2, respectively. The phonemes built for the Czech recognizer can be applied to recognize words in another language. To do this the Czech phonetic form was used, and the acoustic models of the words were constructed by concatenating the corresponding phoneme models. The translation from Slovak or Spanish text to the Czech phonetic representation was automatically done employing the mapping policies shown in Tables 4.1 and 4.2, respectively. For the phonetic representation a mapping of the IPA symbols to a set of ASCII symbols named Phonetic Alphabet for Czech (PAC) was used. We are aware of the fact that there are several encodings of the IPA alphabet such as XSAMPA. However, we have used PAC for two main reasons: firstly, because it has been established as a common base for speech processing research in the Czech Republic (Nouza et al., 1997). Secondly, because it has been successfully employed by the users of MyVoice, providing them with a straightforward language with which they can type their own phonetic transcriptions to customize the pronunciation of the voice commands.

As illustrated in Figure 4.2, the result of this process was a vocabulary with all the words that the speech recognizer accepted. For each word it contained a Slovak or Spanish text form and its Czech phonetic representation. This vocabulary was used by the Czech speech recognizer described in Section 4.3 as if it were a Czech vocabulary, thus not even a single line of code had to be changed in it. As a result, a user can utter a word in Spanish or Slovak, and the speech recognizer uses the Czech models to obtain the best Spanish or Slovak form candidate.

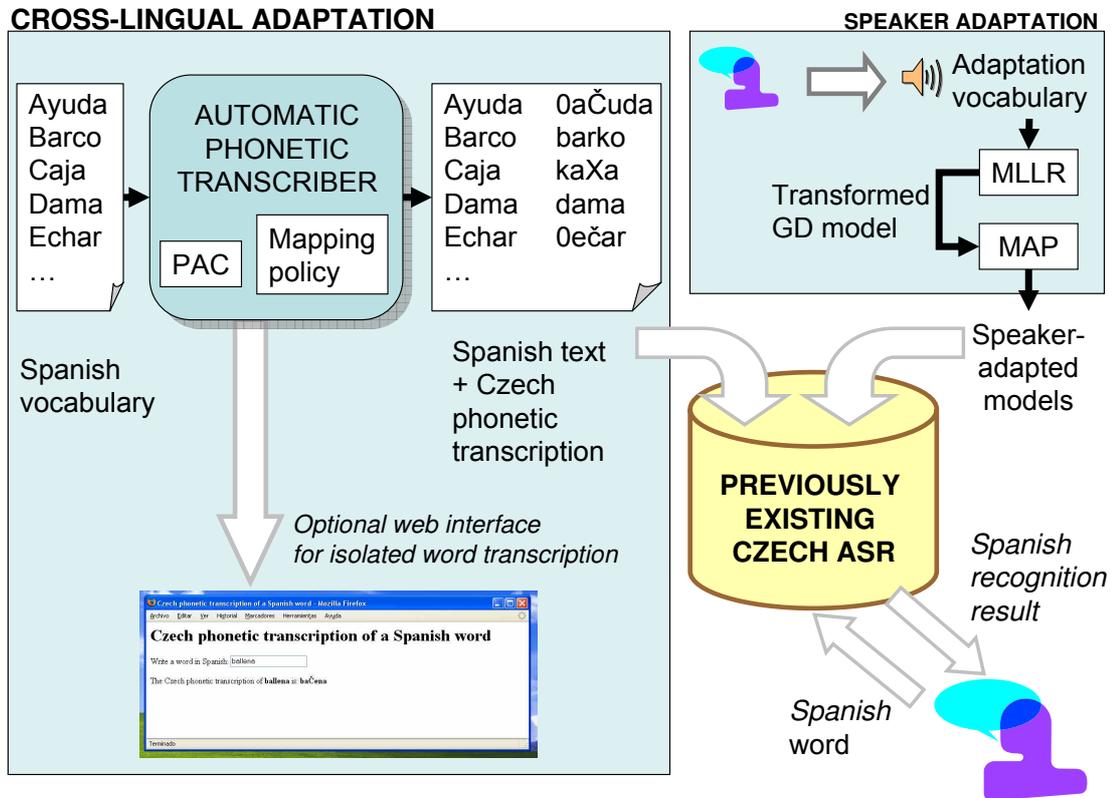


Figure 4.2. Scheme of the cross-language adaptation procedure

To optimize the performance of the recognizer, our proposal is to carry out speaker adaptation as a final step of the cross-language adaptation procedure. This way, the Czech models were tuned to better adapt to the pronunciation that each speaker had of every phoneme in the target languages. This step is not against the initial aim of cost-effective implementation, as it can be performed in a fast and straightforward manner by making the user read a short text when he first uses the recognizer (e.g. the first time he runs MyVoice). The proposal for speaker adaptation is described in Section 4.4.3.

4.4.1. Phoneme mapping between Czech and Slovak

As can be observed in Figure 4.1, Slovak and Czech belong to the same branch of Slavic languages. They share a large portion (about 40%) of their lexical

inventory and many of the remaining words differ slightly, either in spelling or pronunciation. In general, Slovak language sounds softer than Czech, which is caused by different phonotactics and a slightly different set of phonemes. The Slovak language uses four additional phonemes that do not occur in Czech: ‘ĺ’, ‘ľ’, ‘ŕ’ and ‘ŕ’. As shown in Table 4.1, these phonemes were mapped to the closest Czech phonemes. Concretely, the Slovak phonemes represented by characters ‘ĺ’ and ‘ľ’ were mapped to the Czech phoneme ‘l’, Slovak ‘ŕ’ to Czech ‘r’, and Slovak diphthong ‘ŕ’ to the sequence ‘uo’. The effect of this simplification on the performance of the adapted speech recognizer is very small, due to the similarity between the Slovak and Czech phonemes employed.

Czech phoneme	PAC	Slovak phoneme	Czech example	Slovak example
l	l	l, ĺ, ľ	lano	lano, dĺhý, ľud
r	r	ŕ	bere	mŕtvý
uo	uo	ô	duo	môže

Table 4.1. Mapping of the Slovak phonemes that do not exist in Czech to the closest Czech ones

4.4.2. Phoneme mapping between Czech and Spanish

The correspondence between Spanish and Czech phonemes was carried out by one Spanish native speaker and supervised by several Czech native speakers. As stated by Zgank et al. (2004), the accuracy of the correspondence depends on the number of phonemes present in each language and the similarity between them. It is difficult to find a consensus about the exact number of phonemes in the languages used in the experiments. This is especially the case for Spanish, as it has a high number of different varieties, even considering only the European branch and discarding the South American variants. However, in literature, Czech is generally considered to have around 40 phonemes, and Spanish around 20, which shows a big unbalance between both languages.

The result of the mapping is presented in Table 4.2, which only shows the phonemes that were employed for the Spanish recognition. A complete list of Czech phonemes can be found in (Nouza et al., 1997).

Czech phoneme	PAC	Spanish phoneme	Czech example	Spanish example
a	a	a, á	plo ch a	a nillo, á guila
b	b	b, v	b á b a	abuela, v ino
č	č	ch	č ichá	ch arco
ch	X	g, j	ch udý	j aula, g ema
dž	Č	y, ll	rád ž a	llave, yema
d	d	d	je d en	d entro
e	e	e, é	le v	e so, ca f é
f	f	f	f auna	f auna
g	g	g, gu	g uma	g oma, g uisante
i, y	i	i, í	i l, y l	i no, implícito
k	k	c, k, q	k upec	k ilo, q ueso, c asa
l	l	l	de l a	l ibro
m	m	m	m á m a	m adre
n	n	n	vi n o	v ino
ň	ň	ñ	ko ň e	E ñ a
o	o	o, ó	ko l o	ho l a, ca m ió n
p	p	p	p upen	p adre
r	r	r	be r e	a rco
s	s	s	su d	su e lo
t	t	t	du t ý	t eja
u	u	u, ú, ü	du š e	lu n a, ú til, ping ü ino
-	S / s	c, z	-	z umo, c ena
-	R / r	rr	-	r ueda, p er r o

Table 4.2. Mapping of the Spanish phonemes to the closest Czech ones

As can be observed in the last two rows of the table, there are two Spanish phonemes that do not exist in Czech: $/\theta/$ and $/r/$, in IPA representation. For these phonemes two solutions have been studied. The first one was to use the nearest Czech phonemes: s and r , in PAC representation.

This mapping is not so unnatural, as the pronunciation of / θ / as / s / is even present in some varieties of Spanish, for example in Latin America and some areas in Southern Spain. The second option was to adapt these previously existing Czech phonemes to the Spanish pronunciation, by creating two new symbols, S and R , for the PAC table. Experimental results were very similar for both approaches, with a difference in accuracy of only 0.5%. The results reported in Section 4.6 were attained employing the second approach.

Additionally, as our system does not consider allophones, some sounds were ignored, affecting very slightly the experimental results. Moreover, there are differences in stress that make the recognition of Spanish more difficult. Czech words are always stressed in the first syllable, whereas the stress in Spanish varies between words. As noted by Carreiras et al. (1996), differences in stress in Spanish may be of importance not only for automatic speech recognition, but also for human listeners. However, the effect of stress is less important for isolated word recognition; even when some words that are differentiated in Spanish by their stress cannot be distinguished when translated to the Czech pronunciation, for example “este” and “esté”.

4.4.3. Speaker adaptation

The approach proposed for speaker adaptation is a combination of the Maximum A Posteriori (MAP) (Gauvain and Lee, 1994) and the Maximum Likelihood Linear Regression (MLLR) (Gales and Woodland, 1996) methods for speaker adaptation, and is performed in two steps. In the first step, the mean vectors of the Czech gender-dependent models are transformed by the MLLR method. In the second step, these transformed values are used as priors for the MAP based adaptation. The main benefit of this approach is that the models that are not seen in the adaptation are well adapted by the MLLR method; while the MAP ensures that the parameters of the models with a lot of adaptation data can converge to the values of the theoretically best speaker-adapted model.

To carry out the speaker adaptation a 614 words vocabulary was used. It was comprised of a list of the most frequent words in each language (covering all phonemes), along with MyVoice commands. This decision was taken from experimental results, which showed that in most cases, misrecognitions occurred for monosyllabic and short words. Besides, these words, which for

example in Spanish generally are pronouns, determinants and prepositions, are usually the most frequent. Thus, misrecognition of these words had a big impact in the computation of accuracy. Concretely, a 44.9% relative improvement was achieved for Spanish when using this vocabulary for adaptation instead of the 432 MyVoice commands only. Additional experiments showed that this improvement was better than the one obtained using the same amount of phonemes for adaptation, but extracting them from words selected from newspapers instead of using the most frequent words of each language; even when both word lists considered all the phonemes of each language, and these were covered in identical proportions.

4.5 Experimental set-up

Several experiments were carried out with the main objective of testing the viability and performance of the proposed cross-lingual adaptation approach. Additionally, it was measured the impact of several factors in the performance of the employed method, such as the usage of different user adaptation strategies, the size of the recognition dictionary, and the number of words considered for testing, as shown in Figure 4.3.

Firstly, the MyVoice commands were translated into Slovak and Spanish, to measure the performance of the cross-lingual adaptation for a command-and-control application. The speakers used MyVoice to control a PC by spoken commands while they carried out their daily activities. Thus, they were not provided with a specific list of commands to utter. This way, the results were attained from flexible and natural interaction with the system.

As it was described in Section 4.3, the valid vocabulary of MyVoice is restricted at each step to the list of commands in the current group (vocabulary ranging between 5 and 137 commands). As commented before, the largest group was designed for the recognition of spoken characters. This is a very challenging task, mainly due to the acoustic similarity of the characters, which makes them highly confusable. Thus, although the grouping of commands considerably reduces the vocabulary that can be recognized at each time, speech recognition during natural interaction with MyVoice is not trivial, as it can involve the recognition of spoken characters.

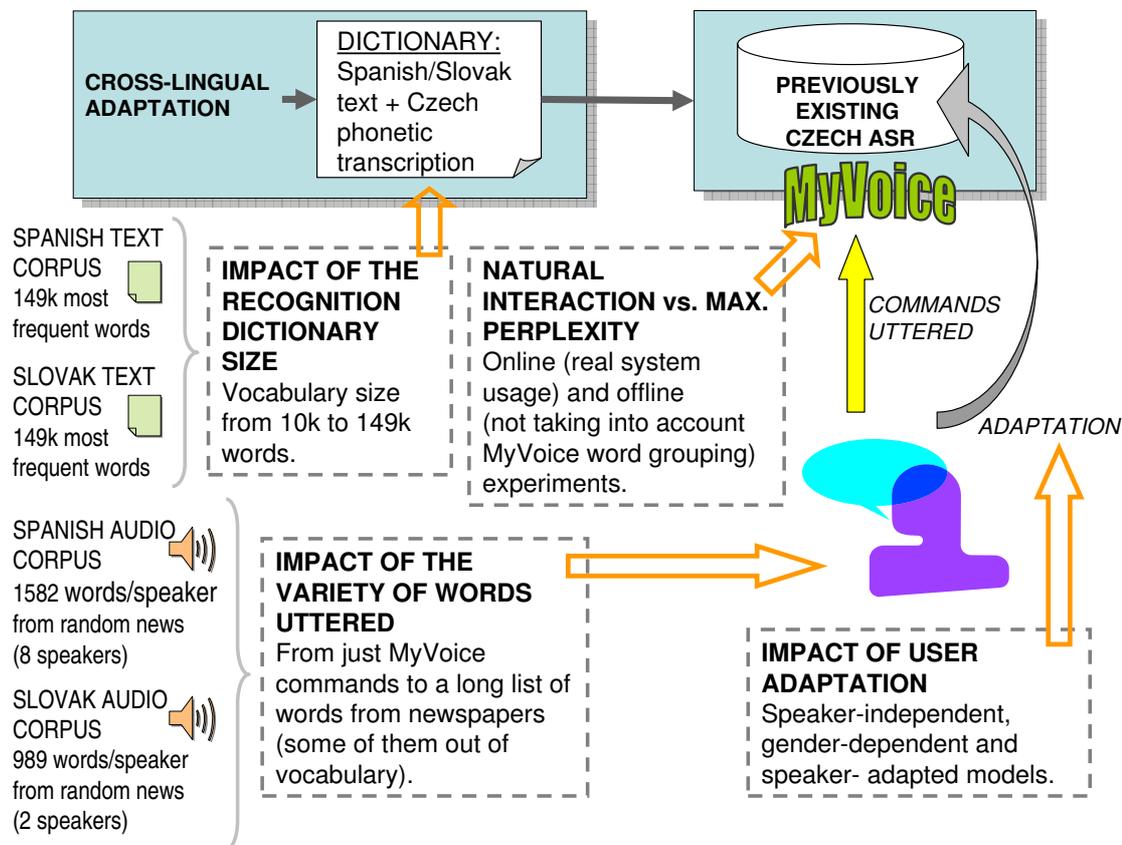


Figure 4.3. Outline of the experimental set-up

To obtain meaningful results from the different speaker models regardless of the groups visited during the interaction, additional experiments were carried out employing the whole MyVoice vocabulary (432 commands). In these experiments the task perplexity was always 432, given that after each command, any word could be uttered.

We also wanted to corroborate that the results obtained from the interaction with MyVoice could be attainable in situations where the accepted vocabulary was larger. Thus, MyVoice commands were extended with a list of the 149k most frequent Spanish and Slovak words, respectively. The two dictionaries were collected from Spanish and Slovak newspapers and contained all the word forms, not only the lexemes. The words were automatically annotated with their appearance frequency and sorted from most frequent to least. In these experiments recognition dictionaries of sizes ranging from 10k

to 149k words were used, which were subsets of the described dictionary, from the most frequent word up to the desired number of words. Our aim was not to build a Slovak or Spanish dictation system, but simply to check whether the cross-lingual approach would also work efficiently for a more complex task with increasingly larger vocabularies. For this purpose, a vocabulary of 149k words is larger than those usually employed in previous cross-lingual studies. For example, Zgank et al. (2004) only varied their vocabulary size from two to several thousand words, whereas Nieuwoudt and Botha (1999) employed 60 sentences as the maximum for their test purposes.

At the same time that the list of MyVoice available commands was extended, it was also augmented the vocabulary employed to test the system. To do this, news from Spanish and Slovak newspapers were randomly selected. They were different from the ones used to collect the recognition dictionaries, and belonged to different categories, namely: politics, economy, culture and sport. Eight Spanish native speakers (four male and four female) aged 23 to 60, and two Slovak native speakers (one male and one female) aged 22 and 24, recorded the isolated words. The experiments with the two Slovak speakers using MyVoice to carry out their daily activities, showed that the WER was almost as small as for native Czech speakers (only 2.5%), and that the space for any further improvement was very small. Therefore, a decision was taken to not to carry out any additional experiments with other Slovak speakers, highlighting that the main contribution is the adaptation to Spanish. Concretely, 1,582 words were recorded by each Spanish speaker, and 989 by each Slovak speaker. For discussing the experimental results we used the average performance values over the number of words for each language, and for all speakers. To obtain results that can be significant for the usage of the proposed method in real conditions, the corpora were not recorded in a closed laboratory environment, but in each speaker's PC in order to reflect the real noise conditions in which the MyVoice system would be used.

Furthermore, in order to study to which extent speaker adaptation allowed us to attain better recognition results, experiments were carried out with speaker-independent, gender-dependent, as well as speaker-adapted models. The improvements achieved by each adaptation step were compared and studied taking into account the impact of out-of-vocabulary (OOV) words and the size of the recognition vocabulary.

4.6 Experimental results

4.6.1. Interaction with MyVoice

In the first experiments the users employed MyVoice to control their PCs in order to carry out their daily activities. The experiments were performed both online and offline. The online results were extracted from natural interactions of the users with the MyVoice system. Thus, as commented before, the perplexity of the recognition task varied from 5 to 137, as the valid vocabulary was comprised of the words in the current command group. In the offline case, the same utterances recorded during the online experiments were used. However, recognition was carried out offline using as valid vocabulary all MyVoice commands. Hence, as all the commands could be uttered any time, the recognition perplexity was 432 (the number of MyVoice commands).

The experimental results are shown in Table 4.3. It can be observed that WER was lower for the online experiments because the vocabulary size was smaller. When using speaker-adapted models, the relative improvements achieved were 24.1% for Slovak and 28.3% for Spanish for the online experiments; whereas they were 46.65% and 56% respectively for the offline experiments. This shows that speaker adaptation considerably reduced the difference in performance observed for the online and offline recognition, as it caused a remarkable improvement in the offline recognition results; which were comparable to the ones obtained in the online experiments after speaker adaptation (around 2% WER for Slovak and 4% for Spanish).

Language	Experiment	Gender-dependent	Speaker-adapted
Slovak	Online	2.9	2.2
	Offline	4.6	2.5
Spanish	Online	6.0	4.3
	Offline	10.0	4.4

Table 4.3. WER [in %] for the command-and-control task

The experiments with Slovak showed a WER almost as small as for native Czech speakers (only 2.5%). The results with Spanish were much better than initially expected. In fact, they differed in less than 2% compared with the ones that could be achieved recognizing the Czech language.

4.6.2. Impact of speaker adaptation

The good results obtained using speaker adaptation during the natural interaction with MyVoice, encouraged us to study to which extent the speaker-adapted models convey an improvement in the proposed cross-lingual approach. To test the performance of the adapted recognizers the Spanish corpus (12,686 words extracted from newspapers) and the Slovak corpus (1,978 words) were used, which were described in Section 4.5. Additionally, we used a 10k words vocabulary for recognition instead of the small vocabulary comprised of 432 commands. As can be observed in Figure 4.4, speaker-independent models yielded a WER of 47% for Slovak and 55.8% for Spanish. These results were improved by using gender-adapted models only in a 7.23% relative for Slovak and 2.15% for Spanish. However, speaker adaptation yielded a 17.6% relative improvement with respect to the speaker-independent models for Slovak, and a remarkable 40.8% relative improvement for Spanish. Most of the recognition errors for Slovak were due to OOV words. Thus, as the objective of this experiment was to measure the impact of speaker adaptation in the proposed cross-language approach, regardless of the dictionary and utterances used for recognition, the recognition results were computed without considering the OOV words. As shown in Figure 4.4, WER decreased for both languages when OOV words were not considered. Concretely, for Slovak the decrement was of 20.9% absolute for the best case. Regarding speaker adaptation, a 54% relative improvement with respect to speaker-independent models was achieved for Spanish, and a 38.2% in the case of Slovak, obtaining for both languages WERs around 20%.

As Slovak is a language very similar to Czech, initially the proposed method attains accuracies around 70% (29% WER) for this language. Hence, speaker adaptation for Slovak only improves accuracy by an absolute 11.1% (38.2% relative) with respect to using speaker-independent models. However, for a language with a very different origin such as Spanish, speaker adaptation enhances the adapted recognizer substantially. The experiments showed that 26.4% absolute improvement (54% relative) can be achieved, with accuracy rates that are only 4.6% worse than the ones obtained for the Slovak language. The proposed cross-lingual approach in combination with speaker adaptation yielded accuracy rates around 80% (17.9% and 22.5% WER) for Slovak and Spanish.

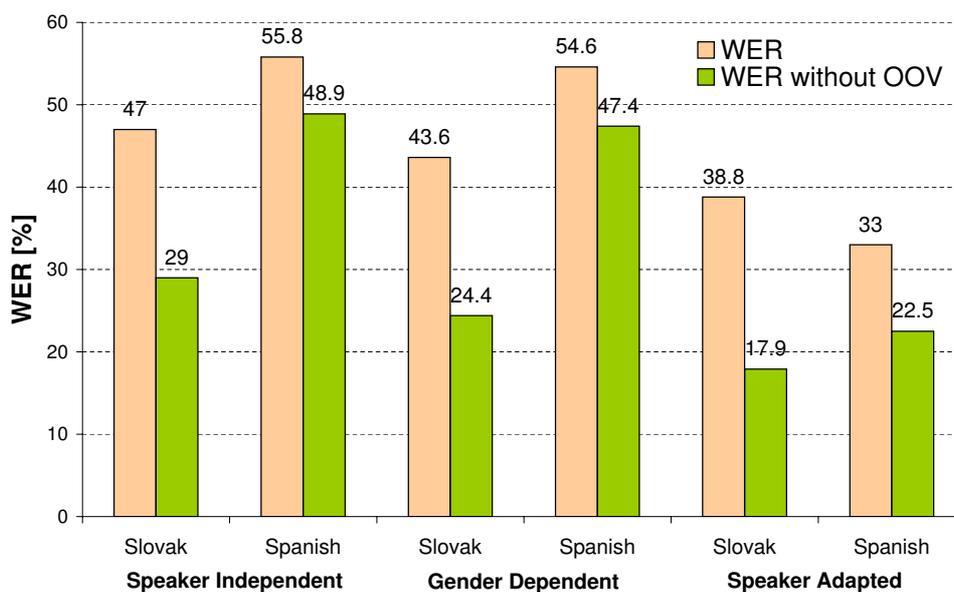


Figure 4.4. Effect of the adaptation technique on the performance of the adapted recognizers

4.6.3. Effect of the size of the recognition dictionary

Finally, we were interested in studying to what extent the experimental results could be affected by increasing the size of the recognition vocabulary up to 149k words. This size is inversely proportional to the number of OOV words that can appear during the recognition process, and at the same time it is directly related to the probability of an uttered word being acoustically similar to others in the vocabulary.

The number of OOV words decrements drastically when the recognition vocabulary is very large. Hence, WER tends to decrease when such a dictionary is employed. As can be observed in Table 4.4, WER decreases 14.7% relative for Slovak when employing a vocabulary comprised of 149k words, compared with 10k, and 15.4% for Spanish.

Vocabulary size	10k	46k	85k	149k
Slovak	38.8	27.3	26.0	24.1
Spanish	33.0	28.4	28.0	27.9

Table 4.4. Effect of dictionary size on WER [in %] taking into account OOV words and speaker adaptation

The experimental results also showed that the smallest WERs were obtained after speaker adaptation, as observed in Figure 4.5. The objective was to check whether other languages could be efficiently recognized using the Czech recognizer, without using a specific vocabulary for recognition. Hence, Figure 4.5 only shows the results obtained by ignoring the OOV words.

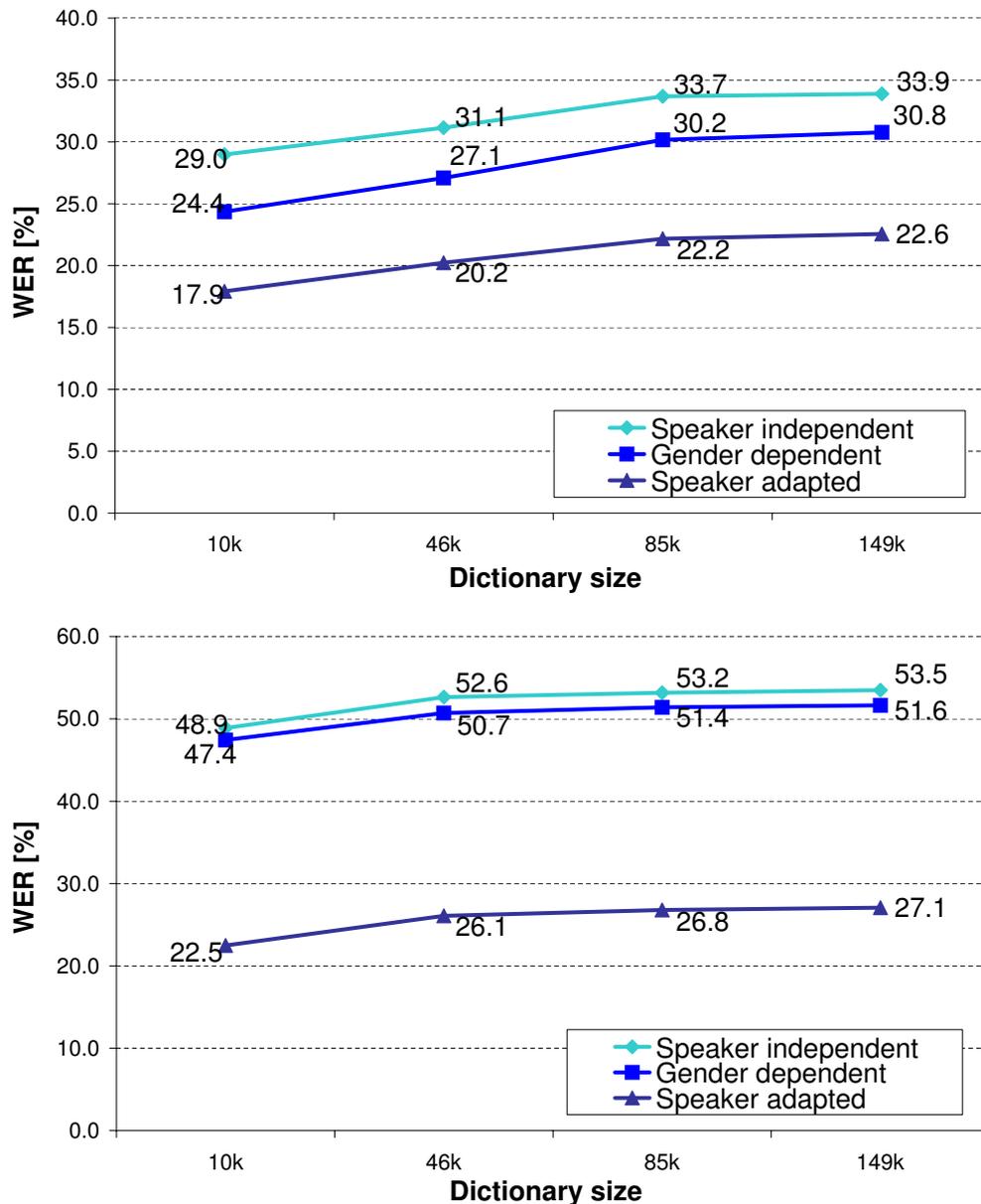


Figure 4.5. Performance of the Slovak (above) and Spanish (below) adapted recognizers

As can be observed, each adaptation type conveys an improvement. This improvement is larger for Spanish, for which it is necessary to use speaker-adapted models to similar results to those obtained for Slovak. As shown in Table 4.5, the relative reduction of WER from speaker-independent to speaker-adapted models, decreases when the accepted vocabulary is larger. This is due to an increment of the probability to find acoustically similar words in the dictionary. However, as can be observed in Figure 4.5, for increasingly larger recognition dictionaries it was found that the WER tends to establish around 23% for Slovak and 27% for Spanish. This shows that the proposed cross-lingual approach yields accuracy rates which are around 70% when adapting Czech to Spanish, i.e. two languages with very different origin. Whereas the accuracy rates are around 80% when adapting Czech to Slovak, i.e. two very phonetically similar languages.

Vocabulary size	10k	46k	85k	149k
Slovak	38.2	35	34.2	33.4
Spanish	54	50.4	49.6	49.4

Table 4.5. *Relative WER reduction [in %] yielded by the adaptation to speakers*

4.7 Conclusions

It has been presented in this chapter a cross-lingual adaptation of a previously-created Czech speech recognizer to Spanish and Slovak. Phonetic cross-linguality is a research area that is gaining increasing interest, especially because it enables resource sharing between languages and thus represents a feasible way of developing systems for minority languages or dialects. As rapid and low-cost development of speech-based systems is essential to foster portability, cross-language adaptation has arisen as one of the main challenges in the area (Gao et al., 2005). However, the state-of-the-art systems are based on complicated and very effort and time demanding linguistic and phonetic studies. Small effort has been devoted so far to study methods to carry out the cross-linguistic adaptation of a speech recognition environment (i.e. acoustic, lexical and linguistic models) in a cost-effective way.

It has been demonstrated that the adaptation of a speech recognizer to another language can be carried out in a straightforward way, employing a mapping between phonemes, and enhancing it with language and speaker adaptation procedures. Moreover, it has been shown that the proposed adaptation method can be used not only with phonetically similar languages, such as Czech and Slovak, but also with languages from very different families, like Czech (Slavic, Czech-Slovak) and Spanish (Italic, West-Iberian).

Several experiments have been carried out using MyVoice, a speech-based application designed for Czech handicapped people. Cross-lingual porting of voice operating systems for such a small group of target users, requires an investment that hardly can be paid back. However, our experimental results show that for a task involving a vocabulary of 432 commands, a 95.6% performance (4.4% WER) can be attained for Spanish and 97.5 (2.5% WER) for Slovak, employing the proposed cost-efficient procedure. Besides, for vocabularies up to 149k words, the proposed scheme yields around 72.9% accuracy (27.1% WER) for Spanish and 77.4% (22.6% WER) for Slovak.

*Ce n'est pas assez de compter les experiences,
il les faut poiser et assortir: et les faut avoir di-
gerees et alambiquees, pour en tirer les raisons
et conclusions qu'elles portent.*

Michel de Montaigne, Essais

5

Field evaluation of spoken dialogue systems

5.1 Introduction

As stated in the introduction of the Thesis, dialogue systems are becoming increasingly attractive for a wide range of applications (McTear, 2004; López-Cózar and Araki, 2005; Wahlster, 2006). In order to minimize costs and optimize results, there is a need for standard methods, architectures and criteria to test, compare and predict the performance and usability of the systems. Several initiatives have arisen since the late 80s to establish these methods. In the USA, the main funding institution for this kind of research is DARPA (Defense Advanced Research Projects Agency), with their project COMMUNICATOR (Walker et al., 2002a), which was aimed to cost-effective development of multimodal dialogue systems. This was achieved by using different plug-and-play components which were evaluated paying special attention to user satisfaction maximization. In Europe, the major institutions concerned with evaluation of dialogue systems have been COCOSDA (Coordinating Committee on Speech Databases and Speech I/O Systems Assessment), which focuses on obtaining corpora that can be shared to study evaluation criteria¹, EAGLES (1996) and DISC (1999). These last two international projects established some best practice guidelines for the development and evaluation of dialogue systems, both at system and component level.

¹<http://www.cocosda.org/>

These research efforts have successfully established a common background of criteria for quantitative evaluation. However, there is still no systematic understanding, nor consensus on the criteria that must be taken into account to optimize the usability of dialogue systems. Some projects have tried to address the problem of predicting system usability and user satisfaction from measurable performance criteria. This is the case of the PARADISE framework (Walker et al., 2000a), which has become one of the reference frameworks for system evaluation.

Because of the complexity and effort demanded by the application of the PARADISE framework, many approaches in the literature apply qualitative and quantitative measures separately. For example, Hartikainen et al. (2004) propose a methodology for subjective evaluation that has been used for evaluating the MUMS Multimodal Route Navigation System (Hurtig, 2004). Recently, the VIRTUAL CO-driver system (Geutner et al., 2002), the MASK multimedia service kiosk (Lamel et al., 2002) and the SAMMIE dialogue system (Becker et al., 2006), have been also evaluated only subjectively. Other authors, for example Robinson et al. (2006), evaluate their systems both with instrumentally-derived measures and quality judgments, but without establishing links between the different evaluation measures employed. In this chapter empirical results are obtained on the relationship between both types of criteria from the evaluation of the UAH system. This is done via correlation studies, which we believe are a reliable method that can be applied to both whole system and component level evaluation. However, when the statistical studies are carried out over a large number of metrics, there is a possibility that some of the findings are due to chance, and thus reliability and significance studies are also reported. This method has been applied successfully for the evaluation of other dialogue systems, e.g. BoRIS (Möller, 2005), yielding some interesting relationships between evaluation criteria.

However, results in the literature are usually based on restricted laboratory interactions, in which some users are asked to interact with the system in accordance with predefined scenarios. In some cases the users are also given evaluation questionnaires in which they express their personal opinion about different interaction aspects. The main disadvantage of this method is that the scenarios may differ from the tasks that a user would have selected in a non-predefined interaction. In contrast, field evaluation requires real users interacting with the final system in their appropriate environments.

Although as stated by Bernsen and Dybkjaer (2000), field tests can fail to be representative of the full functionality of the systems, we believe they offer the most realistic results and cover real user motivations. Field evaluations are not repeatable as the interaction context is highly variable. This is also their main advantage as they gather results from different users (difference in gender, voice, knowledge, experience using the system), who talk on different devices (mobile phones, usual phones or PCs), and in different environments (different noise conditions). As the results obtained from field tests are robust to this heterogeneity, they are more relevant at predicting the real behaviour of the systems. The contribution of the Thesis to the state-of-the-art system evaluation relies on obtaining new empirical evidence by means of a field study carried out employing our spoken dialogue system.

The chapter is structured as follows. Section 5.2 presents an overview of the main evaluation trends that can be found in the literature. Section 5.3 describes the computation of the evaluation criteria, distinguishing between interaction parameters and quality judgments. Section 5.4 presents the statistical studies carried out, whereas Section 5.5 discusses the experimental results obtained. Finally, Section 5.6 presents the conclusions obtained.

5.2 Related work

Evaluation of dialogue systems has been used in the literature for a wide range of purposes, for example, measuring the system's performance, comparing a system with its previous versions to measure the adequacy of changes, comparing different systems and predicting system behaviour.

Regardless of its purpose, evaluation can be carried out using "glass-box" or "black-box" approaches. The former permits access to internal details of the system to measure their contribution to the overall performance. The latter, treats the system as a black box so that the evaluation is based uniquely on the response of the system to the different user inputs. In both cases evaluation can be developed over a complete system or over an individual component. Usually, evaluation is carried out at the component level, the main working areas being the assessment of the speech recognizer, the speech and natural language understanding components, the dialogue manager and the speech output components.

Speech recognition performance is usually assessed in terms of automatically generated measures that calculate the number and importance of the recognition errors. Most of these measures are generalized in their use, as word error rate (WER) and word accuracy (WA), which are complementary in their use. WER is defined as the number of incorrectly determined words (calculated as the sum of the number of substitutions, deletions and substitutions made by the recognizer) divided by the total number of words, and WA can be calculated as $1 - WER$. Apart from the correct functioning of the speech recognition engine, when employing glass-box approaches, internal components of the speech recognizer can also be studied for example language models, lexicons or phonetic models. Additionally, having information about some of the recognizer components, can serve for prediction studies of the final recognition performance. For example in (Persia et al., 2007), they use performance of the source separation algorithm as a prediction of the speech recognition success in noisy environments. A best practices summary for evaluation of speech recognizers can be found in (Lamel et al., 2000a).

Despite the generalized use of the measures, the assessment of different speech recognition systems differs in several experimental conditions which make comparisons very difficult. Some of these influencing factors are the vocabulary characteristics (e.g. whether it is isolated or continuous, the vocabulary size, or the phonetic similarity of the words), the acoustic environment (e.g. noise levels), the transmission characteristics (e.g. transmission errors or signal levels) and the speaker characteristics (e.g. age, gender or cultural background). All these factors have to be taken into account when trying to compare the evaluation results of different recognizers, which makes this task very complicated. Recently, some studies have focused on how to overcome these difficulties by creating corpora that can be shared between the scientific community. The objective is to provide a common testbed for evaluating and subsequently comparing speech recognition and enhancement algorithms, for example the NOIZEUS speech corpus (Hu and Loizou, 2007).

For the assessment of speech and natural language understanding engines, some measures similar to WER have been used (Gupta et al., 2006). For example slot error rate (defined as the number of incorrect slots divided by the number of slots), the update precision (defined as the number of correctly updated slots divided by the number of updated slots) and the concept error rate (CER), which is defined as the number of incorrect slots

divided by the number of filled slots. A more detailed list of metrics can be found in (Higashinaka et al., 2004). However, evaluating understanding modules requires a higher involvement of experts in judging whether the results are correct, and thus there is a need for a higher level of quality judgment measures. One of the seminal studies in this area was developed inside the TSNLP (Test Suites for Natural Language Processing) project (Lehmann et al., 1996). Furthermore, evaluating natural language understanding components is very domain-dependent, as they strongly rely on the semantics of the task and the detail to which it is done (i.e. the number of concepts which are used), and thus it is difficult to compare evaluation results between systems.

Comparison is even more complicated in the case of the speech output components, because their evaluation is almost entirely done using judgment measures about their quality, thus obtaining highly subjective evaluation results. Measures used mainly describe the quality and/or naturalness of the speech sounds, as well as the comprehension of the system message by the user. The results vary greatly depending on the users, that is why most studies are centred in specific population sectors, for example the elderly in (Lines and Hone, 2002). A comprehensive list of intelligibility, prosody and overall quality tests to evaluate speech synthesized outputs can be found in (Gibbon et al., 1997).

Dialogue management is usually evaluated in terms of the quality of the user-system interaction: adequacy of system responses, feedback strategies, duration of the dialogue, average number of turns, adequacy of the initiative, adequacy of confirmation strategies or ability to solve misunderstanding situations. The DISC (1999) work group proposes 6 main groups that gather the criteria to be taken into account when evaluating dialogue managers: correct management of knowledge about the current dialogue context, mapping from the semantically significant units in the user's most recent input, analysing the user's specific contribution, generating of output to the user, specific issues of dialogue management evaluation (e.g. feedback strategies) and global issues of dialogue system evaluation (e.g. time for task completion). It is difficult to separate the evaluation of the dialogue manager from the rest of the system, specially in quality judgments, this is why some authors propose to use manually corrected input to the dialogue manager to be able to have some "gold standard" to represent the dialogue

manager behaviour in isolation. This baseline can be used subsequently for comparison with real situations in which the dialogue manager is affected by errors in the recognition or understanding modules (Roque et al., 2006a). Additionally, the dialogue manager makes intensive use of domain knowledge and has to be able to build and adapt to the interaction context. Thus, some authors like (Hanna et al., 2007), have studied how to carry out evaluation of dialogue managers in terms of modification of the domain-specific expertise and maintenance and reuse of the already existing context knowledge and discourse management behaviour.

There is not a consensus in the literature about the terminology to use for categorizing all the described evaluation criteria. Traditionally, authors have differentiated between objective and subjective evaluation criteria. The former takes into account measures computed from system performance features such as word error rate (WER). The latter considers measures that judge some property, for example intelligibility of the synthesized speech. This notation has been widely used in previous studies, for example, Larsen (2003), Minker et al. (2004b) and Robinson et al. (2006). However, as argued by Möller (2005), human subjects are always involved in determining the systems' performance. In the so-called objective measures human expert evaluators are often used, for example to calculate WER, experts have to compare real user input with the recognizer output). Thus, Möller (2005) proposes to differentiate between quality judgments (subjective), interaction parameters (which can be instrumentally measured or expert derived) and quality predictions (which can be instrumentally derived). This chapter will be focused on the first two categories.

There have been several attempts to create a full list of criteria to be used for evaluation by employing interaction parameters, quality predictions and quality judgments. For example, Dybkjaer and Bernsen (2000) propose a list of 15 criteria to guarantee system usability: adequate use of modalities, accurate input recognition, flexibility of the accepted vocabulary, system voice quality, adequate response generation, adequate domain coverage, and user satisfaction, among others. The Expert Advisory Group on Language Engineering Standards (EAGLES, 1996), proposed quantitative (e.g. system response time) and qualitative measures (e.g. user satisfaction), that were applied and interpreted following an innovative framework. This framework provided guidelines on how to carry out the evaluation and how

to make results available in such a way that they could be easily interpretable and comparable. In the DISC (1999) project there were other best practice guidelines that completed the EAGLES proposal using life cycle development methodologies. Other authors have focused on how to obtain and study speech corpora to compute evaluation measures. These are frequently large corpora extracted from system usage, or from human-to-human dialogues. In the latter case, human behaviour can be used as a baseline to compare with the system behaviour (Paek, 2001). For example, the project EVALDA (Devillers et al., 2004) focuses on evaluation campaigns that consider various aspects of natural language interaction. One of them is the MEDIA campaign, which evaluates the interaction between users and dialogue systems. Their evaluation methodology employs test sets obtained from real corpora along with the commonly used evaluation criteria. Degerstedt and Jönsson (2006) proposed the LINTEST tool to carry out evaluation of dialogue systems using the JUNIT corpus. A very detailed review of the most relevant efforts on generalization of evaluation criteria and practices can be found in Dybkjaer et al. (2004) and in López-Cózar and Araki (2005), whereas Möller et al. (2007) present a review of the de-facto criteria extracted from all these studies and an example of their usage to evaluate a particular dialogue system.

As commented above, PARADISE (Walker et al., 1998b) is the most widely embraced evaluation method proposed so far to specify the relative contribution of various factors to the overall system performance. This method models performance as a weighted function of: task success (exact scenario completion), dialogue efficiency (task duration, system turns, user turns, total turns), dialogue quality (word accuracy, response latency) and user satisfaction (sum of TTS performance, ease of task, user expertise, expected behaviour, future use). More recently, PARADISE has been used to develop models of user satisfaction prediction, again based on the weighted linear combination of different measures (Walker et al., 2000b). The goal of this evaluation method is to maximize user satisfaction by maximizing task success and minimizing interaction costs as shown in Figure 5.1. These costs are quantified using different efficiency and quality measures. The weights of each measure are computed via a multivariable linear regression considering user satisfaction as the dependent variable and task success, efficiency and quality measures as independent variables. Recently, the PARADISE

framework has been enhanced to enable evaluation of multimodal dialogue systems. For example, it was used in the SmartKom Project, as a basis for the so-called PROMISE framework (Beringer et al., 2002).

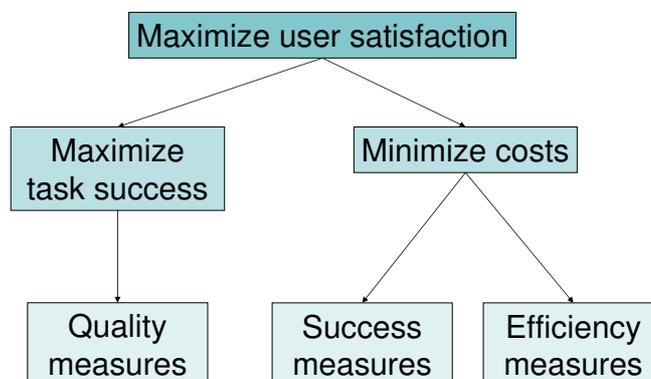


Figure 5.1. *PARADISE architecture model*

The application of PARADISE to evaluate a dialogue system requires dialogue corpora extracted from controlled experiments in which subjects have to evaluate satisfaction on a scale after they have interacted with the system. This approach has been successfully used for evaluating and comparing eight COMMUNICATOR systems (Walker et al., 2002b,a), firstly in controlled laboratory experiments, and secondly in a less restricted context where the systems were accessible on the phone. Strictly, this second evaluation was not an open field study because the authors had control over the users, who were specifically recruited and assigned to the different systems. Nevertheless, the tasks they had to complete were not predefined in all cases. A similar approach was employed in the ARISE project (den Os et al., 1999), where evaluation was based on the responses of subjects who either called a dialogue system from home or interacted with it in the laboratory. In either case, the tasks to be carried out by users were predefined (Sanderman et al., 1998).

It can be also found in the literature a distinction between “internal” and “external” tests regarding whether they were carried out by users from the development team of the dialogue system (internal evaluation) or by users who did not have any previous knowledge about the system (external evaluation). However this is not equivalent to “field” vs. “laboratory” studies distinction, as external tests may involve using predefined scenarios. For example, Rajman et al. (2004) propose a Rapid Dialogue Prototyping

Methodology to produce, for any given application, a quickly deployable dialogue-driven interface which can be later enhanced through an iterative Wizard-of-Oz process. To refine the dialogue models developed using this methodology, the authors propose to use an internal and an external test. The internal test is used to further adapt the prototype and its successive modifications. The external test is employed for the final evaluation of the resulting dialogue interface. In both cases the evaluation is carried out in the form of a satisfaction questionnaire which is submitted to the users after they have interacted with the prototype, on the basis of a set of predefined scenarios involving specific contexts for a restaurant search.

To study the implications of using field tests, some authors have focused on non-restricted evaluation studies. This is the case of the Let's Go system (Raux et al., 2003), which was evaluated using interactions of real users that phoned the system to get information about bus schedules. The evaluation was carried out by reporting results of interaction parameters (Raux et al., 2006). Unfortunately, although these parameters are relatively easy to compute, they do not provide sufficient information on quality. Qualitative judgments, on the other hand, are difficult to extract and compare when they are related to subjective opinions. Only in a few cases, performance parameters which can be measured quantitatively are also able to express quality. Our work focuses on using both quantitative and qualitative de-facto standard measures (Möller et al., 2007) in a field study, to evaluate our spoken dialogue system, which is described in Chapter 2. Our main objective is to empirically obtain relationships between these measures by employing statistical significance studies. Similar methods have been widely used in the area of systems' acceptance, more specifically for predicting the adoption of new technologies, e.g. in risk studies by investing companies. One of the most used models is the Technology Acceptance Model, which relates several user judgments criteria with the final adoption of the technologies by users (Legris et al., 2003). However, no quantitative parameters are considered in this model. In the area of dialogue systems, very few authors have exploited correlation studies to measure such relationships, for example Litman and Pan (2002), Möller (2005) and Schiel (2006), who applied them to controlled laboratory studies.

5.3 Evaluation criteria

The UAH system evaluation was carried out both with interaction parameters and quality judgments. Interaction parameters were employed to measure the system performance (e.g. number of errors made by the speech recognizer), and the dialogue course (e.g. duration of the dialogue or number of turns). These measures allowed to carry out different studies about performance and reliability of the system as well as discovering interaction points which can be improved. Although interaction parameters are a good indicator of the quality of the evaluated interaction, they do not necessarily provide reliable information about user satisfaction (López-Cózar and Araki, 2005). Thus, it is necessary to carry out a qualitative judgment evaluation to register users' opinions about these aspects of the interaction. In the experiments presented in the chapter, the subjective evaluation was carried out by employing user tests.

5.3.1. Interaction parameters

To compute the values for the interaction parameters, the UAH corpus has been used. This corpus consists of 85 dialogues and 422 user turns, with an average of 5 user turns per dialogue. Each dialogue was automatically annotated with two timestamps, corresponding to the call starting and ending times respectively. Each user utterance was stored in .wav format along with information about the recording starting time, the previous system turn, and the speech recognition result, which included confidence scores attached to the recognized words.

Then, it was manually annotated whether each utterance was correctly understood by the system, regardless of the speech recognition errors. For example, if in response to the system prompt: "What type of information do you want?", the user answered: "I want information about a subject", but the recognition result was: "Information about subjects", there are three deletions and one substitution. Regardless of these errors, the utterance was correctly understood by the system, as the semantic values returned by the speech recognition grammar were correct. Hence, the annotator tagged the utterance as "correctly understood".

At the dialogue level, the annotator registered the gender of the speaker, whether the dialogue was complete (i.e. whether the user did not hang up before finishing the dialogue) and whether the dialogue was successful. As it was a field study, there were no predefined tasks for the users to accomplish. Thus, a strategy had to be defined to consider dialogue success. More specifically, it was considered that the dialogues were successful when the user obtained the information he requested.

All the annotations were stored in a database from which the values for the interaction parameters were automatically computed. For example, dialogue duration was computed from the timestamps, and the number of confirmation turns was computed using the information about the previous system turn. Table 5.1 sets out the interaction parameters that were used in the experiments, while Figure 5.2 shows the parameter computation for a sample dialogue.

Parameter	Description	Human annotation needed
Task success	Binary value which indicates whether the user obtained from the system the information he requested	Yes
Dialogue completion	Binary value which indicates whether the user waited until the end of the dialogue before hanging up	Yes
Dialogue duration	Duration of the dialogue in seconds	No
Number of user turns	Number of user turns in the dialogue	No
Average number of words per turn in the dialogue	Avg. number of words along all the user utterances in the dialogue	Yes
WER	Number of incorrectly recognized words divided by the total number of words uttered by the user in the dialogue	Yes
Average recognition confidence	Average confidence score of the words in all the speech recognition results in the dialogue	No
% correctly understood utterances	Percentage of correctly understood utterances with respect to the total number of utterances in the dialogue	Yes
Number of confirmation turns	Number of times the system asked for explicit confirmation during the dialogue	No

Table 5.1. *Interaction parameters employed*

DIALOGUE

S1> Welcome to the UAH system. How can I help you?

U1> I want information about a professor.

S2> Please say the name of the professor.

U2> Peter Smith.

S3> Did you say Peter Smint?

U3> No, Peter Smith.

S4> What information do you want about Professor Smith?

U4> I would like to know his tutoring timetable.

S5> Professor Smith's tutoring hours are on Mondays and Tuesdays from 10 to 12. Do you need any other information?

U5> No, thanks.

S6> Thank you for using the UAH system. Have a nice day!

PHASE 1 :
Automatic call recording and information registration

2006_04_07, 15:02:48 – 15:03:59, Male

Timestamp	Previous system turn	Recognized sentence	Recognition confidence
U1 15:03:09	Initial_menu	I want information about professors	0.89
U2 15:03:20	Ask_prof_name	Peter Smint	0.4
U3 15:03:36	Conf_prof_name	No Peter Smith	0.9
U4 15:03:43	Prof_menu	I would like to know his tutoring timetable	0.85
U5 15:03:57	More_info	No thanks	1

PHASE 2 : Interaction parameters computation

2.1 : Annotation by human expert

Task success	1
Dialogue completion	1
Number of words per turn	U1 – 6 U2 – 2 U3 – 3 U4 – 8 U5 - 2
Number of insertions, deletions and substitutions per turn	U1 – 1 deletion, 1 substitution U2 – 1 substitution
Correctness of the semantic interpretation	U1 – 1 U2 – 0 U3 – 1 U4 – 1 U5 – 1

2.2: Automatic computation

Dialogue duration	71
Number of user turns	5
Average recognition confidence	0.81
Number of confirmation turns	1
Average number of words/turn	4.2
WER	0.14
%correctly understood utterances	0.8

Figure 5.2. Example of the computation of the interaction parameters for an UAH dialogue

5.3.2. Quality judgments

The interaction with the UAH system starts with a welcome message in which the system introduces itself, and asks the user to visit a web page where he can complete a questionnaire with his opinion about the system performance. To be able to link the results of this test with the recordings of the user-system interaction, the user is provided with a dialogue identification number. This number is requested in the questionnaire along with the date he made the telephone call to the system and an approximate time for the start of the interaction.

The English translation of the questionnaire is as follows:

Q1. State on a scale from 1 to 5 your knowledge about new technologies for information access. (1 = "Low", 5 = "High")

Q2. State on a scale from 1 to 5 your previous experience using telephone-based dialogue systems. (1="Low", 5="High")

Q3. How many times have you used the UAH system before?

- I have not used it before.
- times.

Q4. How well did the system understand you?

- Extremely bad.
- Bad.
- Fair.
- Good.
- Excellent.

Q5. How well did you understand the messages generated by the system?

- Extremely bad.
- Bad.
- Fair.
- Good.
- Excellent.

Q6. In your opinion the interaction was:

- Very slow.

- Slow.
- Adequate.
- Fast.
- Very fast.

Q7. Correcting the errors made by the system was:

- Extremely difficult.
- Difficult.
- Easy.
- Extremely easy.
- The system made no errors.

Q8. Was it easy for you to get the information that you requested?

- No, it was impossible.
- Yes, but with great difficulty.
- Yes, but with certain difficulties.
- Yes, it was easy.
- Yes, it was extremely easy.

Q9. Are you satisfied with the system performance?

- Not satisfied at all.
- Not very satisfied.
- Indifferent.
- Satisfied.
- Very satisfied.

Q10. Were you sure about what to say to the system at every moment?

- No, never.
- No, almost never.
- Sometimes.
- Yes, almost always.
- Yes, always.

Q11. Do you believe the system behaved similarly as a human would do?

- No, never.
- No, almost never.
- Sometimes.
- Yes, almost always.
- Yes, always.

The answers to each question were encoded and appropriately saved in the interactions database. All the answers excepting those corresponding to Q3 were assigned a numeric value between one and five (in the same order as they appear in the questionnaire). The values by default were: Q1=1, Q2=1, Q3=1, Q4=3, Q5=3, Q6=3, Q7=5, Q8=3, Q9=3, Q10=3, Q11=3. From the results of the test, the measures listed in Table 5.2 were extracted.

The first three measures listed in Table 5.2 are not quality judgments, but information about users. With the help of these questions, we intended to obtain an approximate idea of the users' background. However, as the UAH users were mainly students and professors of our Faculty, knowledge about new technologies for information access was high in almost all cases, as it is shown in Figure 5.3. Only 36% of our test participants were women.

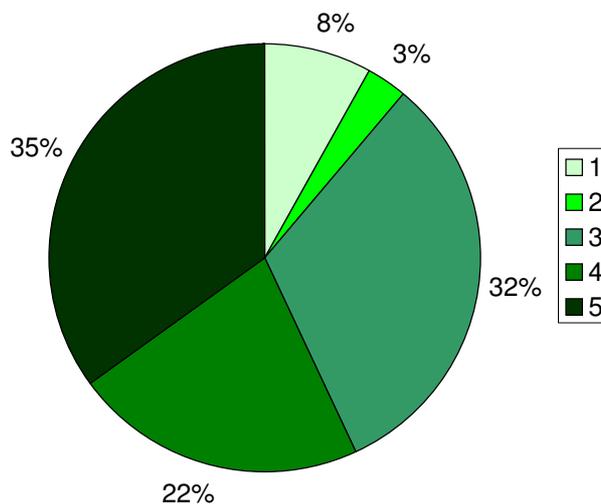


Figure 5.3. *Users' knowledge about new technologies for information access (1 = Low, 5 = High)*

Parameter	Question from which it is extracted
Knowledge about new technologies for information access	Q1
Knowledge about dialogue systems	Q2
Experience using the UAH system	Q3
Perceived extent to which UAH understands the user	Q4
Perceived extent to which the user understands UAH	Q5
Perceived interaction speed	Q6
Perceived presence of errors made by UAH	Q7
Perceived ease of UAH error correction	Q7
Perceived easy of obtaining the requested information	Q8
User satisfaction	Q9
Extent to which the user knew what to say at each moment of the interaction	Q10
Perceived human-like behaviour of the UAH system	Q11

Table 5.2. *Perceived quality and user profile parameters employed*

As our experiments were based on calls made by users who phoned the system on their own initiative, we think that the results obtained are very realistic, given that the interaction was based on a real need of the users. Besides, dialogues were more heterogeneous as they take place in different contexts. The disadvantage of this approach was that, although the users were encouraged to answer the questionnaires, some of them did not do it, and thus there were no quality judgments for all the recorded dialogues. Specifically, only 37 of the 85 dialogues have subjective measures along with the objective ones. Figure 5.4 shows the demographic data of the two types of users: those who answered the subjective test, and those who did not.

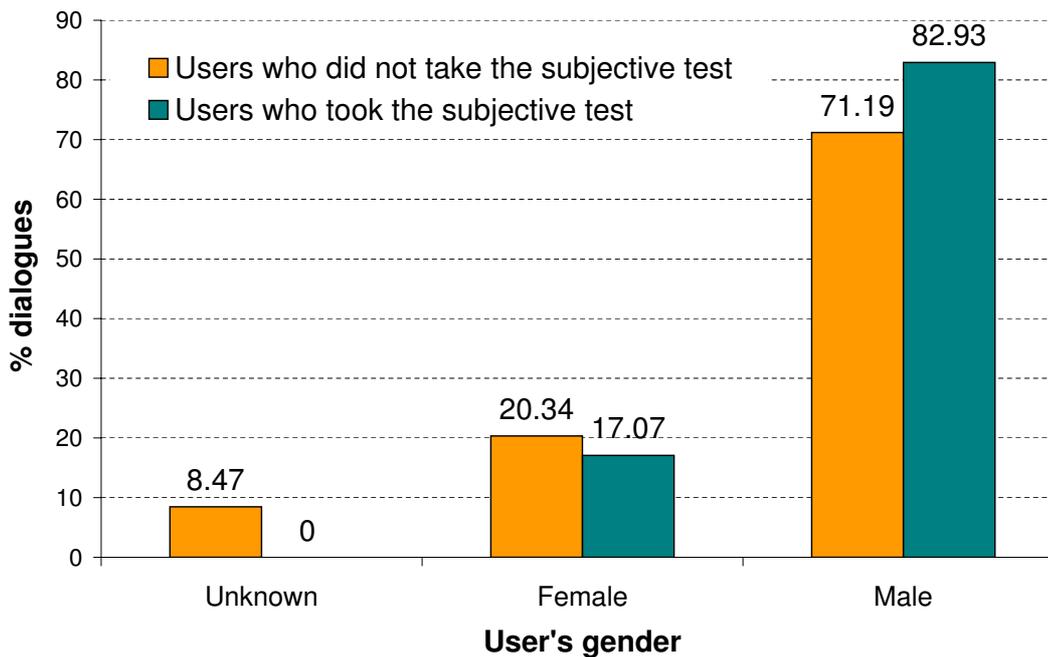


Figure 5.4. Demographic data for the different user types

As can be observed, from the dialogues that corresponded to users who did not fill in the questionnaire, 8.47% were annotated with an unknown gender of the speaker. This is because these users hung up after the first prompt of the system and said nothing in response. The first system prompt clearly stated that the user was about to talk to an automatic system, and that the call was going to be recorded for research purposes. Hence, two plausible reasons why some users hung up before their first turn are that they did not feel confident in talking to a computer, and that they were not happy with having their interactions recorded.

The descriptive statistics of all the parameters regarding the type of users involved are shown in Table 5.3, where the minimum, maximum and range values of all the measures used in our study are indicated. Section 5.5.1 presents a detailed study of the differences in performance and perceived quality between the interactions of these two user groups.

Parameter	User type	Range	Min.	Max.	Avg.	Typ. Dev.	Variance
Knowledge about new technologies for information access	Subj. test	4	1	5	3.77	1.14	1.30
Knowledge about dialogue systems	Subj. test	4	1	5	3.23	1.28	1.65
Experience using the UAH system	Sub. test	9	1	10	2.80	3.20	10.22
Perceived extent to which UAH understands the user	Subj. test	4	1	5	3.69	1.25	1.57
Perceived extent to which the user understands UAH	Subj. test	2	3	5	4.37	0.69	0.48
Perceived interaction speed	Subj. test	3	1	4	2.71	0.62	0.39
Perceived presence of errors made by UAH	Subj. test	1	0	1	0.54	0.50	0.25
Perceived ease of UAH error correction	Subj. test	3	1	4	2.47	0.90	0.82
Perceived easy of obtaining the requested information	Subj. test	4	1	5	3.37	1.437	2.06
User satisfaction	Subj. test	4	1	5	3.63	1.09	1.18
Extent to which the user knew what it was expected from him at each point of the dialogue	Subj. test	3	2	5	4.29	0.893	0.798
Perceived human-like behaviour of the UAH system	Subj. test	4	1	5	3.57	1.04	1.08
Task success	Subj. test	1	0	1	0.77	0.43	0.18
	No subj. test	1	0	1	0.46	0.50	0.25
Dialogue completion	Subj. test	1	0	1	0.74	0.44	0.20
	No subj. test	1	0	1	0.36	0.48	0.23
Dialogue duration	Subj. test	153	21	174	96.66	37.06	1373.70
	No subj. test	297	0	297	90.14	64.65	4179.88
Number of user turns	Subj. test	9	1	10	5.34	2.26	5.11
	No subj. test	16	1	17	4.7	3.94	15.52
Avg. words per turn	Subj. test	3	1	4	1.81	0.69	0.48
	No subj. test	4.33	0	4.33	1.73	0.78	0.61
WER	Subj. test	0.67	0.00	0.67	0.19	0.18	0.03
	No subj. test	0.83	0	0.83	0.25	0.28	0.05
Avg. recognition confidence	Subj. test	0.16	0.82	0.98	0.93	0.04	0.002
	No subj. test	0.23	0.77	1	0.93	0.05	0.003
% correctly understood utterances	Subj. test	0.50	0.50	1	0.95	0.12	0.15
	No subj. test	0.74	0.33	1	0.89	0.19	0.04
Number of confirmation turns	Subj. test	2	0	2	0.80	0.63	0.40
	No subj. test	3	0	3	0.62	0.88	0.77

Table 5.3. Descriptive statistics of the criteria used

5.4 Statistical studies employed for evaluation

In order to find relevant relationships between the criteria used, all the variables were correlated, obtaining the absolute value of the *Pearson correlation coefficient*. However, the value of the correlation coefficient by itself was not

enough to obtain reliable results, as it was also necessary to know the probability of obtaining the results by chance. This was done by computing the significance (or *p-value*) of each correlation coefficient. If the significance level was very small (less than 0.05) then the correlation was significant and the two criteria were considered linearly related.

As most of the variables were inter-correlated, the effect that each criterion had on the significance of the relationships between the rest has been studied. It is possible that two criteria are correlated just because they are both affected by a third one. Thus, when eliminating the effect of this criterion, they would not be significantly correlated. To study the relationships in isolation, eliminating the effect of the rest of the criteria, the *partial correlation coefficients* were computed along with their significance.

The *Pearson correlation coefficient* is suitable for *scale* variables, whose values represent ordered categories with meaningful metrics, such as dialogue duration in seconds, so that distance comparisons between the values are appropriate. However, not only scale variables were used, but also *ordinal* and *dichotomous* variables (a classification can be found in Table 5.4). The values of the ordinal variables represent categories with an intrinsic rating, such as the perceived quality parameters described in Section 5.3.2. Dichotomous variables, such as “task success” or “dialogue completion”, can only have two values (0 or 1 in our case). Thus, in order to obtain reliable results, contingency tables were built for the ordinal criteria. These Tables allow to study these variables and discover associations between them. To measure the strength of their relationships, the *Kendall’s Tau-b* and the *Spearman’s Rho* coefficients were used. The interpretation of these coefficients is equivalent to that of the *Pearson coefficient*. However, as they are based on the ordinal properties of the data, their values and significances may not be the same.

Additionally, analyses of variance (ANOVA) were carried out. Essentially, ANOVA models try to describe a dependent variable as the result of the weighted sum of several factors. Specifically, one-way ANOVA was used, in which there is only one independent variable, and computed the *F coefficient*. When *F*’s critical level is below 0.05, it is possible to discard the average equality and conclude that not all the poblational averages that are being compared are equal. *Eta square* was also obtained, which is an estimation of the degree to which each factor affects the dependent variable.

To obtain more information on which to base our interpretations, especially for the case of dichotomous variables, *Phi* and *Cramer's V* coefficients were also calculated, which allow to contrast the independence hypothesis in contingency tables.

Parameter	Type
Knowledge about new technologies for information access	Ordinal
Knowledge about dialogue systems	Ordinal
Experience using the UAH system	Ordinal
Perceived extent to which UAH understands the user	Ordinal
Perceived extent to which the user understands UAH	Ordinal
Perceived interaction speed	Ordinal
Perceived presence of errors made by UAH	Dichotomous
Perceived ease of UAH error correction	Ordinal
Perceived easy of obtaining the requested information	Ordinal
User satisfaction	Ordinal
Extent to which the user knew what it was expected from him at each point of the dialogue	Ordinal
Perceived human-like behaviour of the UAH system	Ordinal
Task success	Dichotomous
Dialogue completion	Dichotomous
Dialogue duration	Scale
Number of user turns	Scale
Avg. words per turn	Scale
WER	Scale
Avg. recognition confidence	Scale
% correctly understood utterances	Scale
Number of confirmation turns	Scale

Table 5.4. *Type of variables used for the statistical studies*

All the experiments were carried out using the SPSS 14 predictive analysis software². For the experiments in which the aim was to obtain important relationships between all the evaluation criteria including both interaction parameters and quality judgments, the 37 dialogues in which the users answered the subjective test were used. For the experiments in which the possible reasons for the users to take the test or not were studied, both types of dialogues were used (85 in total).

²Statistical Product and Service Solutions - <http://www.spss.com/>

5.5 Evaluation results

This section presents a summary of the numeric results obtained from the statistical studies. Table 5.5 shows a summary of the results obtained with the partial correlations. For reasons of space, we have not reported all the 21 partial correlations tables with their numeric values. Instead, there are only reported all the significant correlations found between all the tables, along with the number of control criteria for which they were significant (i.e. the number of partial correlation tables in which the relationship was significant). For each pair of criteria, Table 5.6 sets out the *Pearson correlation* coefficient and its significance level. Significance levels below 0.05 are marked in blue, those below 0.01 are marked in orange, and non-significant relations are left white.

As can be observed, there were no significant relations regardless of the control criteria used (i.e. none of them appeared in the 21 tables). In fact, the best case was achieved when the relationship between two criteria was shown to be significant when eliminating the effect of 17 of the 21 variables. This showed that all the variables were deeply related. Finally, in Table 5.7 there is a summary of the results for the *Tau-b* and *Rho* coefficients, only emphasizing the relations for which significance differs from those obtained in the Pearson correlation studies. In the following sections the main findings derived from these results are discussed and interpreted.

Criteria relationship		Partial correlation tables in which it was significant
Perc. ease of obtaining the requested information	Perc. extent to which UAH understands the user	17
Perc. human-like behaviour of the UAH system	Perc. extent to which the user understands UAH	17
Knowledge about dialogue systems	Knowledge about new technologies for information access	17
Dialogue duration	Number of user turns	16
Number of confirmation turns	Number of user turns	16
% correctly understood utt.	WER	16
Avg. recognition confidence	WER	16
Task success	Perc. ease of obtaining the requested information	16
Perc. ease of obtaining the requested information	User satisfaction	15
Perc. human-like behaviour of the UAH system	User satisfaction	15
Avg. recognition confidence	% correctly understood utt.	15
Task success	User satisfaction	15
Dialogue completion	Task success	14
Dialogue completion	User satisfaction	14
Perc. ease of UAH error correction	Perc. ease of obtaining the requested information	14
Perc. ease of UAH error correction	Perc. extent to which UAH understands the user	14
Task success	Perc. extent to which UAH understands the user	14
Perc. extent to which UAH understands the user	User satisfaction	14
WER	Avg. words per turn	14
Perc. ease of UAH error correction	User satisfaction	13
Perc. ease of obtaining the requested information	Dialogue completion	13
Perc. ease of obtaining the requested information	Perc. human-like behaviour of the UAH system	13
Dialogue completion	Perc. ease of UAH error correction	12
Dialogue completion	Perc. extent to which UAH understands the user	12
% correctly understood utt.	User satisfaction	12
Perc. ease of UAH error correction	Task success	12
Perc. human-like behaviour of the UAH system	Task success	12
Perc. human-like behaviour of the UAH system	Perc. extent to which UAH understands the user	12
Perc. ease of UAH error correction	Perc. human-like behaviour of the UAH system	11
Dialogue duration	Perc. ease of obtaining the requested information	10
Dialogue duration	Task success	10
Dialogue duration	Perc. extent to which UAH understands the user	10
Dialogue duration	User satisfaction	10
Task success	Number of user turns	10
User satisfaction	Perc. extent to which the user understands UAH	10
Dialogue completion	Dialogue duration	9
Number of user turns	User satisfaction	7
Dialogue completion	Perc. human-like behaviour of the UAH system	6
Number of confirmation turns	Dialogue duration	5
Perc. ease of obtaining the requested information	Perc. extent to which the user understands UAH	5
Number of user turns	Perc. ease of obtaining the requested information	4
User satisfaction	Avg. words per turn	4
Number of confirmation turns	Avg. recognition confidence	3
Dialogue duration	UAH usage	2
Dialogue completion	Number of user turns	1
Dialogue completion	UAH usage	1
Dialogue duration	Perc. ease of UAH error correction	1
Dialogue duration	Perc. human-like behaviour of the UAH system	1
Number of confirmation turns	UAH usage	1
Number of confirmation turns	WER	1
Number of user turns	WER	1
Perc. ease of obtaining the requested information	% correctly understood utt.	1
Perc. ease of obtaining the requested information	Avg. recognition confidence	1
Avg. recognition confidence	Task success	1
Avg. recognition confidence	UAH usage	1
Avg. recognition confidence	User sure	1
Task success	% correctly understood utt.	1
Perc. extent to which UAH understands the user	Number of user turns	1
Perc. extent to which UAH understands the user	% correctly understood utt.	1
UAH usage	WER	1
Knowledge about dialogue systems	Perc. ease of obtaining the requested information	1
Knowledge about dialogue systems	Perc. extent to which UAH understands the user	1
User satisfaction	WER	1

Table 5.5. Significant partial correlations

Criterion 1	Criterion 2	Pearson	Tau-b	Rho
Perc. extent to which UAH understands the user	DS knowledge	0.265 (0.124)	0.276 (0.057)	0.336 (0.049)
Perc. extent to which UAH understands the user	Perc. interaction speed	0.334 (0.050)	0.268 (0.077)	0.299 (0.081)
Perc. extent to which UAH understands the user	Dialogue completion	0.485 (0.003)	0.390 (0.013)	0.426 (0.011)
Perc. extent to which UAH understands the user	Dialogue duration	0.433 (0.009)	0.209 (0.111)	0.278 (0.105)
Perc. extent to which UAH understands the user	Number of user turns	0.340 (0.046)	0.157 (0.255)	0.197 (0.257)
Perc. extent to which UAH understands the user	Number of confirmation turns	0.363 (0.032)	0.291 (0.054)	0.335 (0.049)
Perc. extent to which the user understands UAH	Task success	0.498 (0.002)	0.408 (0.013)	0.424 (0.011)
Perc. human-like behaviour of the UAH system	Perc. interaction speed	0.443 (0.008)	0.355 (0.019)	0.389 (0.021)
Perc. human-like behaviour of the UAH system	Perc. ease of UAH error correction	0.601 (0.006)	0.474 (0.018)	0.523 (0.022)
Dialogue completion	Perc. ease of UAH error correction	0.623 (0.004)	0.559 (0.011)	0.602 (0.006)
Task success	Perc. ease of UAH error correction	0.623 (0.004)	0.559 (0.011)	0.602 (0.006)
Perc. easy of obtaining the required information	Perc. presence of errors made by UAH	-0.326 (0.056)	-0.337 (0.033)	-0.365 (0.031)
User sure	Perc. presence of errors made by UAH	-0.419 (0.012)	-0.429 (0.008)	-0.454 (0.006)
User sure	User satisfaction	0.385 (0.022)	0.291 (0.054)	0.316 (0.064)
Dialogue duration	User satisfaction	0.375 (0.026)	0.245 (0.065)	0.310 (0.070)
% correctly understood utt.	User satisfaction	0.495 (0.002)	0.223 (0.151)	0.248 (0.151)
Perc. easy of obtaining the requested information	Dialogue completion	0.524 (0.001)	0.384 (0.015)	0.416 (0.013)
Dialogue duration	Dialogue completion	0.462 (0.005)	0.350 (0.014)	0.421 (0.012)
Number of user turns	Dialogue completion	0.354 (0.037)	0.274 (0.068)	0.313 (0.067)
Perc. easy of obtaining the requested information	Dialogue duration	0.348 (0.040)	0.151 (0.253)	0.225 (0.194)
Dialogue duration	Task success	0.475 (0.004)	0.362 (0.011)	0.435 (0.009)
Dialogue completion	Number of user turns	0.354 (0.037)	0.274 (0.068)	0.313 (0.067)
User sure	WER	-0.426 (0.011)	-0.337 (0.017)	-0.388 (0.021)
Perc. easy of obtaining the requested information	% correctly understood utt.	0.350 (0.040)	0.244 (0.113)	0.262 (0.129)

Table 5.7. Significance variations between the Pearson, Chramer's Tau-b and Spearman's Rho coefficients

5.5.1. Impact of the interaction performance on the user decision to answer the subjective test

As was described in Section 5.3.2, not all the users answered the subjective test from which the perceived quality criteria were computed. In order to study if there were some interaction parameters that influenced the users' decision to answer the test, we introduced a dichotomous variable indicating whether the user answered the test or not, and carried out Pearson correlation and ANOVA studies to find its relationship with the interaction parameters. Table 5.8 shows the results obtained.

Relationship	ANOVA F (Sig)	Eta square	Pearson (Sig)
Task success	7.156(0.009)	0.079	0.282(0.009)
Dialogue completion	7.775(0.007)	0.086	0.293(0.007)
Dialogue duration	0.245 (0.622)	0.003	0.054 (0.622)
Number of user turns	0.729 (0.396)	0.009	0.093 (0.396)
Avg. recognition confidence	0.122 (0.728)	0.001	-0.159 (0.150)
WER	2.107 (0.150)	0.025	0.010 (0.927)
Avg. words per turn	0.008 (0.927)	0.000	0.038 (0.728)
% correctly understood utt.	3.759 (0.056)	0.043	0.208 (0.56)
Number of confirmation turns	0.592 (0.447)	0.18	0.133 (0.447)

Table 5.8. Significance of the relationship between “The user taking the subjective test” and the interaction parameters

The only relations that were shown to be significant for the “user taking the subjective test” were with the “dialogue completion” and the “task success” metrics. These are two criteria that were also very significantly correlated with each other, with an *ANOVA F* of 180.159, and a 0.000 significance. *Eta square* was 0.685, and as both are dichotomous variables, *Phi* and *Cramer's V* were also calculated, obtaining for both coefficients a value of 0.827 and a 0.000 approximate significance.

One conclusion to be derived from these results is that the users carried out the subjective test mainly when they succeeded in getting the information they wanted. The fact that the successful dialogues were related to dialogue completion might be because unsuccessful dialogues were usually prematurely finished by the user.

To check whether the interaction parameters that affect task success are the same for all the user groups, additional ANOVA studies were carried out, which yielded the results shown in Table 5.9.

Relationship	User group	F	Sig
Dialogue completion - Task success	Users who did not take the subjective test	93.312	0.000
	Users that took the subjective test	19.951	0.000
	All users	180.159	0.000
Dialogue duration - Task success	Users who did not take the subjective test	17.814	0.000
	Users that took the subjective test	9.638	0.004
	All users	21.532	0.000
Number of user turns - Task success	Users who did not take the subjective test	13.025	0.001
	Users that took the subjective test	3.977	0.054
	All users	16.231	0.000
Avg. recognition confidence - Task success	Users who did not take the subjective test	0.105	0.748
	Users that took the subjective test	0.026	0.874
	All users	0.789	0.377
WER - Task success	Users who did not take the subjective test	0.171	0.681
	Users that took the subjective test	0.009	0.925
	All users	0.292	0.590
Avg. words per turn - Task success	Users who did not take the subjective test	12.787	0.001
	Users who took the subjective test	0.964	0.333
	All users	15.452	0.000
% correctly understood utt. - Task success	Users who did not take the subjective test	5.891	0.019
	Users who took the subjective test	3.992	0.054
	All users	12.539	0.001
Number of confirmation turns	Users who did not take the subjective test	0.528	0.471
	Users who took the subjective test	0.789	0.381
	All users	0.963	0.334

Table 5.9. ANOVA table for “task success” and the rest of the interaction parameters regarding the different user groups

As can be observed in the table, the only differences related to task success appeared for its relationships with the number of user turns, the percentage of correctly understood words per turn, and the number of words per turn. The three relationships were significant for the users who did not answer the test, but not for those who answered it, although the first two cases can be considered as almost significant at the 0.05 level. This change might be due to the degree of cooperation of the different types of user. For example, the users who did not answer the test and had unsuccessful dialogues, hung up immediately: 70.37% of the times before the fourth user turn. However, the users who answered the subjective test were more patient and tried to overcome the interaction problems even when in the end they could not obtain the information that they were asking for.

The main difference detected between both user groups was in the relationship between the number of words per turn and task success. For the users who did not answer the test, F had a value of 12.787 and it was significant below the 0.01 level, whereas for those who answered the test, F was 0.964 and it was not significant. This was probably because the distribution of the number of words per turn for the unsuccessful and successful dialogues was more balanced in the case of the users who answered the subjective test. For them, successful and unsuccessful dialogues had a similar number of words per turn. However, the users who did not answer the test employed no more than an average of one word per turn in their unsuccessful dialogues, and more than two turns in the successful ones. Thus, an average of words per turn less or equal to one was an indicator of dialogue failure in the case of users who did not answer the subjective test.

5.5.2. Criteria with highest impact on user satisfaction and task success

Table 5.10 shows the two highest correlation values with user satisfaction, which were obtained in all the statistical studies for the criteria “ease of obtaining information” and “task success”. Thus, as expected, a user was highly satisfied when he found it easy to get the information he wanted. However, it is remarkable that the way of gathering information had the same order of significance with user satisfaction as with the final obtaining of the information. In (Möller, 2005), user satisfaction was also correlated

with the fact that the user finally obtained the information he was looking for. However, Möller’s indicator of ease of communication (which he classified as a comfort factor) did not provide a significant contribution to the overall user satisfaction. This might suggest that ease of interaction is more important for users who have a real need to obtain the information from the system compared with those for whom the interaction is carried out by following predefined scenarios.

Relationship	<i>Pearson</i> (sig)	<i>Tau-b</i> (sig)	<i>Rho</i> (sig)	<i>ANOVA F</i> (sig)
Perceived easy of obtaining the requested information and User satisfaction	0.844 (0.000)	0.750 (0.000)	0.814 (0.000)	31.071 (0.000)
Task success and User satisfaction	0.827 (0.000)	0.732 (0.000)	0.787 (0.000)	33.140 (0.000)

Table 5.10. *Statistical significance of the most important relationships with “user satisfaction”*

In addition, the item of the subjective questionnaire from which the measure “perceived ease of use” is computed, implicitly takes into account the perceived success of the dialogue. Specifically, the answers to question Q8 in the questionnaire (Section 5.3.2) ranged from “No, it was impossible to get the information” to “Yes, it was very easy to get the information”. Thus, there were two different task success measures: an interaction parameter that indicated whether the user was able to get the information that he was looking for, and another that indicated perceived task success. This second measure was extracted from the “ease of obtaining information” parameter by assigning 0 (unsuccessful) to the answer “No, it was impossible” and 1 (success) to the rest.

Contingency tables showed that both task success measures had the same value for all the dialogues. Hence, in our experiments task success was only considered as an interaction parameter. Previous studies such as (Rajman et al., 2004) found that as the users in laboratory tests are not given the possibility to contrast the information provided by the dialogue system, they trust the system responses. For example, they do not check whether the information is correct or useful. Thus, they consider the fact of obtaining a piece of information from the system equivalent to obtaining a correct result.

The authors studied this behaviour by employing laboratory test users who could not discern whether the information about restaurants, menus and prices provided by a dialogue system was correct. In our experiments, the UAH users were provided with real academic information. As they had a real need for this information, they could contrast it and know whether it was accurate or not. Thus, among the unsuccessful dialogues (both from the interaction parameter and the quality perception points of view) there were cases where the system provided information to the user but it was not what they desired, as is shown by the fact that some complete dialogues were unsuccessful. It is a benefit of test fields to allow this separation between the quality of the interaction and the quality of the results.

Within interaction parameters, there is a remarkably high correlation between dialogue completion and task success. As shown in Figure 5.5, although users could hang up when they received the desired information, without waiting for the system to ask if they needed any other information, if the dialogue was successful, they usually waited until the end. Although the percentage of complete and successful dialogues was higher for more collaborative users (i.e. those who answered the questionnaire), both the users that took the subjective test and those who did not take it were patient enough to wait until the end of the dialogue when it was successful.

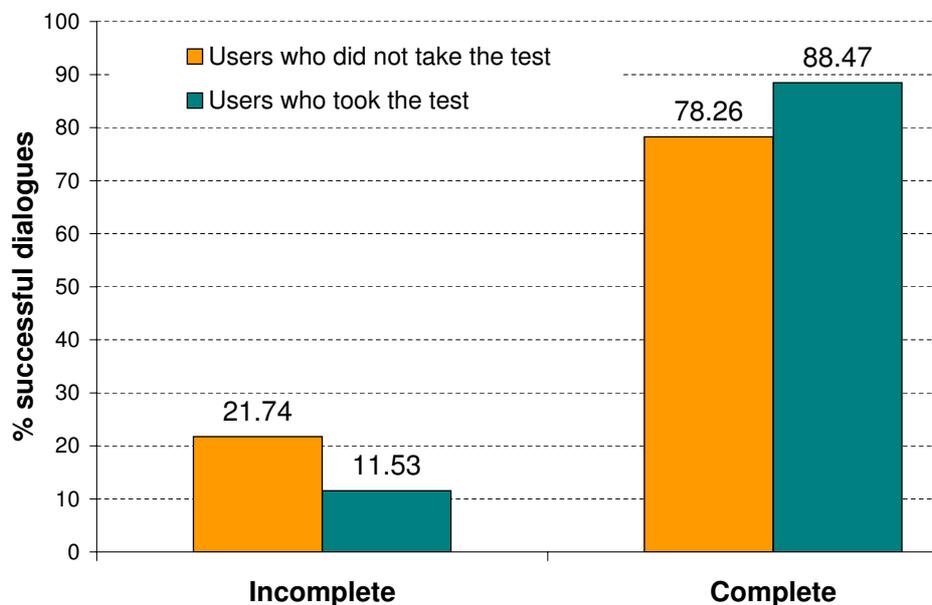


Figure 5.5. Percentage of successful dialogues which are also complete regarding the different user groups

This differs from findings of other authors. For example, Turunen et al. (2006) reported that there were highly significant differences on how the interaction was finished in field and laboratory tests carried out with the Stopman system. In the laboratory tests, 65% of the users employed an explicit request to end the call (e.g. “thank you and goodbye”). On the contrary, in the field tests less than 10% of users waited to the end of the call before hanging up. The number of dialogues in which the users waited until the end of the interaction (i.e. the number of complete dialogues) in our field study is more than 50% higher than in that of Turunen et al. (2006).

Rajman et al. (2004) discuss that a positive attitude of users towards a system does not only depend on its behaviour, but also on the “technophile” or “technophobe” attitude of the users, although they did not control these parameters in their experimentation. In our experiments, 57% of the users rated their knowledge about new technologies for accessing information above 3 in a 1-5 scale, where 1 represented “low” and 5 “high”. Thus, the collaborative nature of our users could be a result of their possible technophile disposition.

Another criterion which is highly correlated with task success and user satisfaction is the perceived ease of error correction. However, the perceived presence of errors is not significantly correlated with any of these criteria. This is probably because although in 48.19% of the successful dialogues the users detected errors, in most cases they managed to circumvent them and obtain the information they were looking for. Specifically, as shown in Figure 5.6, the 69.23% of the users found it “easy” or “very easy” to correct errors in the successful dialogues. However, in the non-successful ones, 83.33% of the users found it “difficult” or “very difficult” to correct the errors.

In (Möller, 2005), the users’ opinion about whether misunderstandings could be easily clarified, which was classified as a contributing factor to dialogue smoothness, was not a good predictor for user satisfaction. Additionally, the author found that user satisfaction could not be fully predicted by task success, and argued that this result could be because of the unrealistic situation of the laboratory experimentation employed. It has been corroborated this finding in our field study, as the subjective user tests could not be replaced by the interaction parameters employed without losing information.

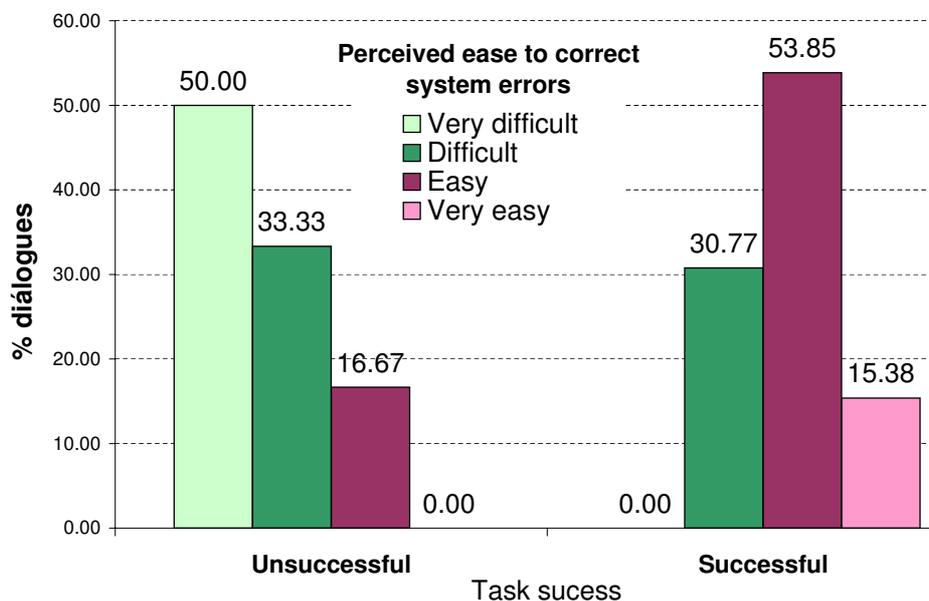


Figure 5.6. *Task success vs. Perceived ease of error correction*

5.5.3. Criteria with highest number of significant relations

The criterion that showed the largest number of significant correlations was the “perceived extent to which UAH understands the user”. On the one hand, it was highly correlated with other quality judgments, like the degree to which the user understands the system, the perceived ease of error correction, the perceived ease of obtaining information, user satisfaction, perceived presence of errors (negative correlation in this case), and the perceived human-like behaviour of the system. Besides, as can be observed in Table 5.6, in most of these relations the significance was highest. On the other hand, this perceived quality criterion was highly correlated with interaction parameters such as completion of the dialogue, task success, dialogue duration or percentage of correctly understood utterances per dialogue.

The most significant relationships between this quality perception and other parameters were with task success and user satisfaction. Perceived system understanding, listed by Möller (2005) as an indicator of speech input quality, was very significantly correlated with user satisfaction in his study.

It is also interesting that the extent to which the user felt that the UAH system understood him was not correlated with the interaction parameters that measure the performance of the speech recognizer, such as WER or confidence scores. However, the percentage of correctly understood utterances was correlated with a significance below 0.01, which indicates that from the user's point of view, speech recognition errors were not important as long as the semantic interpretations were correct and thus these errors were invisible to him. This is reflected in that the perceived presence of errors was related to the percentage of correctly understood utterances and the number of confirmation prompts, but not to WER. However, perceived ease of error correction was not significantly correlated with any of these measures. Both the perceived presence of errors and the perceived ease of correcting them were very highly correlated with the perception that the UAH system understood the user. The perceived presence of errors also negatively affected the user's confidence about what to say next during the interaction.

5.5.4. Impact of user's knowledge and experience

It is noteworthy that the user's knowledge about dialogue systems and new technologies for accessing information were the criteria with the lowest correlation factors with all the others. However, they were significantly correlated with each other. Thus, in our case the knowledge of the user about new technologies for information access was not determinant on the results of the interaction, not in objective terms (e.g. duration, success), nor in perceived terms (e.g. perceived speed, user satisfaction). This may be because the great majority of users had a rather high level of technical knowledge. It is possible that in experiments with other dialogue systems, where users may have more varied backgrounds, these appear to be important criteria.

The previous experience of the user employing the system ("UAH usage") was not correlated with any of the other variables in any statistical study. However, the sign of the correlation parameters indicated that experienced users perceived fewer errors, needed fewer turns to get the information, provoked fewer recognition errors and required fewer confirmation turns.

The fact that previous UAH usage was not significantly correlated with other factors, such as task success or interaction speed, differs from results found in the literature. For example, Turunen et al. (2006) stated

that previous experience in using a system is a very important factor that can help to predict the success and smoothness of the dialogue. Similarly, Park et al. (2007) found that the performance of laboratory test users who had previously employed a system in very strictly predefined interactions was better than for those who had not employed it before. Other authors have studied the effect of user experience on quality judgments. For example Sturm et al. (2005) indicate that a previous prolonged use of the system helps to obtain substantial improvements in quality judgments, such as “ease of use” and “user satisfaction”.

We believe that the impact of the user’s experience is closely related to the type of evaluation carried out. In laboratory tests users are generally trained on how to employ the system, or at least are informed about how to interact with it. In field studies users commonly employ the system without any previous training, and this is why they are less prone to employ characteristics such as help requests (Turunen et al., 2006), of which they are sometimes not aware. However, these characteristics can be very useful to make interaction easier and to recover from error situations. On the other hand, in some particular areas of study, for example spoken dialogue systems for health applications, it has been argued that, contrary to what the previously commented studies suggest, an increasingly richer previous experience using the system does not always imply better performance and perceived quality results. For example Bickmore and Giorgino (2006) report that individuals who intermittently use health dialogue systems on the telephone, compared with those who use them frequently and those who hardly use them at all, obtain the highest satisfaction levels and the best outcomes in terms of the perceived benefits. However, as discussed by Farzanfar et al. (2004), this can be due to the stress of feeling monitored.

5.5.5. Impact of dialogue management initiative

To study the impact of the initiative used for dialogue management, the computations discussed above were repeated, but distinguishing between dialogues with system-directed initiative and dialogues with mixed-initiative. The differences between both initiatives are reported in Table 5.11, where significant correlations are marked with ‘Y’ (yes) and non-significant with ‘N’ (no).

Criterion 1	Criterion 2	Mixed initiative	System-directed initiative
Perc. extent to which the user understands UAH	Perc. extent to which UAH understands the user	N	Y
Perc. interaction speed	Perc. extent to which UAH understands the user	Y	N
Perc. presence of errors made by UAH	Knowledge about dialogue systems	Y	N
Perc. presence of errors made by UAH	Perc. extent to which UAH understands the user	N	Y
User confidence about what to do next	Perc. presence of errors made by UAH	N	Y
User confidence about what to do next	Perc. ease of obtaining the requested information	N	Y
User confidence about what to do next	User satisfaction	N	Y
Perc. human-like behaviour of the UAH system	Perc. presence of errors made by UAH	N	Y
Perc. human-like behaviour of the UAH system	Perc. ease of obtaining the requested information	N	Y
Perc. human-like behaviour of the UAH system	User satisfaction	N	Y
Perc. human-like behaviour of the UAH system	User confidence about what to do next	N	Y
WER	Perc. presence of errors made by UAH	N	Y
Task success	User confidence about what to do next	N	Y
Task success	Perc. ease of obtaining the requested information	N	Y
Task success	Perc. human-like behaviour of the UAH system	N	Y
Dialogue duration	Dialogue completion	Y	N
Dialogue duration	Dialogue completion	N	Y
Dialogue duration	Perc. ease of obtaining the requested information	Y	N
Dialogue duration	User satisfaction	Y	N
Number of user turns	Task success	Y	N
Number of user turns	Task success	Y	N
Number of user turns	User satisfaction	Y	N
Number of user turns	Perc. ease of obtaining the requested information	Y	N
Dialogue completion	Perc. ease of obtaining the requested information	N	Y
Avg. recognition confidence	User satisfaction	Y	N
WER	User confidence about what to do next	N	Y
WER	Dialogue completion	N	Y
WER	Dialogue completion	N	Y
% correctly understood utt.	Number of user turns	Y	N
% correctly understood utt.	Perc. ease of obtaining the requested information	N	Y
% correctly understood utt.	User satisfaction	N	Y
% correctly understood utt.	User confidence about what to do next	N	Y
% correctly understood utt.	Perc. human-like behaviour of the UAH system	N	Y
% correctly understood utt.	Dialogue completion	N	Y
% correctly understood utt.	Task success	N	Y
% correctly understood utt.	Avg. recognition confidence	N	Y
Number of confirmation turns	Perc. presence of errors made by UAH	N	Y
Number of confirmation turns	Dialogue duration	N	Y
Number of confirmation turns	Number of user turns	N	Y
Number of confirmation turns	Avg. recognition confidence	N	Y

Table 5.11. *Criteria that were significantly correlated with one initiative type but not with the other*

It was found that task success was approximately the same for both dialogue management initiatives. This differs from the results in the literature³, where a more flexible initiative led to considerably higher task success rates. In our experiments success was higher for mixed initiative, but the difference between both was practically negligible (77.77% of the mixed-initiative dialogues and 76.92% of the system-directed ones were successful).

However, it was found that task success was related to different factors in each initiative. For example, in mixed-initiative dialogues the user's confidence about what to do next in the dialogue was not correlated with task success, user satisfaction or perceived ease of obtaining information. On the contrary, task success had a significant correlation with user confidence in system-directed dialogues. Probably this is because the user was less constrained in the mixed-initiative interactions, and hence he did not know exactly what he could say (Figure 5.7). This effect did not result in bad interaction results, as task success was not reduced in the case of mixed-initiative interactions.

Correlations of the perceived ease of obtaining information were also very different in the two cases. In the system-directed case it was related to the completion of the dialogue, the number of correctly understood utterances and the opinion that the user had about the human-like behaviour of the system. On the contrary, for mixed-initiative dialogues the perceived ease was not correlated with these measures, but with duration interaction parameters such as dialogue duration or number of user turns. The same happened with satisfaction (judgment) and task success (interaction parameter), which appeared to be highly correlated with duration measures in mixed-initiative interactions, but not in system-directed dialogues. The duration of these dialogues was significantly correlated with user satisfaction, whereas in restricted interaction systems this was not considered so important by users. Besides, as can be observed in Figure 5.8, the average duration of the dialogues was shorter when the interaction was more flexible (mixed-initiative instead of system-directed).

³A comprehensive summary can be found in (Möller, 2005)

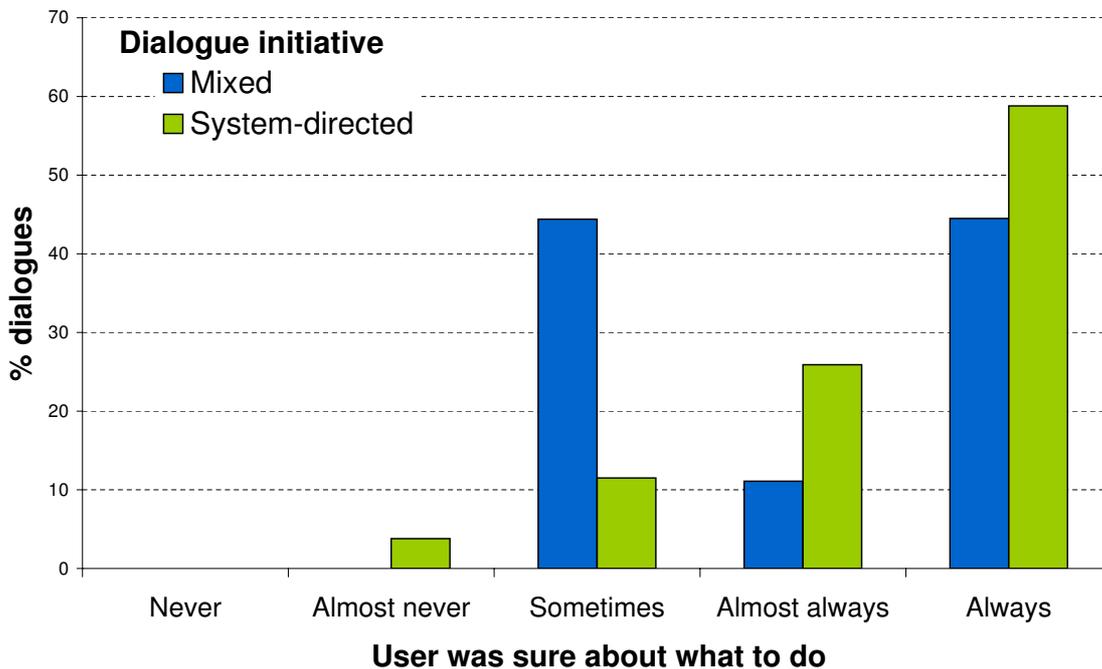


Figure 5.7. Dialogue initiative influence on user confidence

Additionally, the perceived presence of errors was related in mixed-initiative dialogues to the user's knowledge of dialogue systems. This was not the case for system-directed initiative. Besides, it was not correlated with other measures such as user confidence, WER or number of confirmation turns, which were important factors in the system-directed dialogues.

Studies based on laboratory tests like, for example, Rajman et al. (2004) could not find clear quality perception variations with respect to predominance of system or user-driven dialogue management initiatives. Besides, some laboratory tests like the one conducted for the BoRIS system in (Möller, 2005) could not find any significant relationship between the initiative experienced by users and other interaction parameters. However, our results show that the significance of the relationships between the different evaluation criteria, including both interaction parameters and quality judgments, vary depending on the initiative used for dialogue management.

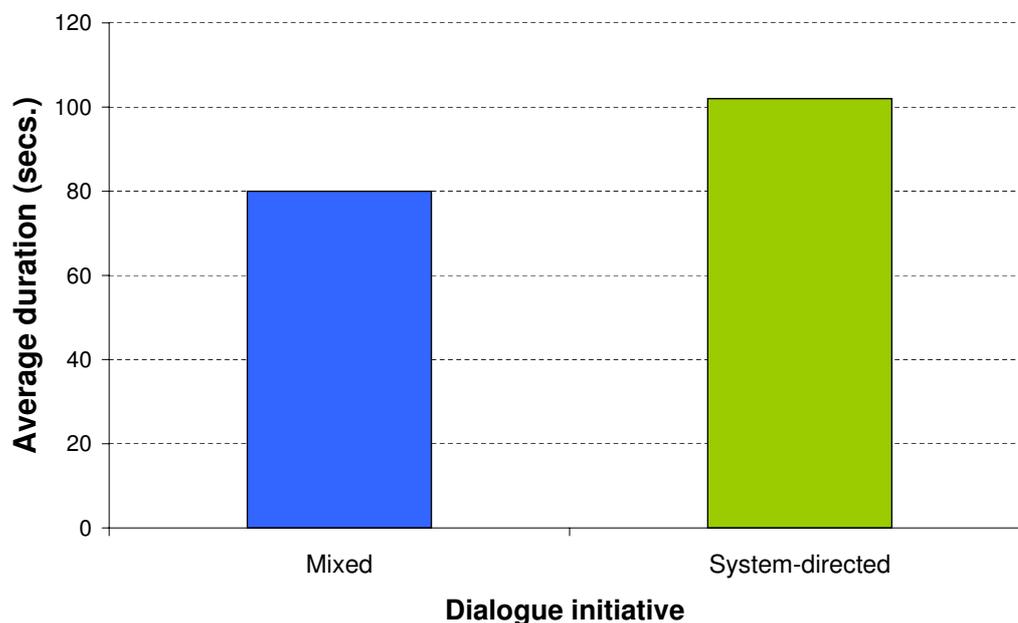


Figure 5.8. Dialogue duration for each dialogue management initiative

5.6 Conclusions

In this chapter it has been presented a study of the relationships between several de-facto standard criteria for the evaluation of a telephone-based spoken dialogue system. Our experimental results are based on a field study using real interactions recorded from users who spontaneously telephoned the system to obtain information, without being recruited to do this.

To carry out our study both interaction parameters (or objective measures) and quality judgments (subjective measures) have been calculated by employing a corpus of real system-user interactions. Specifically, the quantitative criteria employed were: dialogue duration, dialogue completeness, task success, number of user turns, average recognition confidence, average WER, percentage of correctly understood utterances and number of confirmation turns. The qualitative measures were extracted from questionnaires that the users could optionally fill in. The criteria employed were: the extent to which the user felt correctly understood by the system, the extent to which the user understood the system messages, the perceived interaction speed, the per-

ceived ease of error correction, the perceived presence of errors, the extent to which the user was sure about what he should do in every moment of the interaction, the extent to which the user believed the system's behaviour was human-like, and the level of user satisfaction with the interaction. Additionally information about users was also taken into account, namely: user knowledge about new technologies for information access, user knowledge about spoken dialogue systems, and number of times the user had already used the system.

Several statistical studies were developed from which significant relations between all the criteria were extracted. This approach has not been sufficiently exploited in the literature, and some noteworthy empirical findings have been highlighted. Our empirical evidence shows that task success, perceived ease of obtaining information and perceived extent to which the system understands the user are very closely correlated with user satisfaction. These results suggest that obtaining the required information does not completely explain user satisfaction, as in some cases users judged successful dialogues as not satisfying because they found it difficult to obtain the information they were looking for. This is one of the implications derived from the usage of field tests, in which users are very concerned not only with obtaining the information they were looking for, but also with doing it easily.

Furthermore, the relationship between the perceived ease of obtaining information and other criteria varies remarkably with the dialogue management initiative. Our experimental results show that, in the system-directed dialogues, perceived ease was related to the good functioning of the understanding module. On the contrary, in mixed-initiative dialogues, both user satisfaction and the perceived ease of obtaining the information seemed to be related to duration metrics. This had a strong implication in the quality judgments, as task success was highly correlated with user satisfaction in both initiatives. Thus, our results suggest that the prediction of user satisfaction also depends on the dialogue management initiative used. In the mixed-initiative dialogues it seemed to be more directly related to objective measures, such as dialogue duration. However, in more restricted dialogues, subjective measures such as the perceived extent to which the user feels that he is understood by the system, had a bigger impact. This is an important result that could indicate a need to tailor evaluation procedures to the type of interactions being analysed.

Additionally, we have studied the reasons that made some users answer the optional subjective test from which the quality judgments were obtained. It was found that it could be explained mainly in terms of dialogue completion and task success. Thus, the experiments carried out by including the users' perceptions about the quality of the system, corresponded mainly to successful dialogues, in which the users obtained the information they were looking for. This could be one of the reasons why it was found that these users were very cooperative, which yielded high dialogue completion rates rarely reported in previous field test studies. Besides, contrary to what generally happens in laboratory studies, these measures consider that even when the user obtains information from the system, the dialogue cannot be considered successful if the provided information is not correct. Finally, no evidence was encountered of the effect of the users' previous experience employing the system on system performance or task success.

*Hemos llegado al fin y yo inauguro
triste mi paz: la obra está completa.*

Jorge Guillén, *Obra completa*

6

Conclusions and future work

6.1 Summary of contributions

The Thesis has introduced novel contributions directed towards the development of adaptive and portable spoken dialogue systems. Specifically, the Thesis presents approaches to build systems that recognize the users' emotions and adapt to their expectations and needs by detecting them in field evaluation studies. Additionally, portability is fostered by adapting the resources available in a language for the cost-effective development of speech-based systems in other languages.

A **complete spoken dialogue system, known as UAH, was developed** to evaluate the contributions of the Thesis. As described in Chapter 2, it provides academic information on the Dept. of Languages and Computer Systems. The system follows an architecture that is made up of five modules: a speech recognizer, an automatic grammar generation module, a dialogue manager, a database access module and an oral response generator. Given that the information provided by the UAH dialogue system is continually changing, a method was introduced to keep the recognition grammars updated with the last changes in the databases without introducing a delay in the interaction. The technique proposed is called **Triggered Grammar rules Creation (TGC)** and was implemented into the **Grammar Automatic Generation (GAG) module** of the UAH system. The system uses **different confirmation techniques** (explicit and implicit) as well as **several dialogue management initiatives** (system-directed and mixed), in order to study their benefits both in terms of performance and perceived usability. The dialogue manager was developed using **VoiceXML**

documents that are dynamically generated according to the interaction context, which is also employed to tailor the system output to the user's needs.

The system was made public in June 2005. **From the interactions of the users with the system, an evaluation corpus has been generated and semi-automatically annotated.** The users' utterances were recorded in .wav format along with information about the recording starting time, the previous system turn, and the speech recognition result (including confidence scores). This information was stored in a database along with the dialogue starting and ending times. Nine parameters were automatically computed from this information and they were subsequently employed to evaluate the system. Some of them were obtained automatically (e.g. dialogue duration), while others required manual annotation (e.g. dialogue success). Following this methodology an annotated corpus of **85 dialogues (422 user turns)** was obtained from the interactions of a year of user calls to UAH dialogue system. The users were invited to answer a questionnaire in which they could give their personal opinion about different aspects of their interaction with the system. From these opinions, a total of 12 quality criteria were obtained for each utterance including parameters such as the user satisfaction or the perceived interaction speed. Additionally, **the corpus was annotated twice by nine non-expert annotators who classified each utterance in one of the following emotional categories: neutral, doubtful, angry and bored.**

Chapter 3 has described the contributions of the Thesis which are focused on emotion recognition. Firstly, a state of the art on the main approaches used in the literature has been presented. It shows that research on emotion recognition has been mainly centred on how to apply different machine learning algorithms to differentiate between emotions, and less effort has been directed to determine the information in which the learning process should be based on. The most widespread approach consists of employing multimodal (audiovisual) information obtained from acted emotions, in order to retain strict control over the collected data.

The main contribution of the research carried out in the Thesis is the inclusion of contextual information for the **recognition of real emotions in spoken dialogue systems.** This has been a very challenging task, firstly because it had to be based on only one input modality, and secondly

because natural emotions are expressed very subtly and usually the emotional categories encountered are very unbalanced, i.e. there is one predominant emotion, which generally corresponds to the “neutral” state. This raises some problems that, to our knowledge, had not been fully addressed in literature before.

One of the main difficulties was to obtain reliable annotations of the emotional corpora considering that the traditional inter-annotator agreement measures are deeply affected by the skewing of the corpora obtained. In Chapter 3 we have presented a **detailed discussion on how to reliably calculate and interpret kappa coefficients with corpora of real emotions**. The proposed methodology has been evaluated using the UAH emotional corpus and the significance of the results has been statistically computed. The results of this evaluation show that, on the one hand, **our proposal makes possible to obtain annotator agreement values that are closer to the maximum attainable**. It also allows non-expert annotators to detect more non-neutral emotions, and the annotation result is less affected by differences among the annotators. On the other hand, **automatic emotion recognition obtains values that can be more than 40% better than those obtained with the approaches described in the literature**, which are based on the acoustic features without taking into account the proposed contextual information. The fact that all the proposals have been tested with a corpus of real emotions makes **the results directly applicable to practical operations**, as all **the algorithms proposed can be used dynamically during the execution of the dialogue systems**.

Chapter 4 has presented the work carried out during a three months stay in the Laboratory of Computer Speech Processing in the Technical University of Liberec under the supervision of Prof. Jan Nouza. As stated in the chapter, the application of the resources that are already available for a specific language to recognize a different language facilitates the development of new speech recognizers, specially in the case of minority languages and dialects. The approaches that can be found in literature usually demand laborious phonetic and linguistic studies of the languages involved. **A new method has been proposed to adapt a speech recognizer to other languages in a efficient way**. Our proposal has been evaluated using the MyVoice system, which was developed in the Technical University of Liberec

for handicapped Czech users, and whose translation to the Spanish language is presented in the Thesis. It has been empirically demonstrated that **the proposed cross-lingual adaptation can be performed in a relatively short time by carrying out an expert-driven correspondence between both languages' phonetic alphabets**. Experimental results show that for a task involving a vocabulary of 432 commands, 95.6% performance can be attained for Spanish and 97.5% for Slovak after adapting a Czech recognizer. For vocabularies up to 149k words, 72.9% and 77.4% accuracy rates were obtained for Spanish and Slovak respectively. The results are very promising given that they show that that portability of speech recognizers can be ensured fairly simply and that **the approach can achieve good results with similar languages such as Czech and Slovak, and also for phonetically different languages such as Czech (Slavic) and Spanish (Italic)**.

In Chapter 5 a **field evaluation of the UAH system** has been presented. Typically, in the literature, evaluation is carried out under restricted laboratory conditions, in which users are recruited to test the systems following a predefined list of scenarios. The problem with this method is that the scenarios may differ from the tasks that a user would have selected in a real interaction. In the field study that has been carried out in the Thesis, the users interacted with the system on their own initiative. Thus, the dialogues appeared as a need of the users for the information that the system provides. Additionally, most of the evaluation approaches cover interaction parameters and quality judgments independently. On the contrary, our results measure the relationship between these two factors, determining the **statistically significant relationships**. To do so, several coefficients adapted to the type of information processed have been employed, such as *Pearson correlation coefficients*, *ANOVA studies*, *Tau-b*, *Rho* and *Eta-square coefficients*. Thus, our study provides **new empirical evidence** which is more relevant to predict real systems behaviour, yielding very interesting results **on the criteria that affected more deeply the user satisfaction and task success, the impact of users knowledge and experience or the suitability of different dialogue management initiatives**. These interesting empirical relationships can be taken into account to enhance system development and also for the evaluation of the systems performance and usability.

6.2 Future work

6.2.1. Recognition of non-acted emotions

The future work includes evaluating the proposed technique using other corpora, to guarantee its independence of the application domain. Additionally, we are currently integrating the emotion recognizer into the UAH system architecture. This module will work in parallel with the speech recognizer, accepting the voice signal as an input that comes from the telephony card, and obtaining the emotion recognized, which will be sent to the dialogue manager. Then, the dialogue manager will adapt its behaviour to the emotion recognized. The initial objective is to use only the emotions described in the Thesis: *angry*, *bored* and *doubtful*. Dialogue management for doubtful users will require providing them a more detailed help and using system-directed initiatives, so that it will be clear to the user what he is required to do in the interaction step. On the other hand, when the user is bored, the optimal strategy could be to make the system prompts shorter, change the prosody of the synthesized speech, and employ more implicit confirmations in which the user will not be explicitly prompted to confirm the information he provides. For the *angry* emotional state, error recovery strategies must be considered and incorporated to the system, with some feedback on misunderstandings.

6.2.2. Cross-lingual adaptation of speech recognizers

Our immediate future objective is to compare the adapted recognizers with a plain Spanish and Slovak system built from the scratch. We have already carried out several experiments with the MS Vista Spanish recognizer. The results of this evaluation show an error rate of around 20%, as against 24% in the adapted system. This indicates that the adapted system is not too far off a native-language system. Further experiments on this topic will be carried out in the near future.

The promising findings described in Chapter 4 encourage us to consider the future application of the proposed cross-language phonetic adaptation for languages with very small speech and linguistic resources. A particularly interesting task would be to study the suitability of the proposed cross-lingual adaptation technique for minority languages that do not belong to the Indo-European family.

6.2.3. Field evaluation of spoken dialogue systems

Besides the dialogues used in the Thesis, we have acquired another corpus under laboratory conditions, i.e. by means of recruited users employing the UAH system using predefined scenarios. Our objective is to carry out an evaluation of the results obtained with this method, and compare them with the results of the field evaluation described in Chapter 5.

We also believe that statistical analysis such as the ones presented in the Thesis lead to interesting empirical relationships that can be taken into account to enhance system development and evaluation. Such studies can serve to evaluate systems as a whole instead of individual components. Another future line of work will focus on incorporating factor analysis studies to group different criteria to obtain the major trends necessary to optimize the performance and foster user satisfaction. For this purpose a more extensive list of criteria will be compiled. Once the factors are computed, they will be analysed to obtain their interrelationships so as to be able to build a criteria taxonomy that can then be compared with other state-of-the-art taxonomies, for example the Quality-Of-Service proposed by (Möller, 2002).

New criteria will be introduced and studied to evaluate the affective intelligence. In this way, it will be possible to measure the benefits of adding the proposed emotional intelligence mechanism to the UAH dialogue system (Section 6.2.1), both objectively (e.g. in terms of task success) and subjectively (considering the users opinions).



Publications

The research described in the Thesis has been published in the national and international conferences and journals listed below.

There are several publications describing different stages of the design and development of the UAH spoken dialogue system (Chapter 2). The next paper presents the UAH spoken dialogue system, describing all its modules and the innovative approaches employed for its development.

- (Callejas and López-Cózar, 2005b) *Callejas, Z., López-Cózar, R., 2005. Implementing modular dialogue systems: a case study. In: Proc. of Applied Spoken Language Interaction in Distributed Environments (ASIDE 05). Aalborg, Denmark.*

The following papers correspond to a description and a demonstration of the GAG module respectively. They correspond to the first version of the module in which the grammar rules were designed and generated dynamically using a web interface, but the automatic update with the database changes was still not developed:

- (Callejas and López-Cózar, 2005c) *Callejas, Z., López-Cózar, R., 2005. Nueva técnica de generación automática de gramáticas para sistemas de diálogo. Procesamiento del Lenguaje Natural (35), 205 - 212.*
- (Callejas and López-Cózar, 2005a) *Callejas, Z., López-Cózar, R., 2005. GAG: Generación automática de gramáticas en un sistema conversacional de interacción oral. Procesamiento del Lenguaje Natural (35), 457 - 458.*

The final version of the GAG tool, in which the process of grammar creation and updating was fully automatic, was presented in the following paper, in which the TGC technique was also fully designed, developed and evaluated.

- (Callejas and López-Cózar, 2007a) *Callejas, Z., López-Cózar, R., 2007. Automatic creation of ASR grammar rules for unknown vocabulary applications. In: Proc. of 8th International Workshop on Electronics, Control, Modelling, Measurement and Signals (ECMS'07). Liberec, Czech Republic. pp.50-55*

Regarding the research on emotion recognition described in Chapter 3, the next paper presents a state on the art on emotion recognition paying special attention to the latest international projects in the area, putting our research into context:

- (Callejas and López-Cózar, 2007c) *Callejas, Z., López-Cózar, R., 2007. Emotion recognition for spoken dialogue systems. In: Proc. of I Simposio en Desarrollo de Software (SDS'07). Granada, Spain. pp. 59-68*

A complete study on Kappa coefficients and how they are affected by the skewness of the corpora of non-acted emotions is presented in:

- (Callejas and López-Cózar, 2008b) *Callejas, Z., López-Cózar, R., 2008. On the use of kappa coefficients to measure the reliability of the annotation of non-acted emotions. Lecture Notes in Computer Science (5078), 221 - 232.*

The next paper presents the first stage on the development of the proposed automatic emotion recognition approach using contextual information. It presents the preliminary version of the two-step algorithm described in Section 3.5:

- (Callejas and López-Cózar, 2007b) *Callejas, Z., López-Cózar, R., 2007. Decisive factors in the annotation of emotions for spoken dialogue systems. Advances in Soft Computing (45), 747 - 754.*

The main results of the research described in Chapter 3 are presented in the following paper. It describes the work on both human and automatic emotion recognition using contextual information. It includes the optimal version of the two-step method for machine-learned recognition.

- (Callejas and López-Cózar, 2008a) *Callejas, Z., López-Cózar, R., 2008. Influence of contextual information in emotion annotation for spoken dialogue systems. Speech Communication 50 (5), 416-433.*

Regarding the research on the cross-lingual adaptation of a Czech speech recognizer described in Chapter 4, the following paper corresponds to a demonstration of the Spanish version of the MyVoice system recognizing Spanish with the adapted Czech speech recognizer (Chapter 4).

- (Callejas et al., 2007) *Callejas, Z., Nouza, J., Cerva, P., López-Cózar, R., 2007. Myvoice goes Spanish. Cross-lingual adaptation of a voice-controlled PC tool for handicapped people. Procesamiento del Lenguaje Natural (39), 277 - 278.*

Finally, the research done on field evaluation of SDSs (Chapter 5) is presented in.

- (Callejas and López-Cózar, 2008c) *Callejas, Z., López-Cózar, R. Relations between de-facto criteria in the evaluation of a spoken dialogue system. Speech Communication. In press. Available online since 15th April 2008. DOI: 10.1016/j.specom.2008.04.004*

Bibliography

- Adell, J., Bonafonte, A., Escudero, D., 2005. Analysis of prosodic features towards modelling of emotional and pragmatic attributes of speech. *Procesamiento de Lenguaje Natural* 35, 277–284.
- Ai, H., Litman, D., Forbes-Riley, K., Rotaru, M., Tetreault, J., Purandare, A., 2006. Using systems and user performance features to improve emotion detection in spoken tutoring dialogs. In: *Proc. of the 9th International Conference on Spoken Language Processing (Interspeech'06-ICSLP)*. Pittsburgh, USA, pp. 797–800.
- Alexandersson, J., Becker, T., 2001. Overlay as the basic operation for discourse processing in a multimodal dialogue system. In: *Proc. of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*. Seattle, USA, pp. 8–14.
- Andeani, G., Fabbriozio, D. D., Gilbert, M., Gillick, D., Hakkani-Tur, D., Lemon, O., 2006. Let's DISCOH: Collecting an Annotated Open Corpus with Dialogue Acts and Reward Signals for Natural Language Helpdesks. In: *Proc. of IEEE 2006 Workshop on Spoken Language Technology (SLT'06)*. Palm Beach, Aruba, pp. 218–221.
- Ang, J., Dhillon, R., Krupski, A., Shriberg, E., Stolcke, A., 2002. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In: *Proc. of the 7th International Conference on Spoken Language Processing (Interspeech'02-ICSLP)*. Denver, USA, pp. 2037–2040.
- Artstein, R., Poesio, M., 2005. *kappa₃ = alpha (or beta)*. Tech. rep., University of Essex.
- AUBADE, 2005. <http://www.aubade-group.com/>.
- Augmented Multiparty Interaction Project, 2007. <http://www.amiproject.org/>.

- Baggia, P., Castagneri, G., Danieli, M., 2000. Field trials of the Italian ARISE train timetable system. *Speech Communication* 31, 355–367.
- Bangalore, S., Hakkani-Tur, D., Tur, G., 2006. Introduction to the Special Issue on Spoken Language Understanding in Conversational Systems. *Speech Communication* 48 (3-4), 233–3238.
- Batliner, A., Hacker, C., Steidl, S., Nöth, E., D’Arcy, S., Russel, M., Wong, M., 2004. Towards multilingual speech recognition using data driven source/target acoustical units association. In: *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP’04)*. Montreal, Quebec, Canada, pp. 521–524.
- Baudoin, F., Bretier, P., Corruble, V., 2005. A dialogue agent with adaptive and proactive capabilities. In: *Proc. of IEEE/WIC/ACM International Conference on Intelligent Agent Technology*. Compiègne, France, pp. 293–296.
- Bayeh, R., Lin, S., Chollet, G., Mokbel, C., 2004. You stupid tin box! - children interacting with the aibo robot: a cross-linguistic emotional speech corpus. In: *Proc. of the 4th International Conference on Language Resources and Evaluation (LREC’04)*. Lisbon, Portugal, pp. 171–174.
- Becker, T., Gerstenberger, C., Kruijff-Korbayova, I., Korthauer, A., Pinkal, M., Pitz, M., Poller, P., Schehl, J., 2006. Natural and intuitive multimodal dialogue for In-Car Applications: The SAMMIE System. In: *Proc. of the 4th European Conference of Prestigious Applications of Intelligent Systems (PAIS’06)*. Riva del Garda, Italy, pp. 612–616.
- Beringer, N., Kartal, U., Louka, K., Schiel, F., Tük, U., 2002. PROMISE: A Procedure for Multimodal Interactive System Evaluation. In: *Proc. of the Workshop on Multimodal Resources and Multimodal Systems Evaluation*. Las Palmas de Gran Canaria, Spain, pp. 77–80.
- Bernsen, N., Dybkjaer, L., 1997. The DISC project. *ELRA Newsletter* 2 (2), 6–8.
- Bernsen, N., Dybkjaer, L., Dybkjaer, H., 1994. A dedicated task-oriented dialogue theory in support of spoken language dialogue system design. In:

-
- Proc. of the 3rd International Conference on Spoken Language Processing (ICSLP'94). Yokohama, Japan, pp. 875–878.
- Bernsen, N. O., Dybkjaer, L., 2000. A methodology for evaluating spoken language dialogue systems and their components. In: Proc. of the 2nd International Conference on Language Resources and Evaluation (LREC'00). Athens, Greece, pp. 183–188.
- Bickmore, T., Giorgino, T., 2004. Some novel aspects of health communication from a dialogue systems perspective. In: Proc. of AAAI Fall Symposium on Dialogue Systems for Health Communication. Washington DC, USA, pp. 275–291.
- Bickmore, T., Giorgino, T., 2006. Health dialog systems for patients and consumers. *Journal of Biomedical Informatics* 39 (5), 556–571.
- Bishop, C. M., 2006. *Pattern Recognition and Machine Learning*. Springer.
- Boehner, K., DePaula, R., Dourish, P., Sengers, P., 2007. How emotion is made and measured. *International Journal of Human-Computer Studies*, Special Issue on Evaluating Affective Interactions 65 (4), 275–291.
- Boersma, P., 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. Tech. rep., Institute of Phonetic Sciences, University of Amsterdam.
- Bohus, D., Grau, S., Huggins-Daines, D., Keri, V., Krishna, G., Kumar, R., Raux, A., Tomko, S., 2007. Conquest - an Open-Source Dialog System for Conferences. In: Proc. of Human Language Technologies'07: The Conference of the North American Chapter of the Association for Computational Linguistics. Rochester, NY, USA, pp. 9–12.
- Bohus, D., Rudnicky, A., 2002. LARRI: A Language-Based Maintenance and Repair Assistant. In: Proc. of Multi-Modal Dialogue in Mobile Environments Conference (IDS'02). Kloster Irsee, Germany, pp. 203–218.
- Bonaventura, P., Gallochio, F., Micca, G., 1997. Multilingual speech recognition for flexible vocabularies. In: Proc. of 5th European Conference on Speech Communication and Technology (Eurospeech'97). Rhodes, Greece, pp. 355–358.

- Bos, J., Klein, E., Lemon, O., Oka, T., 1999. The verbmobil prototype system - a software engineering perspective. *Journal of Natural Language Engineering* 5(1), 95–112.
- Boves, L., Os, E. D., 2002. Multimodal services, a MUST for UMTS. Tech. rep., EURESCOM.
- Brown, M. K., 1999. Grammar Representation Requirements for Voice Markup Languages. Tech. rep., W3C.
- Burkhardt, F., van Ballegooy, M., Englert, R., Huber, R., 2005. An emotion-aware voice portal. In: *Proc. of Electronic Speech Signal Processing*. Prague, Czech Republic, pp. 123–131.
- Cahill, L., Tiberius, C., 2002. Cross-linguistic phoneme correspondences. In: *Proc. of 19th International Conference on Computational Linguistics (COLING'02)*. Taipei, Taiwan, pp. 1–5.
- Callas - Conveying Affectiveness in Leading-Edge Living Adaptive Systems, 2007. <http://www.callas-newmedia.eu/>.
- Callejas, Z., López-Cózar, R., 2005a. GAG: Generación automática de gramáticas en un sistema conversacional de interacción oral. *Procesamiento del Lenguaje Natural* 35, 457–458.
- Callejas, Z., López-Cózar, R., 2005b. Implementing modular dialogue systems: a case study. In: *Proc. of Applied Spoken Language Interaction in Distributed Environments (ASIDE'05)*. Aalborg, Denmark.
- Callejas, Z., López-Cózar, R., 2005c. Nueva técnica de generación automática de gramáticas para sistemas de diálogo. *Procesamiento del Lenguaje Natural* 35, 205–212.
- Callejas, Z., López-Cózar, R., 2007a. Automatic creation of ASR grammar rules for unknown vocabulary applications. In: *Proc. of the 8th International Workshop on Electronics, Control, Modelling, Measurement and Signals (ECMS'07)*. Liberec, Czech Republic, pp. 50–55.
- Callejas, Z., López-Cózar, R., 2007b. Decisive factors in the annotation of emotions for spoken dialogue systems. *Advances in Soft Computing* 45, 747–754.

-
- Callejas, Z., López-Cózar, R., 2007c. Emotion recognition for spoken dialogue systems. In: Proc. of I Simposio en Desarrollo de Software (SDS'07). Granada, Spain, pp. 59–68.
- Callejas, Z., López-Cózar, R., 2008a. Influence of contextual information in emotion annotation for spoken dialogue systems. *Speech Communication* 50 (5), 416–433.
- Callejas, Z., López-Cózar, R., 2008b. On the use of kappa coefficients to measure the reliability of the annotation of non-acted emotions. *Lecture Notes in Computer Science* 5078, 221–232.
- Callejas, Z., López-Cózar, R., 2008c. Relations between de-facto criteria in the evaluation of a spoken dialogue system. *Speech Communication* In Press. Available online since 15th April 2008. DOI: 10.1016/j.specom.2008.04.004.
- Callejas, Z., Nouza, J., Cerva, P., López-Cózar, R., 2007. Myvoice goes Spanish. Cross-lingual adaptation of a voice-controlled PC tool for handicapped people. *Procesamiento del Lenguaje Natural* 39, 277–278.
- Camurri, A., Mazzarino, B., Volpe, G., 2004. Expressive interfaces. *Cognition, Technology and Work* 6 (1), 15–22.
- Carreiras, M., García-Albea, J. E., Sebastián-Gallés, N., 1996. *Language processing in Spanish*. Lawrence Erlbaum Associates.
- Castagneri, G., Baggia, P., Danieli, M., 1998. Field trials of the Italian ARISE train timetable system. In: Proc. of the Interactive Voice Technology for Telecommunications Applications Workshop (IVTTA'98). Torino, Italy, pp. 97–102.
- Catizone, R., Setzer, A., Wilks, Y., 2003. Multimodal Dialogue Management in the COMIC Project. In: Proc. of EACL'03 Workshop on Dialogue Systems: interaction, adaptation, and styles of management. Budapest, Hungary, pp. 25–34.
- Cerrato, L., 2002. A comparison between feedback strategies in human-to-human and human-machine communication. In: Proc. of the 8th Inter-

- national Conference on Spoken Language Processing (ICSLP'02). Vol. 2. Denver, USA, pp. 557–560.
- Cerva, P., Nouza, J., 2007. Design and development of voice controlled aids for motor-handicapped persons. In: Proc. of the 11th International Conference on Spoken Language Processing (Interspeech'07-Eurospeech). Antwerp, Belgium, pp. 2521–2524.
- Chambers, N., Allen, J., 2004. Stochastic language generation in a dialogue system: Toward a domain independent generator. In: Proc. of the 5th SIGdial Workshop on Discourse and Dialogue. Boston, USA, pp. 9–18.
- Cheyen, A., Julia, L., 1995. Multimodal maps: An agent based approach. In: Proc. of International Conference on Cooperative Multimodal Cooperation. Eindhoven, Holland, pp. 111–121.
- CHIL - Computers In the Human Interaction Loop, 2007. <http://chil.server.de/servlet/is/101/>.
- Chu, S.-W., O'Neill, I., Hanna, P., McTear, M., 2005. An approach to multi-strategy dialogue management. In: Proc. of 9th International Conference on Spoken Language Processing (Interspeech'05-Eurospeech). Lisbon, Portugal, pp. 865–868.
- Cicchetti, D. V., Feinstein, A. R., 1990. High agreement but low Kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology* 43 (6), 551–558.
- Cohen, J., 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20 (3), 37–46.
- Cohen, J., 1968. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* 70 (4), 213–220.
- Cole, R., Hirschman, L., Atlas, L., Beckman, M., Biermann, A., Bush, M., Clements, M., Cohen, J., Garcia, O., Hanson, B., Hermansky, H., Levinson, S., McKeown, K., Morgan, N., Novick, D. G., Ostendorf, M., Oviatt, S., Price, P., Silverman, H., Spitz, J., Waibel, A., Weinstein, C., Zahorian, S., Zue, V., 1995. *The Challenge of Spoken Language Systems: Research*

- Direction for the Nineties. *IEEE Transactions on Speech and Audio Processing* 3, 1–20.
- Cole, R., Mariani, J., Uszkoreit, H., Varile, G. B., Zaenen, A., Zampolli, A., Zue, V. (Eds.), 1997. *Survey of the State of the Art in Human Language Technology*. Cambridge University Press.
- Companions, 2007. <http://nlp.shef.ac.uk/companions/>.
- Corradini, A., Bernsen, N., Fredriksson, M., Johanneson, L., Königsmann, J., Mehta, M., 2004. Towards believable behavior generation for embodied conversational agents. In: *Proc. of the Workshop on Interactive Visualisation and Interaction Technologies (IV&IT 2004)*. Krakow, Poland, pp. 946–953.
- Corradini, A., Mehta, M., Bernsen, N. O., Charfuelán, M., 2005. Animating an interactive conversational character for an educational game system. In: *Proc. of the 2005 International Conference on Intelligent User Interfaces*. San Diego, CA, USA, pp. 183–190.
- CoSy Home, 2007. <http://www.cs.bham.ac.uk/research/projects/cosy/>.
- Cowie, R., 2000. Describing the emotional states expressed in speech. In: *Proc. of the International Speech Communication Association (ISCA) Workshop on Speech and Emotion*. Newcastle, Northern Ireland, UK, pp. 11–18.
- Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schröder, M., 2000. FEELTRACE: An Instrument for Recording Perceived Emotion in Real Time. In: *Proc. of the International Speech Communication Association (ISCA) Workshop on Speech and Emotion*. Newcastle, Northern Ireland, UK, pp. 19–24.
- Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J., 2001. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine* 18 (1), 32–80.
- Cowie, R., Schröder, M., 2005. Piecing Together the Emotion Jigsaw. *Lecture Notes in Computer Science* 3361, 305–317.

- Craggs, R., Wood, M. M., 2003. Annotating emotion in dialogue. In: Proc. of the 4th SIGdial Workshop on Discourse and Dialogue. Sapporo, Japan, pp. 218–225.
- Critchley, H. D., Rotshtein, P., Nagai, Y., O'Doherty, J., Mathias, C. J., Dolana, R. J., 2005. Activity in the human brain predicting differential heart rate responses to emotional facial expressions. *NeuroImage* 24, 751–762.
- Cuayáhuitl, H., Renals, S., Lemon, O., Shimodaira, H., 2006. Reinforcement learning of dialogue strategies with hierarchical abstract machines. In: Proc. of IEEE/ACL Workshop on Spoken Language Technology (SLT'06). Palm Beach, Aruba, pp. 182–186.
- Dale, R., September 2003. Next-generation spoken dialog systems. Technology Trends Seminar Series, Macquarie University, <http://www.ict.csiro.au/MU/Trends/2003.htm>.
- DARPA, 1992. Speech and Natural Language Workshop. Defense Advanced Research Projects Agency (DARPA), San Mateo, USA.
- DARPA, 1994. Speech and Natural Language Workshop. Defense Advanced Research Projects Agency (DARPA), San Mateo, USA.
- Davies, M., Fleiss, J. L., 1982. Measuring agreement for multinomial data. *Biometrics* 38 (4), 1047–1051.
- de Melo, C., Paiva, A., 2005. Environment expression: Expressing emotions through cameras, lights and music. In: Proc. of the 1st International Conference on Affective Computing and Intelligent Interaction (ACII'05). Beijing, China, pp. 715–722.
- Degerstedt, L., Jönsson, A., 2006. LinTest, A development tool for testing dialogue systems. In: Proc. of the 9th International Conference on Spoken Language Processing (Interspeech'06-ICSLP). Pittsburgh, USA, pp. 489–492.
- den Os, E., Boves, L., Lamel, L., Baggia, P., 1999. Overview of the ARISE project. In: Proc. of the European Conference on Speech Technology (Eurospeech'99). Budapest, Hungary, pp. 1527–1530.

- Devillers, L., Maynard, H., Rosset, S., 2004. The French Media/Evalda project: the evaluation of the understanding capability of spoken language dialog systems. In: Proc. of the 4th International Conference on Language Resources and Evaluation (LREC'04). Vol. 6. Lisbon, Portugal, pp. 2131–2134.
- Devillers, L., Vidrascu, L., Lamel, L., 2005. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks* 18 (4), 407–422.
- DISC, 1999. DISC Final Report covering the period from 1.6.98 to 28.2.99. Deliverable D5.2. Tech. rep., The DISC Consortium.
- Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P., 2003. Emotional speech: towards a new generation of databases. *Speech Communication* 40, 33–60.
- Dunn, G., 1989. Design and analysis of reliability studies: the statistical evaluation of measurement errors. Edward Arnold.
- Dybkjaer, L., Bernsen, N. O., 2000. Usability issues in spoken language dialogue systems. *Natural Language Engineering* 6, 243–271.
- Dybkjaer, L., Bernsen, N. O., Minker, W., 2004. Evaluation and usability of multimodal spoken language dialogue systems. *Speech Communication* 43, 33–54.
- EAGLES, 1996. Evaluation of Natural Language Processing Systems. Final report. Document EAG-EWG-PR2. Tech. rep., Center for Sprogteknologi, Copenhagen, Denmark.
- EUROPA - CORDIS: Community Research and Development Information Service, 2006. <http://cordis.europa.eu/>.
- Farzanfar, R., Frishkopf, S., Migneault, J., Friedman, R., 2004. Telephone-linked care for physical activity: A qualitative evaluation of the use patterns of an information technology program for patients. *Journal of Biomedical Informatics* 38 (3), 220–228.

- Feinstein, A. R., Cicchetti, D. V., 1990. High agreement but low Kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology* 43 (6), 543–549.
- Fleiss, J. L., 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76 (5), 378–382.
- Fleiss, J. L., Cohen, J., 1973. The equivalence of weighted kappa and the interclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement* 33, 613–619.
- Fluhr, C., Schmit, D., Andrieux, C., Ortet, P., Bisson, F., Combet, V., 1999. Crosslingual interrogation of multilingual catalogs. *Lecture Notes in Computer Science* 1696, 294–310.
- Forbes-Riley, K., Litman, D. J., 2004a. Predicting emotion in spoken dialogue from multiple knowledge sources. In: *Proc. of the Human Language Technology Conference - North American chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL'04)*. Boston, USA, pp. 201–208.
- Forbes-Riley, K. M., Litman, D., 2004b. Modelling user satisfaction and student learning in a spoken dialogue tutoring system with generic, tutoring, and user affect parameters. In: *Proc. of the Human Language Technology Conference - North American chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL'06)*. New York, USA, pp. 264–271.
- Gales, M. J. F., Woodland, P. C., 1996. Mean and Variance Adaptation Within the MLLR Framework. *Computer Speech and Language* 10, 249–264.
- Gao, Y., Gu, L., Jeff, H.-K., 2005. Portability challenges in developing interactive dialogue systems. In: *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)*. Vol. 5. Philadelphia, PA, USA, pp. 18–23.
- Gauvain, J., 1999. The limsi sdr system for trec. In: *Proc. of the 8th Text Retrieval Conference (TREC 8)*. Gaithersburg, Maryland, USA, pp. 475–482.

- Gauvain, J. L., Lee, C. H., 1994. Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. *IEEE Transactions on Speech and Audio Processing* 2, 291–298.
- Gebhard, P., Klesen, M., Rist, T., 2004. Coloring multi-character conversations through the expression of emotions. In: *Proc. of Tutorial and Research Workshop on Affective Dialogue Systems*. Kloster Irsee, Germany, pp. 128–141.
- Gerfen, C., 2002. Andalusian codas. *Probus* 14, 247–277.
- Geutner, P., Steffens, F., Manstetten, D., 2002. Design of the VICO Spoken Dialogue System: Evaluation of User Expectations by Wizard-of-Oz experiments. In: *Proc. of the 3rd International Conference on Language Resources and Evaluation (LREC'02)*. Las Palmas de Gran Canaria, Spain, pp. 385–400.
- Gibbon, D., Mertins, I., Moore, R., 2000. *Handbook of Multimodal and Spoken Dialogue Systems. Resources, Terminology and Product Evaluation*. Kluwer Academic Publishers (Kluwer International Series in Engineering and Computer Science, 565).
- Gibbon, D., Moore, R., Winski, R., 1997. *Handbook of Standards and Resources for Spoken Language Systems*. Walter de Gruyter.
- Glass, J., Flammia, G., Goodine, D., Phillips, M., Polifroni, J., Sakai, S., Seneff, S., Zue, V., 1995. Multilingual spoken-language understanding in the MIT Voyager system. *Speech Communication* 17 (1-2), 1–18.
- Glass, J. R., 1999. Challenges for Spoken Dialogue Systems. In: *Proc. of the Automatic Speech Recognition and Understanding Workshop (ASRU'99)*. Keystone, Colorado, USA.
- González, G. M., 1999. Bilingual computer-assisted psychological assessment: An innovative approach for screening depression in chicanos/latinos. Tech. Rep. 39, University of Michigan, Ann Arbor, USA.
- Griol, D., 2007. *Desarrollo y Evaluación de diferentes Metodologías para la Gestión Automática del Diálogo*. Ph.D. thesis, Universidad Politécnica de Valencia, Valencia, Spain.

- Gupta, N., Tur, G., Hakkani-Tur, D., Bangalore, S., Riccardi, G., Gilbert, M., 2006. The AT&T spoken language understanding system. *IEEE Transactions on Audio, Speech and Language Processing* 14 (1), 213–222.
- Gustafson, J., Bell, L., Beskow, J., Boye, J., Carlson, R., Edlund, J., Granström, B., House, D., Wirén, M., 2000. AdApt - a multimodal conversational dialogue system in an apartment domain. In: *Proc. of the 6th International Conference on Spoken Language Processing (ICSLP'00)*. Vol. 2. Beijing, China, pp. 134–137.
- Gustafson, J., Lindberg, N., Lundeberg, M., 1999. The August spoken dialogue system. In: *Proc. of the 6th European Conference on Speech Communication and Technology (EuroSpeech'99)*. Budapest, Hungary, pp. 1151–1154.
- Gut, U., Bayerl, P. S., 2004. Measuring the reliability of manual annotations of speech corpora. In: *Proc. of the 2nd International Conference on Speech Prosody (SP'04)*. Nara, Japan, pp. 565–568.
- Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research* 3, 1157–1182.
- Haas, J., Gallwitz, F., Horndasch, A., Huber, R., Warnke, V., 2005. Telephone-based speech dialog systems. *Lecture Notes in Computer Science* 3663, 125–132.
- Hall, L., Woods, S., Aylett, R., Paiva, A., Newall, L., 2005. Achieving empathic engagement through affective interaction with synthetic characters. In: *Proc. of the 1st International Conference on Affective Computing and Intelligent Interaction (ACII'05)*. Beijing, China, pp. 731–738.
- Hamerich, S., de Córdoba, R., Schless, V., d'Haro, L., Schubert, V., Kocsis, O., Igel, S., Pardo, J. M., 2004. The GEMINI Platform: Semi-Automatic Generation of Dialogue Applications. In: *Proc. of the 8th International Conference on Spoken Language Processing (Interspeech-ICSLP)*. Vol. 4. Jeju, Korea, pp. 2629–2632.
- Hanna, P., O'Neill, I., Wootton, C., McTear, M., 2007. Promoting extension and reuse in a spoken dialog manager: an evaluation of the Queen's

-
- Communicator. *ACM Transactions on Speech and Language Processing* 4 (3).
- Hansen, J. H. L., 1996. Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. *Speech communication* 20 (2), 151–170.
- Hardy, H., Biermann, A., Inouye, R., McKenzie, A., Strzalkowski, T., Ursu, C., Webb, N., Wu, M., 2006. The Amitiés system: Data-driven techniques for automated dialogue. *Speech Communication* 48, 354–373.
- Hartikainen, M., Salonen, E.-P., Turunen, M., 2004. Subjective Evaluation of Spoken Dialogue Systems Using SERVQUAL Method. In: *Proc. of 8th International Conference on Spoken Language Processing (Interspeech'04-ICSLP)*. Jeju Island, Korea, pp. 2273–2276.
- Haseel, L., Hagen, E., 2005. Adaptation of an automotive dialogue system to users' expertise. In: *Proc. of 9th International Conference on Spoken Language Processing (Interspeech'05-Eurospeech)*. Lisbon, Portugal, pp. 222–226.
- Heim, J., Nilsson, E. G., Skjetne, J. H., 2007. User Profiles for Adapting Speech Support in the Opera Web Browser to Disabled Users. *Lecture Notes in Computer Science* 4397, 154–172.
- Higashinaka, R., Miyazaki, N., Nakano, M., Aikawa, K., 2004. Evaluating discourse understanding in spoken dialogue systems. *ACM Transactions on Speech and Language Processing* 1, 1–20.
- Holmes, W., Huckvale, M., 1994. Why have HMMs been so successful for automatic speech recognition and how might they be improved? *Speech Hearing and Language* 8, 1875–1878.
- Hozjan, V., Kacic, Z., Moreno, A., Bonafonte, A., Nogueiras, A., 2002. Interface databases: Design and collection of a multilingual emotional speech database. In: *Proc. of the 3rd International Conference on Language Resources and Evaluation (LREC'02)*. Las Palmas de Gran Canaria, Spain, pp. 385–400.

- Hu, Y., Loizou, P. C., 2007. Subjective comparison and evaluation of speech enhancement algorithms. *Speech Communication* 49, 588–601.
- Hubal, R. C., Frank, G. A., Guinn, C. I., 2000. AVATALK Virtual Humans for Training with Computer Generated Forces. In: *Proc. of the 9th Conference on Computer Generated Forces and Behavioral Representation*. Orlando, Florida, USA, pp. 617–623.
- Humaine emotion-research.net, 2007. <http://emotion-research.net/>.
- Hurtig, T., 2004. Visualization and multimodality: a mobile multimodal dialogue system for public transportation navigation evaluated. In: *Proc. of the 8th Conference on Human-computer interaction with mobile devices and services (MobileHCI'06)*. Helsinki, Finland, pp. 251–254.
- Interspeech 2007 Special Session on Speech and language technology for less-resourced languages, 2007. http://www.interspeech2007.org/Technical/less_resourced_languages.php.
- Intrepid Project, 2004. <http://www.intrepid-project.org/>.
- Iriondo, I., Gaus, R., Rodríguez, A., Lázaro, P., Montoya, N., Blanco, J. M., Bernadas, D., Oliver, J. M., Tena, D., Longhi, L., 2000. Validation of an acoustical modelling of emotional expression in spanish using speech synthesis techniques. In: *Proc. of the International Speech Communication Association (ISCA) Workshop on Speech and Emotion*. Newcastle, Northern Ireland, UK, pp. 161–166.
- Johnstone, T., 1996. Emotional speech elicited using computer games. In: *Proc. of the 4th International Conference on Spoken Language Processing (ICSLP'96)*. Vol. 3. Philadelphia, PA, pp. 1985–1988.
- Jokinen, K., 2003. Natural interaction in spoken dialogue systems. In: *Proc. of the Workshop Ontologies and Multilinguality in User Interfaces*. Crete, Greece, pp. 730–734.
- Jokinen, K., Kanto, K., Rissanen, J., 2004. Adaptive User Modelling in AthosMail. *Lecture Notes in Computer Science* 3196, 149–158.

-
- Jurafsky, D., Martin, J. H., 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. Prentice-Hall.
- Kacic, Z., 1999. Advances in spoken dialogue systems development. In: *Proc. of IEEE International Symposium on Industrial Electronics (ISIE'99)*. Vol. 1. Bled, Slovenia, pp. 169–172.
- Kirchhoff, K., Bilmes, J. A., 1999. Statistical acoustic indications of coarticulation. In: *Proc. of International Congress of Phonetic Sciences*. San Francisco, California, USA, pp. 1729–1732.
- Kirchhoff, K., Vergyri, D., 2005. Cross-dialectal data sharing for acoustic modeling in Arabic speech recognition. *Speech Communication* 46, 37–51.
- Krippendorff, K., 2003. *Content Analysis: An Introduction to its Methodology*. Sage Publications, Inc.
- Krsmanovic, F., Spencer, C., Jurafsky, D., Ng, A. Y., 2006. Have we meet? MDP Based Speaker ID for Robot Dialogue. In: *Proc. of the 9th International Conference on Spoken Language Processing (Interspeech'06-ICSLP)*. Pittsburgh, USA, pp. 461–464.
- Kumar, C. S., Mohandas, V. P., Haizhou, L., 2005. Multilingual speech recognition: A unified approach. In: *Proc. of the 9th International Conference on Spoken Language Processing (Interspeech'05-Eurospeech)*. Lisbon, Portugal, pp. 3357–3360.
- Kwon, O., Yoo, K., Suh, E., 2005. ubiES: An Intelligent Expert System for Proactive Services Deploying Ubiquitous Computing Technologies. In: *Proc. of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05)*. Hawaii, pp. 85–86.
- Lamel, L., Bennacef, S., Gauvain, J., Dartigues, H., Temem, J., 2002. User evaluation of the MASK kiosk. *Speech Communication* 38 (1-2), 131–139.
- Lamel, L., Minker, W., Paroubek, P., 2000a. Towards best practice in the development and evaluation of speech recognition components of a spoken language dialog system. *Natural Language Engineering* 6 (3-4), 305–322.

- Lamel, L., Rosset, S., Gauvain, J., Bennacef, S., Garnier-Rizet, M., Prouts, B., 2000b. The LIMSI ARISE system. *Speech Communication* 31, 339–353.
- Landis, J. R., Koch, G. G., 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174.
- Langner, B., Black, A., 2005. Using speech in noise to improve understandability for elderly listeners. In: *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'05)*. San Juan, Puerto Rico, pp. 392–396.
- Lantz, C. A., Nebenzahl, E., 1996. Behavior and interpretation of the κ statistic: Resolution of the two paradoxes. *Journal of Clinical Epidemiology* 49 (4), 431–434.
- Larsen, L. B., 2003. Issues in the evaluation of spoken dialogue systems using objective and subjective measures. In: *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'03)*. St. Thomas, U.S. Virgin Islands, USA, pp. 209–214.
- Lee, C., Yoo, S. K., Park, Y. J., Kim, N. H., Jeong, K. S., Lee, B. C., 2005. Using Neural Network to Recognize Human Emotions from Heart Rate Variability and Skin Resistance. In: *Proc. of the 27th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS'05)*. Shanghai, China, pp. 5523–5525.
- Lee, C. M., Narayanan, S. S., 2005. Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing* 13 (2), 293–303.
- Legris, P., Ingham, J., Collerette, P., 2003. Why do people use information technology: A critical review of the technology acceptance model. *Information and Management* 40, 191–204.
- Lehmann, S., Oepen, S., Regnier-Prost, S., Netter, K., Lux, V., Klein, J., Falkedal, K., Fouvry, F., Estival, D., Dauphin, E., Compagnion, H., Baur, J., Balkan, L., Arnold, D., 1996. TSNLP — Test Suites for Natural Language Processing. In: *Proc. of the 16th International Conference on Computational Linguistics (COLING'96)*. Copenhagen, Denmark, pp. 711–716.

- Lemon, O., Bracy, A., Gruenstein, A., Peters, S., 2001. The Witas Multi-Modal Dialogue System. In: Proc. of Eurospeech'01. Aalborg, Denmark, pp. 1559–1562.
- Lemon, O., Georgila, K., Henderson, J., 2006. Evaluating Effectiveness and Portability of Reinforcement Learned Dialogue Strategies with real users: the TALK TownInfo Evaluation. In: Proc. of IEEE-ACL Workshop on Spoken Language Technology (SLT 2006). Palm Beach, Aruba, pp. 178–181.
- Ligozat, A.-L., Grau, B., Robba, I., Vilnat, A., 2006. Evaluation and improvement of cross-lingual question answering strategies. In: Proc. of the 11th EACL Workshop on Multilingual Question Answering (MLQA'06). Trento, Italy, pp. 23–30.
- Lines, L., Hone, K. S., 2002. Older adults' evaluations of speech output. In: Proc. of the 5th International ACM SIGCAPH Conference on Assistive Technologies (ASSETS'02). Edinburgh, Scotland, pp. 170–177.
- Liscombe, J., Riccardi, G., Hakkani-Tür, D., 2005. Using context to improve emotion detection in spoken dialog systems. In: Proc. of the 9th International Conference on Spoken Language Processing (Interspeech'05-Eurospeech). Lisbon, Portugal, pp. 1845–1848.
- Litman, D. J., Forbes-Riley, K., 2006. Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. *Speech communication* 48 (5), 559–590.
- Litman, D. J., Pan, S., 2002. Designing and evaluating an adaptive spoken dialogue system. *User modelling and user-adapted interaction* 12, 111–137.
- Litman, D. J., Silliman, S., 2004. ITSPOKE: An Intelligent Tutoring Spoken Dialogue System. In: Proc. of the Human Language Technology Conference: 4th Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL). Boston, USA, pp. 233–236.
- López-Cózar, R., Araki, M., 2005. Spoken, Multilingual and Multimodal Dialogue Systems. Development and Assessment. John Wiley and Sons.

- LREC 2008 Workshop on Collaboration: interoperability between people in the creation of language resources for less-resourced languages, 2008. <http://ixa2.si.ehu.es/saltmil/en/home/>.
- MagiCster Project Pages, 2007. <http://www.ltg.ed.ac.uk/magicster/>.
- Mahlke, S., 2006. Emotions and EMG measures of facial muscles in interactive contexts. In: Proc. of the 2006 Conference on Human Factors in Computer Systems (CHI'06). Montreal, Canada.
- Mangold, H., 2001. Speech technology in reality - Applications, their challenges and solutions. In: Proc. of the 4th International Conference on Text, Speech and Dialogue (TSD'01). Pilsen, Czech Republic, pp. 197–200.
- Martinovski, B., Traum, D., 2003. Breakdown in human-machine interaction: the error is the clue. In: Proc. of the ISCA Tutorial and Research Workshop on Error Handling in Dialogue Systems. Chateau d'Oex, Vaud, Switzerland, pp. 11–16.
- Martín-Valdivia, M. T., Martínez-Santiago, F., Ureña-López, L., 2005. Merging strategy for cross-lingual information retrieval systems based on learning vector quantization. *Neural Processing Letters* 22 (2), 149–161.
- Mattasoni, M., Omologo, M., Santarelli, A., Svaizer, P., 2002. On the Joint Use of Noise Reduction and MLLR Adaptation for In-Car Hands-Free Speech Recognition. In: Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP'02). Orlando, USA, pp. 289–292.
- McTear, M. F., 2004. Spoken dialogue technology. Springer.
- MEGA Project, 2001. <http://www.megaproject.org/>.
- Melin, H., Sandell, A., Ihse, M., 2001. CTT-bank: A speech controlled telephone banking system - an initial evaluation. *TMH-QPSR* 1, 1–27.
- Menezes, P., Lerasle, F., Dias, J., Germa, T., 2007. Towards an interactive humanoid companion with visual tracking modalities. In: Hackel, M. (Ed.), *Humanoid robots: Human-like machines*. Itech, pp. 367–398.

- Mihalcea, R., Leong, B., 2006. Toward communicating simple sentences using pictorial representations. In: Proc. of the 7th Conference of the Association for Machine Translation in the Americas (AMTA'06). Boston, MA, USA, pp. 119–127.
- Minker, W., 1998. Stochastic versus rule-based speech understanding for information retrieval. *Speech Communication* 25(4), 223–247.
- Minker, W., Albalade, A., Bühler, D., Pittermann, A., Pittermann, J., Strauss, P.-M., Zaykovskiy, D., 2006a. Recent trends in spoken language dialogue systems. In: ITI 4th International Conference on Information and Communications (ICICT'06). Montreal, Canada, pp. 1–16, invited paper.
- Minker, W., Haiber, U., Heisterkamp, P., Scheible, S., 2004a. The Seneca Spoken Language Dialogue System. *Speech Communication* 43, 1–2.
- Minker, W., Haiber, U., Heisterkamp, P., Scheible, S., 2004b. The SENECA spoken language dialogue system. *Speech Communication* 43 (1-2), 89–102.
- Minker, W., Pittermann, J., Pittermann, A., Strauss, P.-M., Bühler, D., 2006b. Next-generation human-computer interfaces - Towards intelligent, adaptive and proactive spoken language dialogue systems. In: Proc. of the 2nd IET International Conference on Intelligent Environments (IE'06). Vol. 1. Athens, Greece, pp. 213–219.
- Mäkelä, K., Salonen, E., M., T., Hakulinen, J., Raisamo, R., 2001. Evaluating the User Interface of a Ubiquitous Computing system Doorman. In: Proc. of the 3rd International Conference on Ubiquitous Computing (UbiComp'01). Atlanta, USA.
- Möller, S., 2002. A new taxonomy for the quality of telephone services based on spoken dialogue systems. In: Proc. of the 3rd Workshop on Discourse and Dialogue (SIGDial'02). Philadelphia, USA, pp. 142–153.
- Möller, S., 2005. Quality of telephone-based spoken dialogue systems. Springer.
- Möller, S., Smeele, P., Boland, H., Krebber, J., 2007. Evaluating spoken dialogue systems according to de-facto standards: A case study. *Computer Speech and Language* 21 (1), 26–53.

- Montero, J. M., Gutiérrez-Arriola, J., Enríquez, E., Pardo, J. M., 1999. Analysis and modelling of emotional speech in Spanish. In: Proc. of the 14th International Conference of Phonetic. San Francisco, USA, pp. 957–960.
- Montoro, G., Alamán, X., Haya, P. A., 2004. Spoken interaction in intelligent environments: a working system. In: Ferscha, A., Hoertner, H., Kotsis, G. (Eds.), *Advances in Pervasive Computing*. Austrian Computer Society (OCG), pp. 747–754.
- Montoro, G., Haya, P. A., Alamán, X., López-Cózar, R., Callejas, Z., 2006. A proposal for an XML definition of a dynamic spoken interface for ambient intelligence. In: *International Conference on Intelligent Computing (ICIC'06)*. Kunming, China, pp. 711–716.
- Morgan, N., Fosler, E., Mirghafori, N., 1997. Speech recognition using on-line estimation of speaking rate. In: *Proc. of the 5th European Conference on Speech Communication and Technology (Eurospeech'97)*. Rhodes, Greece, pp. 2079–2082.
- Morrison, D., Wang, R., Silva, L. C. D., 2007. Ensemble methods for spoken emotion recognition in call-centers. *Speech communication* 49 (2), 98–112.
- Mostow, J., 2008. Experience from a reading tutor that listens: Evaluation purposes, excuses, and methods. In: Kinzer, C., Verhoeven, L. (Eds.), *Interactive Literacy Education: Facilitating Literacy Environments Through Technology*. New York: Lawrence Erlbaum Associates, Taylor and Francis, pp. 117–148.
- Mourão, M., Cassaca, R., Mamede, N., 2004. An Independent Domain Dialogue System Through a Service Manager. In: *Proc. of the 4th International Conference on Advances in Natural Language Processing*. Alicante, Spain, pp. 161–171.
- Nakamura, S., Markov, K., Jitsuhiro, T., Zhang, J.-S., Yamamoto, H., Kikui, G., 2004. Multi-lingual speech recognition system for speech-to-speech translation. In: *Proc. of 8th International Conference on Spoken Language Processing (Interspeech'04-ICSLP)*. Jeju Island, Korea, pp. 146–154.

- Narayanan, S., Fabbriozio, G. D., Kamm, C., Hubbell, J., Buntschuh, B., Ruscitti, P., Wright, J., 2000. Effects of dialog initiative and multi-modal presentation strategies on large directory information access. In: Proc. of the 6th International Conference on Spoken Language Processing (IC-SLP'00). Vol. 2. Beijing, China, pp. 636–639.
- NECA project, 2005. <http://www.ofai.at/research/nlu/NECA>.
- Nguyen, A., Wobcke, W., 2006. Extensibility and Reuse in an Agent-Based Dialogue Model. In: Proc. of IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology. pp. 367–371.
- NICE project - Main page, 2007. <http://www.niceproject.com/>.
- Nielsen, J., 1994. Usability Engineering. Morgan Kaufmann Ed., San Francisco, USA.
- Nieuwoudt, C., Botha, E. C., 1999. Cross-language adaptation of acoustic models in automatic speech recognition. In: Proc. of 5th Africon Conference in Africa (IEEE Africon 1999). Cape Town, South Africa, pp. 181–184.
- Nieuwoudt, C., Botha, E. C., 2002. Cross-language use of acoustic information for automatic speech recognition. *Speech Communication* 38, 101–113.
- Németh, G., Zainkó, C., 2003. Multilingual statistical text analysis, Zipf's law and Hungarian speech generation. *Acta Linguistica Hungarica* 49, 385–405.
- Nouza, J., Nouza, T., Cerva, P., 2005. A multi-functional voice-control aid for disabled persons. In: Proc. of International Conference on Speech and Computer (SPECOM'05). Patras, Greece, pp. 715–718.
- Nouza, J., Psutka, J., Uhlír, J., 1997. Phonetic Alphabet for Speech Recognition of Czech. *Radioengineering* 6 (4), 16–20.
- Oh, A., Rudnicky, A., 2000. Stochastic language generation for spoken dialogue systems. In: Proc. of ANLP/NAACL 2000 Workshop on Conversational Systems. Seattle, USA, pp. 27–32.
- O'Neill, P., 2005. Utterance final /s/ in Andalusian Spanish. The phonetic neutralization of a phonological contrast. *Language Design* 7, 151–166.

- Ortony, A., G. L. Clore, A. C., 1988. *The cognitive structure of emotions*. Cambridge University Press.
- Ostendorf, M., Digalakis, V., Kimball, O., 1996. From hmm's to segment models: a unified view of stochastic modelling for speech recognition. *IEEE Transactions on Speech and Audio Processing* 4 (5), 360–378.
- Oviatt, S., DeAngeli, A., Kuhn, K., 1997. Integration and synchronization of input modes during multimodal human-computer interaction. In: *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*. Atlanta, Georgia, USA, pp. 415–422.
- Paek, T., 2001. Empirical methods for evaluating dialog systems. In: *Proc. of the Workshop on Evaluation for Language and Dialogue Systems*. Vol. 9. Toulouse, France, pp. 1–8.
- Park, W., Han, S. H., Park, Y. S., Park, J., Yang, H., 2007. A framework for evaluating the usability of spoken language dialogue systems (SLDSs). *Lecture Notes in Computer Science* 4559, 398–404.
- Pérez, G., Amores, G., Manchón, P., 2006. A multimodal architecture for home control by disabled users. In: *Proc. of IEEE/ACL Workshop on Spoken Language Technology (SLT)*. Palm Beach, Aruba, pp. 134–137.
- Persia, L. D., Yanagida, M., Rufiner, H. L., Milone, D., 2007. Objective quality evaluation in blind source separation for speech recognition in a real room. *Signal Processing* 87, 1951–1963.
- PF-STAR home page, 2004. <http://pfstar.itc.it/>.
- Pfau, T., Ruske, G., 1998. Estimating the speaking rate by vowel detection. In: *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'98)*. Vol. 2. Seattle, Washington, USA, pp. 945–948.
- Picard, R. W., 1997. *Affective Computing*. The MIT Press, Cambridge, Massachusetts.
- Picard, R. W., Daily, S. B., 2005. Evaluating affective interactions: Alternatives to asking what users feel. In: *Proc. of the 2005 Conference on Human*

-
- Factors in Computer Systems (CHI'05), Workshop: Evaluating Affective Interfaces-Innovative Approaches. Portland, Oregon, USA.
- Pieraccini, R., Levin, E., Eckert, W., 1997. AMICA: The AT&T mixed initiative conversational architecture. In: Proc. of European Conference on Speech Communications and Technology (Eurospeech'97). Rhodes, Greece, pp. 1875–1878.
- Pitterman, J., Pitterman, A., 2006. Integrating emotion recognition into an adaptive spoken language dialogue system. In: Proc. of the 2nd IEEE International Conference on Intelligent Environments. Athens, Greece, pp. 197–202.
- Plutchik, R., 1980. EMOTION: A psychoevolutionary synthesis. Harper and Row publishers.
- Polzin, T., Waibel, A., 2000. Emotion-sensitive human-computer interfaces. In: Proc. of ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion. Newcastle, Northern Ireland, UK, pp. 201–206.
- Prendinger, H., Mayer, S., Mori, J., Ishizuka, M., 2003. Persona effect revisited. using bio-signals to measure and reflect the impact of character-based interfaces. In: Proc. of the 4th International Working Conference on Intelligent Virtual Agents (IVA'03). Kloster Irsee, Germany, pp. 283–291.
- Radev, D., Teufel, S., Saggion, H., Lam, W., Blitzer, J., Celebi, A., Qi, H., Drabek, E., Liu, D., 2001. Evaluation of text summarization in a cross-lingual information retrieval framework. Tech. rep., Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA.
- Rajman, M., Bui, T. H., Rajman, A., Seydoux, F., Trutnev, A., Quarteroni, S., 2004. Assessing the usability of a dialogue management system designed in the framework of a rapid dialogue prototyping methodology. *Acta Acustica united with Acustica* 90, 1906–1111.
- Raux, A., Bohus, D., Black, A. W., Eskenazi, M., 2006. Doing Research on a Deployed Spoken Dialogue System: One Year of Let's Go! Experience. In: Proc. of the 9th International Conference on Spoken Language Processing (Interspeech'06-ICSLP). Pittsburgh, USA, pp. 65–68.

- Raux, A., Langner, B., Black, A., Eskenazi, M., 2005. Let's Go Public! Taking a Spoken Dialog System to the Real World. In: Proc. of the 9th International Conference on Spoken Language Processing (Interspeech'05-Eurospeech). Lisbon, Portugal, pp. 885–888.
- Raux, A., Langner, B., Black, A. W., Eskenazi, M., 2003. LET'S GO: Improving Spoken Dialog Systems for the Elderly and Non-natives. In: Proc. of the European Conference on Speech Technology (Eurospeech'03). Geneva, Switzerland, pp. 753–756.
- Rayner, M., Hockey, B., Renders, J.-M., Chatzichrisafis, N., Farrell, K., 2005. A voice enabled procedure browser for the International Space Station. In: 43th Annual Meeting of the Association for Computational Linguistics. Ann Arbor, USA, pp. 29–32.
- Riccardi, G., Hakkani-Tür, D., 2005. Grounding emotions in human-machine conversational systems. *Lecture Notes in Computer Science* 3814, 144–154.
- Rich, C., Sidner, C., 1998. COLLAGEN: A Collaboration Manager for Software Interface Agents. *User Modeling and User-Adapted Interaction* 8, 315–350.
- Robinson, S. M., Roque, A., Vaswani, A., Traum, D., 2006. Evaluation of a spoken dialogue system for virtual reality call for fire training. In: Proc. of the 25th Army Science Conference. Orlando, USA.
- Roque, A., Ai, H., Traum, D., 2006a. Evaluation of an information state-based dialogue manager. In: Proc. of the 10th Workshop on the Semantics and Pragmatics of Dialogue (Brandial'06). Potsdam, Germany, pp. 181–182.
- Roque, A., Leuski, A., Rangarajan, V., Robinson, S., Vaswani, A., Narayanan, S., Traum, D., 2006b. Radiobot-CFF: A Spoken Dialogue System for Military Training. In: Proc. of the 9th International Conference on Spoken Language Processing (Interspeech'06-ICSLP). Pittsburgh, USA, pp. 477–480.
- Rotaru, M., Litman, D. J., 2006. Discourse structure and speech recognition problems. In: Proc. of the 9th International Conference on Spoken Language Processing (Interspeech'06-ICSLP). Pittsburgh, USA, pp. 53–56.

- Rudnicky, A., Thayer, E., Constantinides, P., Tchou, C., Shern, R., Lenzo, K., Xu, W., Oh, A., 1999. Creating natural dialogs in the Carnegie Mellon Communicator system. In: Proc. of European Conference on Speech Communications and Technology (Eurospeech'99). Vol. 1. pp. 1531–1534.
- Rumelhart, D. E., Hinton, G. E., Williams, R. J., 1986. PDP: Computational models of cognition and perception, I. MIT Press.
- Russell, J. A., 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39, 1161–1178.
- Sanderman, A., Sturm, J., den Os, E., Boves, L., Cremers, A., 1998. Evaluation of the Dutch train timetable information system developed in the ARISE project. In: Proc. of IEEE 4th Workshop on Interactive Voice Technology for Telecommunications Applications (IVTTA'98). Torino, Italy, pp. 91–96.
- Sarukkai, R., Hunter, C., 1997. Integration of eye fixation information with speech recognition systems. In: Proc. of the 5th European Conference on Speech Communication and Technology (Eurospeech'97). Rhodes, Greece, pp. 1639–1643.
- Schalkwyck, J., Story, L. H. E., 2003. Speech recognition with Dynamic Grammars Using Finite-State Transducer. In: Proc. of Eurospeech'03. Geneva, Switzerland, pp. 1969–1972.
- Scherer, K. R., 2005. What are emotions? and how can they be measured? *Social Science Information* 44 (4), 694–729.
- Schiel, F., 2006. Evaluation of multimodal dialogue systems. In: Wahlster (2006), pp. 617–643.
- Schneider, M., 2004. Towards a Transparent Proactive User Interface for a Shopping Assistant. In: Proc. of Workshop on Multi-User and Ubiquitous User Interfaces (MU3I). Vol. 3. Funchal, Madeira, Portugal, pp. 10–15.
- Schultz, T., Kirchhoff, K., 2006. *Multilingual Speech Processing*. Elsevier.
- Schultz, T., Waibel, A., 1998. Language independent and language adaptive large vocabulary speech recognition. In: Proc. of the 5th International

- Conference of Spoken Language Processing (ICSLP'98). Vol. 5. Sidney, Australia, pp. 1819–1822.
- Scott, W., 1955. Reliability of content analysis: the case of nominal scale coding. *Public opinion quarterly* 19 (3), 321–325.
- Seneff, S., Hurley, E., Lau, R., Pao, C., Schmid, P., Zue, V., 1998. Galaxy-II: A Reference Architecture for Conversational System Development. In: Proc. of the 5th International Conference on Spoken Language Processing (ICSLP'98). Vol. 3. Sydney, Australia, pp. 931–934.
- Seneff, S., Polifroni, J., 2000. Dialogue management in the Mercury flight reservation system. In: Proc. of ANLP-NAACL Workshop on Conversational systems. Vol. 3. Seattle, Washington, USA, pp. 11–16.
- Shafran, I., Mohri, M., 2005. A comparison of classifiers for detecting emotion from speech. In: Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05). Philadelphia, PA, USA, pp. 341–344.
- Shafran, I., Riley, M., Mohri, M., 2003. Voice signatures. In: Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'03). St. Thomas, U.S. Virgin Islands, USA, pp. 31–36.
- Sim, J., Wright, C. C., 2005. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical Therapy* 85 (3), 257–268.
- Skantze, G., Edlund, J., Carlson, R., 2006. Talking with Higgins: Research challenges in a spoken dialogue system. In: Proc. of Perception and Interactive Technologies (PIT'06). Kloster Irsee, Germany, pp. 193–196.
- Stern, R., Liu, F., Ohshima, Y., Sullivan, T., Acero, A., 1992. Multiple approaches to robust speech recognition. In: Proc. of the 5th DARPA Speech and Natural Language Workshop. Harriman, New York, USA, pp. 274–279.
- Stewart, J., 1922. An electrical analog of the vocal tract. *Nature* 110, 311–312.

- Stibbard, R., 2000. Automated extraction of tobi annotation data from the reading/leeds emotional speech corpus. In: Proc. of the International Speech Communication Association (ISCA) Workshop on Speech and Emotion. Newcastle, Northern Ireland, UK, pp. 60–65.
- Sturm, J., Cranen, B., Terken, J., Bakx, I., 2005. Effects of prolonged use on the usability of a multimodal form-filling interface. In: Minker, W., Bühler, D., Dybkjaer, L. (Eds.), Spoken multimodal human-computer dialogue in mobile environments. Springer, pp. 329–348.
- Sun, D., 1997. Statistical modeling of co-articulation in continuous speech based on data driven interpolation. In: Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'97). Munich, Germany, pp. 1751–1754.
- The Safira Project - DFKI Page, 2002. <http://www2.dfki.de/imedia/safira/>.
- Traum, D., Larsson, S., 2003. The Information State Approach to Dialogue Management. Current and New Directions in Discourse and Dialogue. Kluwer Academic Publishers.
- TRINDIConsortium, 2001. TRINDI (Task Oriented Instructional Dialogue) Book Draft. <http://www.ling.gu.se/projekt/trindi/book.ps> .
- Truillet, P., Grisvard, O., Goujon, B., 2004. SCOPE - CARE II Innovative WP3 -R3 - Model of English. Tech. rep., European Organisation for the Safety of Air Navigation.
- Turing, A., 1950. Computing machinery and intelligence. *Mind* 236, 433–460.
- Turunen, M., Hakulinen, J., Kainulainen, A., 2006. Evaluation of a spoken dialogue system with usability tests and long-term pilot studies: Similarities and differences. In: Proc. of the 9th International Conference on Spoken Language Processing (Interspeech'06-ICSLP). Pittsburgh, USA, pp. 1057–1060.
- Turunen, M., Salonen, E., Hartikainen, M., Hakulinen, J., Black, W., Ramsay, A., Funk, A., Conroy, A., Thompson, P., Stairmand, M., Jokinen, K., Rissanen, J., Kanto, K., Kerminen, A., Gamback, B., Cheadle, M.,

- Olsson, F., Sahlgren, M., 2004. Athosmail: a multilingual adaptive spoken dialogue system for the e-mail domain. In: Proc. of Workshop on Robust and Adaptive Information Processing for Mobile Speech Interfaces. Geneva, Switzerland, pp. 77–86.
- Vaquero, C., Saz, O., Lleida, E., Marcos, J., Canalís, C., 2006. VOCALIZA: An application for computer-aided speech therapy in spanish language. In: Proc. of IV Jornadas en Tecnología del Habla. Zaragoza, Spain, pp. 321–326.
- Ververidis, D., Kotropoulos, C., 2006. Emotional speech recognition: resources, features and methods. *Speech communication* 48, 1162–1181.
- VICTEC in Lynne Hall web page, 2005. <http://osiris.sunderland.ac.uk/~cs01ha/Research/victec.html>.
- Vidrascu, L., Devillers, L., 2005. Real-life emotion representation and detection in call centers data. *Lecture Notes in Computer Science* 3784, 739–746.
- Villing, J., Larsson, S., 2006. Dico - A Multimodal Menu-based In-vehicle Dialogue System. In: Proc. of the 10th Workshop on the Semantics and Pragmatics of Dialogue (BRANDIAL'06). Potsdam, Germany, pp. 187–188.
- Vogt, T., André, E., 2005. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In: Proc. of IEEE International Conference on Multimedia and Expo. pp. 474–477.
- Wahlster, W. (Ed.), 2006. *SmartKom: Foundations of Multimodal Dialogue Systems*. Springer.
- Waibel, A., Suhm, B., Vo, M., Yang, J., 1997. Multimodal Interfaces for Multimedia Information Agents. In: Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'97). Munich, Germany, pp. 167–170.
- Walker, M., Cahn, J., Whittaker, S., 1997. Improvising linguistic style: Social and affective bases of agent personality. In: Proc. of the 1st International Conference on Autonomous Agents (Agents'97). Marina del Rey, CA, USA, pp. 96–105.

- Walker, M., Fromer, J., Fabbriozio, G., Mestel, C., Hindle, D., 1998a. What can I say? Evaluating a spoken language interface to Email. In: Proc. of ACM CHI 98 Conference on Human Factors in Computing Systems. Los Angeles, USA, pp. 582–589.
- Walker, M., Kamm, C. A., Litman, D. J., 2000a. Towards developing general models of usability with paradise. *Natural Language Engineering* 6 (3-4), 363–377.
- Walker, M., Langkilde, I., wright, J., Gorin, A., Litman, D., 2000b. Learning to predict problematic situations in a spoken dialogue system: Experiments with How May I Help You. In: Proc. of the North American Meeting of the Association for Computational Linguistics. Seattle, USA, pp. 210–217.
- Walker, M., Litman, D., Kamm, C., Abella, A., 1998b. Evaluating spoken dialogue agents with PARADISE: Two case studies. *Computer Speech and Language* 12 (4), 317–347.
- Walker, M., Rudnicky, A., Prasad, R., Aberdeen, J., Bratt, E. O., Garofolo, J., Hastie, H., Le, A., Pellom, B., Potamianos, A., Passonneau, R., Roukos, S., Sanders, G., Seneff, S., Stallard, D., 2002a. DARPA Communicator: Cross-System Results for The 2001 Evaluation. In: Proc. of the 7th International Conference on Spoken Language Processing (Interspeech'02-ICSLP). Vol. 1. Denver, USA, pp. 269–272.
- Walker, M., Rudnicky, A., Prasad, R., Aberdeen, J., Bratt, E. O., Garofolo, J., Hastie, H., Le, A., Pellom, B., Potamianos, A., Passonneau, R., Roukos, S., Sanders, G., Seneff, S., Stallard, D., 2002b. DARPA Communicator Evaluation: Progress from 2000 to 2001. In: Proc. of the 7th International Conference on Spoken Language Processing (Interspeech'02-ICSLP). Vol. 1. Denver, USA, pp. 273–276.
- Weizenbaum, J., 1966. ELIZA - A computer program for the study of natural language communication between man and machine. *Communications of the ACM* 9 (1), 36–45.
- Weng, F., Varges, S., Raghunathan, B., Ratiu, F., Pon-Barry, H., Lathrop, B., Zhang, Q., Scheideck, T., Bratt, H., Xu, K., Purver, M., Mishra, R., Raya, M., Peters, S., Meng, Y., Cavedon, L., Shriberg, L., 2006. CHAT: A

- Conversational Helper for Automotive Tasks. In: Proc. of the 9th International Conference on Spoken Language Processing (Interspeech'06-ICSLP). Pittsburgh, USA, pp. 1061–1064.
- Wilks, Y., 2006. Artificial companions as a new kind of interface to the future internet. Tech. Rep. 13, Oxford Internet Institute.
- Williams, J., Young, S., 2007. Partially Observable Markov Decision Processes for Spoken Dialog Systems. *Computer Speech and Language* 21(2), 393–422.
- Wilting, J., Kraemer, E., Swerts, M., 2006. Real vs. acted emotional speech. In: Proc. of the 9th International Conference on Spoken Language Processing (Interspeech'06-ICSLP). Pittsburgh, USA, pp. 805–808.
- Witten, I. H., Frank, E., 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.
- Zeng, Z., Hu, Y., Fu, Y., Huang, T. S., Roisman, G. I., Wen, Z., 2006. Audio-visual emotion recognition in adult attachment interview. In: Proc. of the 8th International Conference on Multimodal Interfaces (ICMI'06). Banff, Alberta, Canada, pp. 828–831.
- Zgank, A., Kaèiè, Z., Diehl, F., Vicsi, K., Szaszak, G., Juhar, J., Lihan, S., 2004. The COST278 MASPER initiative - Crosslingual speech recognition with large telephone databases. In: Proc. of the 4th International Conference on Language Resources and Evaluation (LREC'04). Lisbon, Portugal, pp. 2107–2110.
- Zong, Y., Dohi, H., Ishizuka, M., 2000. Multimodal presentation markup language mpml with emotion expression functions attached. In: Proc. of the 2nd International Symposium on Multimedia Software Engineering. pp. 359–365.
- Zue, V., Seneff, S., Glass, J., Polifroni, J., Pao, C., Hazen, T., Hetherington, L., 2000. JUPITER: A telephone-based conversational interface for weather information. In: *IEEE Transactions on Speech and Audio Processing*. Vol. 8. pp. 85–96.

Zue, V. W., Glass, J. R., 2000. Conversational Interfaces: Advances and Challenges. Proc. of the IEEE 88 (8), 1166–1180.