SmartKom-English: From Robust Recognition to Felicitous Interaction

David Gelbart[1], John Bryant[1], Andreas Stolcke[1], Robert Porzel[2], Manja Baudis[2] and Nelson Morgan[1]

[1] International Computer Science Institute (ICSI), Berkeley, USA
texttt{gelbart,jbryant,stolcke,morgan}@icsi.berkeley.edu
[2] European Media Laboratory GmbH [EML], Heidelberg, Germany
{robert.porzel,manja.baudis}@eml-d.villa-bosch.de

Summary. This chapter describes the English-language SMARTKOM-Mobile system and related research. We explain the work required to support a second language in SMARTKOM and the design of the English speech recognizer. We then discuss research carried out on signal processing methods for robust speech recognition and on language analysis using the Embodied Construction Grammar formalism. Finally, the results of human-subject experiments using a novel *Wizard and Operator* model are analyzed with an eye to creating more felicitous interaction in dialogue systems.

1 Introduction

The SMARTKOM-Mobile application provides navigation and tourism information using either a handheld personal digital assistant (PDA) interface or an in-car interface. Some images from the application's display are shown in Fig. 1. A user communicates with SMARTKOM-Mobile using pointing gestures and natural speech, and the system responds with speech (from an animated agent, displayed on-screen) and the display of images and text. The natural and reliable conversational interaction that this calls for provided motivation for a range of research. In this section and in Sect. 2 we describe the work required to port SMARTKOM-Mobile to the English language and the design of the English speech recognizer. In the following sections, we describe research on robust speech recognition, language analysis, and human–computer interaction carried out by the SMARTKOM-English team.

The development of an English-language SMARTKOM-Mobile verified the language portability of the SMARTKOM architecture and facilitated the demonstration of SMARTKOM at international conferences. Staff and visiting researchers at ICSI and staff at DFKI took the lead roles in the creation of the English-language system, and important contributions came from several other SMARTKOM partner sites.

The English speech recognizer was developed completely independently from the German one; it is a hybrid connectionist system descended from the one used in

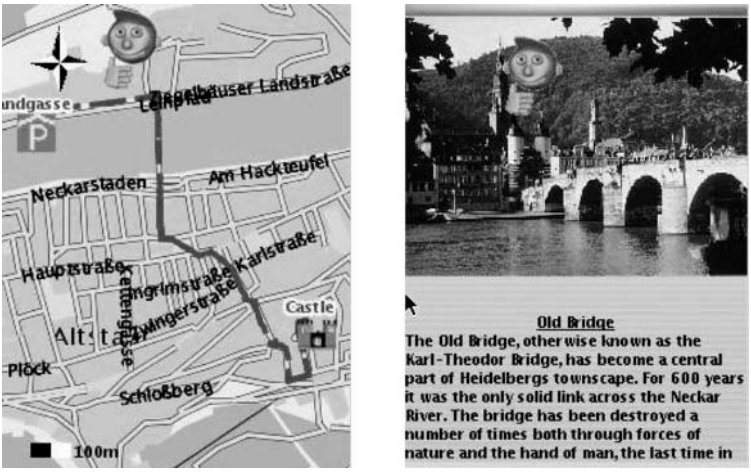

Fig. 1. SMARTKOM-Mobile screen shots showing pedestrian navigation (*left*) and tourist site information (*right*)

the BERP dialogue system project (Jurafsky et al., 1994). All other modules in the English SMARTKOM-Mobile system were based on, or identical to, modules in the German-language SMARTKOM-Mobile.

The modular architecture of SMARTKOM greatly eased porting to English by encapsulating language dependencies in specific modules. Most SMARTKOM modules required no modification to support English. Of the modules requiring modification, only speech recognition and speech synthesis required significant changes to software source code. The speech analyzer (which parses the recognized speech) and text generator (which creates the system's output sentences) required only a change in their template (grammar) files and otherwise used the same software engines for both German and English. The lexicon module had to be aware of the current language in order to provide the correct word pronunciations. Some displayed text provided by the pedestrian and vehicle navigation modules (such as tourist site information and map labels) was translated to English. If dynamic help had been included in the English system, some additional displayed text would have required translation. The speech analyzer outputs a language-independent semantic representation of the user input, and so modules which tracked dialogue state and user intention did not need to be language-aware.

2 The SmartKom-English Speech Recognizer

2.1 Overview

The recognizer uses the hidden Markov model (HMM) approach to speech recognition illustrated in Fig. 2. This approach models speech as a sequence of observations

sampled from different possible probability distributions, with a distinct distribution corresponding to each member of a finite set of possible hidden states. In our recognizer the hidden states represent phones, and the observations are assumed to be acoustic realizations of those phones. Each observation represents a single frame of time. The time step from the start of one frame to the start of the next frame is 16 ms and the length of each frame is 32 ms; the resulting overlap between frames is useful since the frame boundaries are not necessarily aligned with phone boundaries. The observations being modeled are not the original audio but rather are the output of a feature extraction process intended to reduce dimensionality and discard irrelevant variation. We used perceptual linear prediction (PLP) feature extraction (Hermansky, 1990), which captures the envelope of the frame power spectrum but discards some spectral detail. The probability of a particular observation for each possible hidden state is determined by an acoustic modeling stage, which we carry out using a multilayer perceptron (MLP), following the hybrid connectionist approach of Bourlard and Morgan (1993). The probabilities determined by the acoustic modeling stage for all frames are used by a decoding stage that searches for the most likely sentence, taking into account a dictionary of word pronunciations and a language model (tables of word transition probabilities).

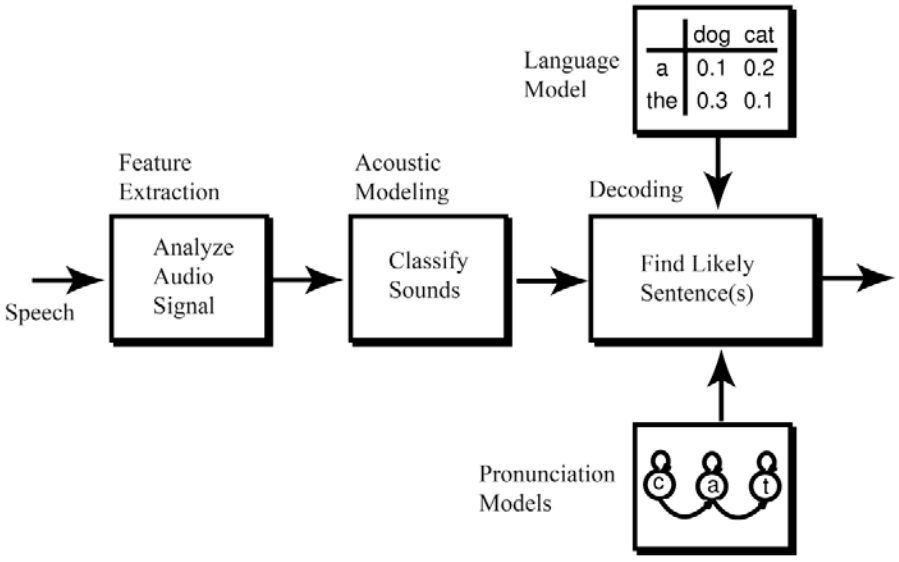

Fig. 2. Speech recognizer architecture

The English speech recognizer was built as a chain of small tools used in a pipeline, communicating with each other using Unix pipes. This simple, modular design makes adding or upgrading speech recognizer components easy. A wrapper program starts the pipeline and handles communication with other SMARTKOM mod-

ules. Feature extraction is performed by ICSI's RASTA tool. Utterance-level mean and variance normalization of the features and MLP output calculation is performed by ICSI's FFWD-NORM tool. Decoding was performed by the NOWAY tool (Renals and Hochberg, 1996); changes made to the NOWAY tool to support the needs of the SMARTKOM project are described below. Source code for all these tools is available free for research use.

2.2 Language Modeling and Decoding

The speech recognizer uses a trigram language model with backoff to bigrams and unigrams. The SRI language modeling toolkit (Stolcke, 2002) was used to estimate the language model from training data that consisted of sample dialogues, partly edited by hand to achieve good coverage of what were perceived to be natural user inputs. Since no naturally collected English data was available, we relied on English translations of German dialogues, taking care to produce idiomatic, rather than literal translations.

To meet the needs of SMARTKOM, we modified the NOWAY decoder to use the C++ libraries associated with the SRI LM toolkit to access language models. This modification allowed NOWAY to be used with class-based language models. Class-based LMs can include class labels as part of N-grams, which are then expanded by a list of class member words. Class-based models have two key advantages for SMARTKOM. First, by using classes the LM generalizes better to novel word sequences, which is especially important given the scarcity of training data. For example, a DIRECTION class was used to stand for possible map directions (e.g., left, right, up, down, east, west, north, and south) in training sentences. To achieve generalization, the word classes are defined by hand based on task knowledge, and the appropriate words are replaced by class labels in the training data.

The second key function of word classes is that they allow new class members to be added on the fly while SMARTKOM is running, without reestimating the entire language model. In the English Mobile application this occurs with parking garage names, which are retrieved from the car navigation module. Parking garage names can occur any time the class name GARAGE occurs in the language model, thereby covering sentences such as

- Can you tell me more about GARAGE
- I would like to know more about GARAGE
- I'd like to know more about GARAGE

3 Signal Processing for Robust Speech Recognition

3.1 Introduction

Compared to recordings made using a close-talking microphone placed near the user's mouth (e.g., a headset microphone), recordings from more distant microphones have higher levels of background noise relative to the speech level (since

the speech level is lower) and are subject to reverberation and other effects due to the longer and sometimes indirect paths taken by the traveling sound waves. While these degradations affect human speech recognition performance as well, current automatic speech recognition systems are much more sensitive to them. However, close-talking microphones are often inconvenient, and improvements in the recognition accuracy that can be achieved without them are very likely to increase adoption of speech recognition technology.

The SMARTKOM-Mobile application can be used inside a car (using an installed display and microphones) or on foot (using the microphone and display on a handheld computer or PDA).[3] Of these two circumstances, ICSI research focused on the in-car case, making use of the SpeechDatCar (Moreno et al., 2000) corpus. The SpeechDatCar corpus is available in several languages; in this article we will only describe results for SpeechDatCar-German. This contains in-car recordings of German connected digit strings made simultaneously with close-talking and hands-free microphones, in various noise conditions: "Stop Motor Running," "Town Traffic," "Low-Speed Rough Road," and "High-Speed Good Road." Our speech recognition experiments with this corpus were performed for three cases: the well-matched case used all microphone types and noise conditions in both training and test data, the medium-matched case used only the hands-free microphone and tests using only the "High-Speed Good Road" noise condition (which is excluded from the training data in this case), and the highly mismatched case used close-talking microphone training data and hands-free microphone test data.

3.2 Noise Reduction and Deconvolution

Figure 3 shows a model of how acoustic degradation is caused by reverberation and background noise. Reverberation and other acoustic effects related to the transmission of speech from talker to microphone, together with the frequency response of the microphone itself, are modeled as a linear time-invariant system with impulse response $c(n)$ and frequency response $C(\omega)$. Background noise (including noise internal to the microphone itself) is additive after this system. We investigated the effectiveness of some signal processing techniques based on this model.

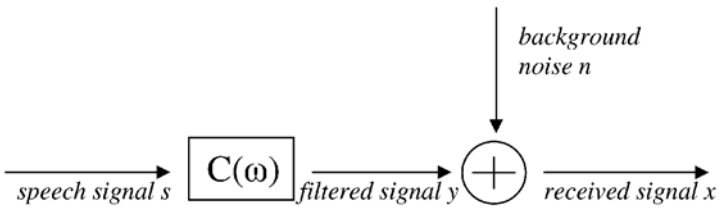

Fig. 3. Model of acoustic degradation

[3] In in-car applications, non–close-talking microphones are sometimes referred to as *hands-free* microphones, as in hands-free mobile phone operation.

3.2.1 Noise Reduction

We used a noise reduction implementation developed for an Aurora (Hirsch and Pearce, 2000) front-end proposal, a joint effort between ICSI, OGI, and Qualcomm engineers, described in Adami et al. (2002). The algorithm performs Wiener filtering with modifications such as a noise overestimation factor, smoothing of the filter response, and a spectral floor. It calculates an instantaneous noise power spectral estimate $|\acute{N}(m,k)|^2$ (where k is the frequency bin and m is the frame index) by averaging the noisy power spectra $|X(m,k)|^2$ over an initial period before speech starts as well as later frames, which are judged to be nonspeech because their energy level falls below a threshold. This estimate is used to calculate a filter

$$|H(m,k)| = max\left(\frac{|X(m,k)|^2 - \alpha|\acute{N}(m,k)|^2}{|X(m,k)|^2}, \beta\right),$$

where α is an SNR-dependent oversubtraction factor and the spectral floor parameter β is used to avoid negative and very small filter values. This filter $|H(m,k)|$ varies from frame to frame. To reduce artifacts in the noise-reduced output, the filter is smoothed over time and frequency. Then the smoothed filter is applied to the noisy power spectra X to obtain an estimate of the noise-free power spectra Y:

$$|\acute{Y}(m,k)|^2 = max(|X(m,k)|^2 * |H(m,k)|^2, \beta_{final} * |\acute{N}(m,k)|^2),$$

where β_{final} is a second spectral floor parameter, which specifies the floor as a fraction of the estimated noise power.

This Wiener filtering approach is based on the assumption that the noise power spectrum is steady. This is probably a fairly good assumption for engine and wind noise during unchanging driving conditions.

Table 1 shows word error rates (WER) on SpeechDatCar-German using the Aurora baseline speech recognizer described in Hirsch and Pearce (2000), with and without noise reduction preprocessing. The noise reduction is very effective in the medium-mismatch condition. In the high mismatch condition it is counterproductive; perhaps the application of noise reduction to the close-talking microphone training data makes that data an even worse match for the hands-free microphone test data. For these experiments, the noise estimation was performed independently for each utterance, and we used overlap-add resynthesis to create noise-reduced output waveforms. This allowed the noise reduction to be used with existing feature extraction code without modifying that code. This noise reduction approach was added to the SMARTKOM-English system as a new pipeline stage preceding the RASTA feature extraction tool.

3.2.2 Deconvolution by Mean Subtraction

Deconvolution by mean subtraction is commonly employed in speech recognition systems, most often via the cepstral mean subtraction (CMS) algorithm. The reasoning behind it is as follows. Consider a discrete-time speech signal $s(n)$ (the origin

Table 1. SpeechDatCar-German word error rates (WER) with and without noise reduction. The well-matched test data contain 5009 words, the medium-matched test data contain 1366 words, and the highly mismatched test data contain 2162 words.

WER (%)	Without noise reduction	With noise reduction
Well-matched	8.0	7.0
Medium-matched	20.6	14.9
Highly mismatched	15.4	17.5

of speech as a continuous-time signal is ignored here for simplicity) sent over a linear time-invariant channel with impulse response $c(n)$, producing a filtered signal $y(n)$. Then the product property holds for the spectra of s and c: $Y(\omega) = S(\omega)C(\omega)$. Assume that the filtered signal is processed using a windowed discrete Fourier transform, giving $Y(m,\omega)$, where m is the frame index determining around which sample the window function was centered. If the window function is long and smooth enough relative to $c(n)$, then the product property approximately still holds (Avendano, 1997): $X(m,\omega) \approx S(m,\omega)C(\omega)$. Note that the channel is assumed not to vary over time. Taking the logs of the magnitudes of both sides of the previous equation, we find that $log|X(m,\omega)| \approx log|S(m,\omega)| + log|C(\omega)|$. Therefore, by subtracting the mean over m (i.e., over time) of $log|X(m,\omega)|$ from $log|X(m,\omega)|$, we remove $log|C(\omega)|$ and the mean over time of $log|S(m,\omega)|$. If $s(n)$ is speech, and the mean is calculated over a large enough number of frames, then we expect the mean of $log|S(m,\omega)|$ to contain little linguistic information, and therefore its removal need not be detrimental to speech recognition performance. Thus the subtraction can be used to compensate for the magnitude response of the channel (for various reasons, there is usually no attempt to compensate the phase response). Speech recognition feature extraction is usually essentially based on windowed discrete Fourier transforms. However, following the transform it is typical to do further processing like Mel/Bark-scale filter bank integration and cepstral transformation, and most often the mean subtraction is done after that processing. The cepstral transformation, being linear, does not affect the reasoning above, but the filter bank integration implies an additional assumption that the channel frequency response is close to constant across the frequency bins that are being integrated.

Table 2 shows WER on SpeechDatCar-German with and without mean subtraction methods added to the Aurora baseline speech recognizer described in Hirsch and Pearce (2000), which uses Mel-frequency cepstral coefficient (MFCC) feature extraction and by default does not perform mean subtraction.

Cepstral mean subtraction (CMS) removes the mean across frames from the MFCCs. The table shows CMS was most helpful in the well-matched and highly mismatched cases, where both close-talking and hands-free recordings are used. This is not surprising because there is a channel mismatch between close-talking and hands-free recordings. The reasoning behind mean subtraction assumes that all frames contain speech, while in fact the data is a mix of speech and pauses. Calculating the mean only over frames judged by a multilayer perceptron classifier to contain speech resulted in a significant performance improvement, which is consistent with the results

for speech recognition over telephone connections in Mokbel et al. (1996). We also tried the log-DFT mean normalization (LDMN) proposed in Neumeyer et al. (1994). In this method the mean subtraction occurs midway through the MFCC computation, before the filter bank integration, so the assumption of channel constancy across bins being integrated is not required. At ICSI, we have sometimes found this to be more effective than CMS, but on this task it does not perform better (in fact, there is a slight, though not statistically significant, drop in performance).

Table 2. SpeechDatCar-German word error rates (WER)

WER (%)	Baseline	Noise reduction alone	Noise red. and CMS	Noise red. and CMS; mean taken over speech frames	Noise red. and LDMN; mean taken over speech frames
Well-matched	8.0	7.0	6.7	6.1	6.1
Medium-matched	20.6	14.9	15.7	14.6	15.2
Highly mismatched	15.4	17.5	14.7	11.3	11.0

For further results, including other mean subtraction methods and other data sets, see Gelbart (2004).

3.3 Gabor Filtering

The noise reduction and mean subtraction approaches described above are intended to increase the robustness of existing feature extraction methods by adding additional processing. Another approach, which can be complementary, is to create new feature extraction methods which have desirable properties. We collaborated with Michael Kleinschmidt of the Universität Oldenburg on his project on Gabor filter feature extraction for automatic speech recognition (Kleinschmidt and Gelbart, 2002; Kleinschmidt, 2002). Gabor filters are a family of two-dimensional filters, which Kleinschmidt proposed to use for feature extraction by convolving Gabor filters with a time-frequency representation such as a mel-band spectrogram (see Fig. 4). Depending on what Gabor filters are used, this can behave similarly to short-term spectral envelope-based feature extraction approaches like the popular MFCC and PLP methods, or to the TRAPS (Jain and Hermansky, 2003) approach of long-term analysis in narrow frequency bands, or it can look for patterns at an oblique angle to the time and frequency axes. In Kleinschmidt's approach the Gabor filters used are chosen by a data-driven selection procedure which searches for Gabor filters that appear likely to give good classification performance.

The second column of Table 3 gives WER on SpeechDatCar-German using the back end of the Aurora baseline speech recognizer (Hirsch and Pearce, 2000) with the QIO-NoTRAPS feature extraction module developed by Qualcomm, ICSI, and OGI (Adami et al., 2002), which calculates robust MFCC features using techniques

such as noise reduction by Wiener filtering and mean subtraction. When features derived from Gabor analysis were concatenated to the robust MFCC features to form a longer feature vector, the error rate decreased, as shown in the third column. We found that for good performance with the Gabor filters it was necessary to pass them through a stage of nonlinear discriminant analysis by MLP (the back end used a different acoustic modeling approach). For source code for Gabor feature extraction, and additional information about this approach, please refer to our website. [4]

Table 3. SpeechDatCar-German word error rates (WER)

WER (%)	QIO-NoTRAPS	With Gabor analysis
Well-matched	5.8	5.4
Medium-matched	12.1	11.7
Highly mismatched	12.0	11.6

4 Robust, Semantically Rich Language Analysis

The approach to analysis (parsing and semantic analysis) of recognized speech normally used in the SMARTKOM system was Ralf Engel's SPINmodule (Engel, 2002). The SMARTKOM project also funded investigation into an alternative approach, aimed at robust and semantically rich language analysis. The alternative approach, described here, makes use of a linguistically sophisticated language formalism called Embodied Construction Grammar (ECG (Bergen and Chang, 2002; Chang et al., 2002). ECG is a construction-based grammar formalism (Goldberg, 1995) that uses embodied primitives like frames (Fillmore, 1982), image schemas (Lakoff, 1987) and executing schemas (Narayanan, 1997) as its semantic representation. In addition to supporting these cognitive primitives, ECG is an extension of unification-based formalisms like HPSG (Pollard and Sag, 1994), and as a consequence, it is precise enough for computational models of language analysis.

ECG is an extremely expressive grammar formalism, and as such new algorithms needed to be designed that could take advantage of the wealth of semantic information contained in an ECG grammar, and yet still efficiently and robustly process each utterance. These algorithms are implemented within the so-called constructional analyzer (Bryant, 2003).

The first key innovation employed in the design of the constructional analyzer is how it combines linguistic knowledge with process. Instead of treating each construction as a passive piece of grammatical knowledge, the analyzer compiles each construction into an active unit called a construction recognizer. Each construction recognizer is responsible for applying both the form and the semantic constraints associated with its construction. The recognizer then generates an instance of the construction if an acceptable set of constituents is found.

[4] http://www.icsi.berkeley.edu/Speech/papers/icslp02-gabor/

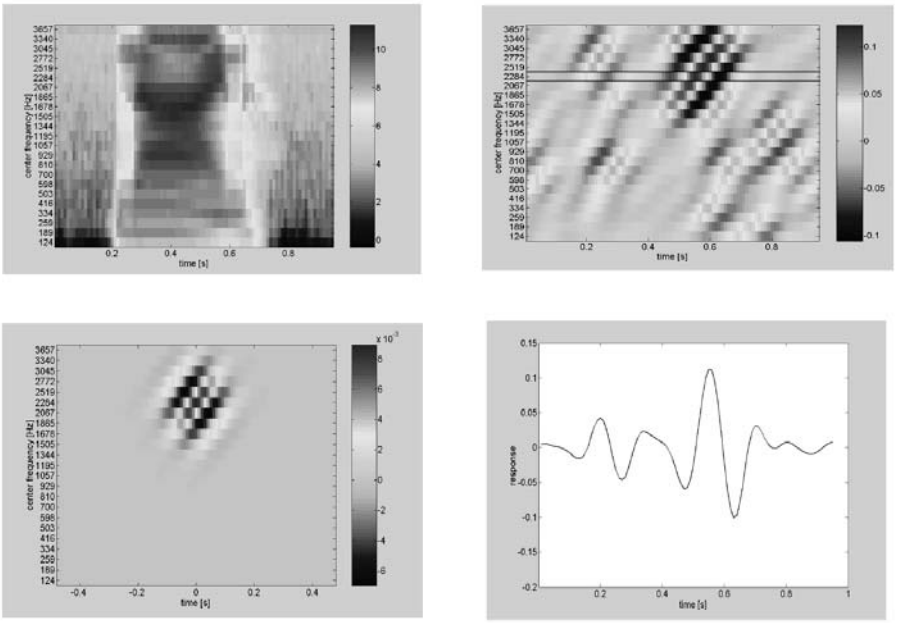

Fig. 4. *Upper left*: A log mel-band spectrogram for the spoken word "nine". *Lower left*: The real part of a Gabor filter. The Gabor filters are complex-valued. Rather than doing acoustic modeling with complex-valued features, Kleinschmidt chose to use the real part, imaginary part, or magnitude of the results of correlation as features. *Upper right*: The real part of the log mel-band spectrogram after correlation with the Gabor filter. *Lower right*: The values of the filtered log mel-band spectrogram at the center frequency of the Gabor filter (it is those that would be used as features for speech recognition)

The analyzer itself must manage the interaction between each of the recognizers, while looking for semantically and formally complete analyses of the whole utterance. The analyzer also has the responsibility for robustly responding when an analysis of the whole utterance cannot be found because of an unforeseen syntactic pattern. This scenario leads us to the second key conceptual innovation employed in the constructional analyzer: leveraging semantics for robust analysis.

Given the rich semantics found in an ECG grammar, the strategy for robust behavior is twofold. First, the analyzer needs a way to infer likely identifications between frames. Such a strategy is specified by the Parsimony Principle (Kay, 1987). It states that the ideal reader unifies compatible frames whenever possible. Applying this principle to a collection of frames and schemas simulates the identifications that were likely to result had a complete analysis been found.

Second, the analyzer needs a mechanism for choosing between the competing analyses that result from application of the Parsimony Principle. The heuristic used for this task is called semantic density (Bryant, 2003). Semantic density defines the completeness of an analysis as the ratio of filled frame roles to total frame roles.

Analyses that are more semantically dense specify more of the frame roles in the utterance than those that are less semantically dense, and the analyzer thus prefers denser analyses.

The constructional analyzer was tested by integrating it with the SMARTKOM domain and context model (Porzel et al., 2006), taking semantic information from that module's ontology. Not only did it successfully analyze questions from the tourist domain, but it also capitalized on the structure found with the ontology to perform linguistic type-coercion known as construal (Porzel and Bryant, 2003).

5 Wizard and Operator Study of Felicitous Human Computer Interaction[5]

5.1 Introduction

End-to-end evaluations of conversational dialogue systems with naive users are currently uncovering severe usability problems that result in low task completion rates. Preliminary analyses suggest that these problems are related to the system's dialogue management and turn-taking behavior. We present the results of experiments designed to take a detailed look at the effects of that behavior. Based on the resulting findings, we spell out a set of criteria which lie orthogonal to dialogue quality, but nevertheless constitute an integral part of a more comprehensive view on dialogue *felicity* as a function of dialogue quality and efficiency.

Research on dialogue systems in the past has focused on engineering the various processing stages involved in dialogical human–computer interaction (HCI), e.g., robust automatic speech recognition, intention recognition, natural language generation, or speech synthesis (Allen et al., 1996; Cox et al., 2000; Bailly et al., 2003). Alongside these efforts, the characteristics of computer-directed language have also been examined as a general phenomenon (Zoeppritz, 1985; Wooffitt et al., 1997; Fraser, 1993; Darves and Oviatt, 2002). The flip side, i.e., computer–human interaction (CHI), has received very little attention as a research question by itself.

The intuitive usability of such conversational dialogue systems can be demonstrated by usability experiments with real users that employ the PROMISE evaluation framework (Beringer et al., 2002), which offers some multimodal extensions over the PARADISE framework (Walker et al., 2000). The work described herein constitutes a starting point for a scientific examination of the "whys" and "wherefores" of the challenging results stemming from such end-to-end evaluations of conversational dialogue systems.

One of the potential reasons for the problems thwarting task completion stems from the problem of *turn overtaking*, which occurs when users rephrase questions or make a second remark to the system while it is still processing the first one. After such occurrences a dialogue becomes asynchronous, meaning that the system responds to the second-last user utterance while in the user's mind that response con-

[5] Robert Porzel and Manja Baudis were the principal authors of this section.

cerns the last. Given the state of the art regarding the dialogue handling capabilities of HCI systems, this inevitably causes dialogues to fail completely.

5.2 Wizard and Operator Study

Here, we describe a new experimental paradigm and the first corresponding experiments tailored toward examining the effects of the computer's communicative behavior on its human partner. More specifically, we will analyze the differences in human–human interaction (HHI) and HCI/CHI turn-taking and dialogue management strategies, which constitutes a promising starting point for an examination of the effects of the computer's communicative behavior on the felicity and intuitiveness of dialogue systems. The overall goal of analyzing these effects is for systems to become usable by exhibiting a more felicitous communicative behavior. After reporting on the results of the experiments in Sect. 5.3, we highlight a set of hypotheses that can be drawn from them and finally point toward future experiments that need to be conducted to verify these hypotheses in Sect. 5.4.

For conducting the experiments we developed a new paradigm for collecting telephone-based dialogue data, called *Wizard and Operator Test* (WOT), which contains elements of both Wizard-of-Oz (WOZ) experiments (Francony et al., 1992) as well as Hidden Operator Tests (HOT (Rapp and Strube, 2002)). This procedure also represents a simplification of classical end-to-end experiments, as it is much like WOZ and HOT experiments conductable without the technically very complex use of a real conversational system. As postexperimental interviews showed, this did not limit the feeling of *authenticity* regarding the simulated conversational system by the human subjects. The WOT setup is described in detail by Porzel and Baudis (2004) and Gurevych and Porzel (2006). It consists of two major phases that begin after the subject has been given a set of tasks to be solved with the telephone-based dialogue system. In the first phase the human assistant is acting as a wizard who is simulating the dialogue system by operating a speech synthesis interface. In the second phase, which starts immediately after a system breakdown has been simulated by means of beeping noises transmitted via the telephone, the human assistant is acting as a *human* operator asking the subject to continue. In our experiments, subjects used the simulated dialogue system to gather information related to tourism in the city of Heidelberg. Simulating a telephone-based dialogue system (rather than a local multimodal dialogue system such as the SMARTKOM-Mobile demonstrator) allowed a natural-seeming switchover from computer–human interaction to human–human interaction.

The experiments were conducted in the English language at ICSI in California. A total of 25 sessions were recorded. At the beginning of the WOT, a person acting as test manager told the subject that they were testing a novel, telephone-based dialogue system that supplies tourist information on the city of Heidelberg. In order to avoid the usual paraphrases of tasks worded too specifically, the manager gave the subjects an overall list of 20 very general tourist activities, such as *visit museum* or *eat out*, from which each subject had to pick 6 tasks that were to be solved in the experiment. The manager then removed the original list, dialed the system's number on

the phone, and exited from the room after handing over the telephone receiver. The subject was always greeted by the system's standard opening ply: *Welcome to the Heidelberger tourist information system. How I can help you?* After three tasks were finished (some successful, some not) the assistant simulated the system's breakdown and came onto the telephone line saying *Excuse me, something seems to have happened with our system, may I assist you from here on*, and finishing the remaining three tasks with the subjects.

5.3 Experimental Results

The PARADISE framework (Walker et al., 1997, 2000) proposes distinct measurements for dialogue quality, dialogue efficiency, and task success metrics. The remaining criterion, i.e., user satisfaction, is based on questionnaires and interviews with subjects and cannot be extracted (sub)automatically from log-files. The measurements described here mainly revolve around dialogue efficency metrics in the sense of Walker et al. (2000). As we will show below, our findings show that a felicitous dialogue is not only a function of dialogue quality, but critically hinges on a minimal threshold of efficiency and overall dialogue management as well. While these criteria lie orthogonal to the Walker et al. (2000) criteria for measuring dialogue quality (such as recognition rates), we regard them to constitute an integral part of an aggregate view on dialogue quality and efficiency, here referred to as *dialogue felicity*. For examining dialogue felicity we will provide detailed analyses of efficiency metrics per se as well as additional metrics for examining the number and effect of pauses, the employment of feedback and turn-taking signals, and the amount of overlaps.

First of all, we apply the classic Walker et al. (2000) metric for measuring dialogue efficiency, by calculating the number of turns over dialogue length on the collected data. (The average length of a dialogue was 6 minutes. The subjects featured approximately uniform distributions of gender, age (12–71), and computer expertise.) As the discrepancy between the dialogue efficiency in phase 1 (HHI) versus phase 2 (HCI) of the experiment might be accounted for by latency times alone, we calculated the same metric with and without pauses. For these analyses, pauses are very conservatively defined as silences during the conversation that exceeded one second.

The overall comparison, given by Porzel and Baudis (2004), shows that naturally latency times severely decrease dialogue efficiency, but also that they alone do not account for the difference in efficiency between human–human and human–computer interaction. This means that even if latency times were to vanish completely, yielding actual real-time performance, we would still observe less efficient dialogues in HCI. While it is obvious that the existing latency times increase the number and length of pauses of the computer interactions as compared to the human operator's interactions, there are no such obvious reasons why the number and length of pauses in the human subjects' interactions should differ in the two phases. However, they do differ substantially.

Next to this *pause effect*, which contributes greatly to dialogue efficiency metrics by increasing dialogue length, we have to take a closer look at the individual turns and their nature. While some turns carry propositional information and constitute utterances proper, a significant number solely consist of specific particles used to exchange signals between the communicative partners, or combinations of such communicative signals with propositional information. We differentiate between dialogue-structuring signals and feedback signals in the sense of Yngve (1970). Dialogue-structuring signals — such as hesitations like *hmm* or *ah* as well as expressions like *well*, *yes*, *so* — mark the intent to begin or end an utterance, or to make corrections or insertions. Feedback signals, while sometimes phonetically alike — such as *right*, *yes* or *hmm* — do not express the intent to take over or give up the speaking role, but serve as a means to stay in contact, which is why they are sometimes referred to as *contact signals*. All dialogues were annotated manually for dialogue structuring and feedback particles.

The data show that feedback particles almost vanish from the human–computer dialogues — a finding that corresponds to those described above. This linguistic behavior, in turn, constitutes an adaptation to the employment of such particles by the respective interlocutor. Striking, however, is that the human subjects still attempted to send dialogue structuring signals to the computer, which — unfortunately — would have been ignored by today's "conversational" dialogue systems. (In the data the subject's employment of dialogue structuring particles in HCI even slightly surpassed that of HHI.)

Most overlaps in human–human conversation occur during turn changes, with the remainder being feedback signals that are uttered during the other interlocutor's turn (Jefferson, 1983). In the collected data the HHI dialogues featured significantly more overlap than the HCI ones, which is partly due to the respective presence and absence of feedback signals as well as to the fact that in HCI turn-taking is accompanied by pauses rather than immediate overlapping handovers.

Lastly, our experiments yielded negative findings concerning the type-token ratio and syntax. This means that there was no statistically significant difference in the linguistic behavior with respect to these factors. We regard this finding to strengthen our conclusions, that emulating human syntactic and semantic behavior does not suffice to guarantee effective and therefore felicitous human–computer interaction.

5.4 Analysis of the Results

The results presented above enable a closer look at dialogue efficiency as one of the key factors influencing overall dialogue felicity. As our experiments show, the difference between the human–human efficiency and that of the human–computer dialogues is not solely due to the computer's response times. There is a significant amount of *white noise*, for example, as users wait after the computer has finished responding. We see these behaviors as a result of a mismanaged dialogue. In many cases users are simple unsure whether the system's turn has ended or not and consequently wait much longer than necessary.

The situation is equally bad at the other end of the turn-taking spectrum, i.e., after a user has handed over the turn to the computer, there is no signal or acknowledgment that the computer has taken the baton and is running with it — regardless of whether the user's utterance is understood or not. Insecurities regarding the main question, i.e., *whose turn is it anyways*, become very notable when users try to establish contact, e.g., by saying *hello* —pause— *hello*. This kind of behavior certainly does not happen in HHI, even when we find long silences.

Examining why silences in human–human interaction are unproblematic, we find that such silences are being announced, e.g., by the human operator employing linguistic signals, such as *just a moment please* or *well, I'll have to have a look in our database* in order to communicate that he is holding on to the turn and finishing his round.

To push the relay analogy even further, we can look at the differences in overlap as another indication of crucial dialogue inefficiency. Since most overlaps occur at the turn boundaries, thereby ensuring a smooth (and fast) handover, their absence constitutes another indication why we are far from having winning systems. As the primary effects of the human-directed language exhibited by today's conversational dialogue systems, the experiments showed that dialogue efficiency decreased significantly even beyond the effects caused by latency times. Additionally, human interlocutors ceased in the production of feedback signals, but still attempted to use his or her turn signals for marking turn boundaries, which, however, go ignored by the system. Last, an increase in pausing is observable, caused by waiting and uncertainty effects, which are also manifested by missing overlaps at turn boundaries.

Generally, we can conclude that a felicitous dialogue needs some amount of extrapropositional exchange between the interlocutors. The complete absence of such dialogue controlling mechanisms by the nonhuman interlocutors alone literally causes the dialogical situation to get out of control, as observable in turn-taking and turn-overtaking phenomena. As evaluations show, this way of behaving does not serve the intended end, i.e., efficient, intuitive, and felicitous human–computer interaction.

Acknowledgments

The English Mobile team at ICSI was Johno Bryant, David Gelbart, Eric Lussier, Bhaskara Marthi, Robert Porzel (visiting from EML), Thilo Pfau, Andreas Stolcke, and Chuck Wooters. Contributions to the development, integration, and testing of the English Mobile system also came from Tilman Becker, Ralf Engel, Gerd Herzog, Norbert Reithinger, Heinz Kirchmann, Markus Loeckelt, Stefan Merten, Christian Pietsch, and Hans-Joerg Kroner at DKFI; Antje Schweizter at IMS; Silke Goronzy, Juergen Schimanowski, and Marion Freese at Sony; Hidir Aras at EML; and Andre Berton at DaimlerChrysler. Funding for ICSI and EML participation in the SMARTKOM project was provided by the German Federal Ministry for Education and Research (BMBF). Some of the work described in this chapter received additional support from other sources, including the Canadian Natural Sciences and Engineering Research Council, the Klaus Tschira Foundation, and Qualcomm.

References

A. Adami, L. Burget, S. Dupont, H. Garudadri, F. Grezl, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan, and S. Sivadas. Qualcomm-ICSI-OGI Features for ASR. In: *Proc. ICSLP-2002*, Denver, CO, 2002.

J.F. Allen, B. Miller, E. Ringger, and T. Sikorski. A Robust System for Natural Spoken Dialogue. In: *Proc. 34th Annual Meeting of the Association for Computational Linguistics*, pp. 62–70, Santa Cruz, CA , June 1996.

C. Avendano. *Temporal Processing of Speech in a Time-Feature Space*. PhD thesis, Oregon Graduate Institute, 1997.

G. Bailly, N. Campbell, and B. Mobius. ISCA Special Session: Hot Topics in Speech Synthesis. In: *Proc. EUROSPEECH-03*, pp. 37–40, Geneva, Switzerland, 2003.

B. Bergen and N. Chang. Embodied Construction Grammar in Simulation Based Language Understanding. Technical Report TR-02-004, ICSI, 2002.

N. Beringer, U. Kartal, K. Louka, F. Schiel, and U. Türk. PROMISE: A Procedure for Multimodal Interactive System Evaluation. In: *Proc. Workshop "Multimodal Resources and Multimodal Systems Evaluation"*, pp. 77–80, Las Palmas, Spain, 2002.

H. Bourlard and N. Morgan. *Connectionist Speech Recognition: A Hybrid Approach*. Kluwer Academic, Dordrecht, The Netherlands, 1993.

J. Bryant. Constructional Analysis. Master's thesis, University of California Berkeley, 2003.

N. Chang, S. Narayanan, and M. Petruck. From Frames to Inference. In: *Proc. Scalable Natural Language Understanding(SCANALU)*, Heidelberg, Germany, 2002.

R.V. Cox, C.A. Kamm, L.R. Rabiner, J. Schroeter, and J.G. Wilpon. Speech and Language Processing for Next-Millenium Communications Services. *Proc. IEEE-2000*, 88(8):1314–1337, 2000.

C. Darves and S. Oviatt. Adaptation of Users' Spoken Dialogue Patterns in a Conversational Interface. In: *Proc. ICSLP-2002*, Denver, CO, 2002.

R. Engel. SPIN: Language Understanding for Spoken Dialogue Systems Using a Production System Approach. In: *Proc. ICSLP-2002*, pp. 2717–2720, Denver, CO, 2002.

C. Fillmore. Frame Semantics. In: Linguistics Society of Korea (ed.), *Linguistics in the Morning Calm*, Seoul, Korea, 1982. Hanshin.

J.M. Francony, E. Kuijpers, and Y. Polity. Towards a Methodology for Wizard of Oz Experiments. In: *Proc. 3rd Conf. on Applied Natural Language Processing - ANLP-92*, Trento, Italy, 1992.

N. Fraser. Sublanguage, Register and Natural Language Interfaces. *Interacting with Computers*, 5, 1993.

D. Gelbart. Mean Subtraction for Automatic Speech Recognition in Reverberation. Technical Report TR-04-003, ICSI, 2004.

A. Goldberg. *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press, Chicago, IL, 1995.

I. Gurevych and R. Porzel. Empirical Studies for Intuitive Interaction, 2006. In this volume.

H. Hermansky. Perceptual Linear Predictive (PLP) Analysis of Speech. *Journal of the Acoustical Society of America*, 87(4), 1990.

H.G. Hirsch and D. Pearce. The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems Under Noisy Conditions. In: *Proc. ISCA ITRW ASR2000*, Paris, France, 2000.

P. Jain and H. Hermansky. Beyond a Single Critical-Band in TRAP Based ASR. In: *Proc. EUROSPEECH-03*, Geneva, Switzerland, 2003.

G. Jefferson. Two Explorations of the Organisation of Overlapping Talk in Conversation. *Tilburg Papers in Language and Literature*, 28, 1983.

D. Jurafsky, C. Wooters, G. Tajchman, J. Segal, A. Stolcke, E. Fosler, and N. Morgan. The Berkeley Restaurant Project. In: *Proc. ICSLP-94*, Yokohama, Japan, 1994.

P. Kay. Three Porperties of the Ideal Reader. In: R.O. Freedle and R.P. Durán (eds.), *Cognitive and Linguistic Analyses of Test Performance*, pp. 208–224, Norwood, NJ, 1987. Ablex.

M. Kleinschmidt. *Robust Speech Recognition Based on Spectro-Temporal Processing*. PhD thesis, Carl von Ossietzky-Universität, Oldenburg, Germany, 2002.

M. Kleinschmidt and D. Gelbart. Improving Word Accuracy With Gabor Feature Extraction. In: *Proc. ICSLP-2002*, Denver, CO, 2002.

G. Lakoff. *Women, Fire, and Dangerous Things*. University of Chicago Press, Chicago, IL, 1987.

C. Mokbel, D. Jouvet, and J. Monné. Deconvolution of Telephone Line Effects for Speech Recognition. *Speech Communication*, 19(3), 1996.

A. Moreno, B. Lindberg, C. Draxler, G. Richard, K. Choukri, S. Euler, and J. Allen. SpeechDat-Car: A Large Speech Database for Automotive Environments. In: *Proc. 2nd Int. Conf. on Language Resources and Evaluation (LREC 2000)*, Athens, Greece, 2000.

S. Narayanan. *Knowledge-Based Action Representations for Metaphor and Aspect*. PhD thesis, University of California, Berkeley, CA, 1997.

L. Neumeyer, V. Digalakis, and M. Weintraub. Training Issues and Channel Equalization Techniques for the Construction of Telephone Acoustic Models Using a High-Quality Speech Corpus. *IEEE Transactions on Speech and Audio Processing*, 2(4), 1994.

C. Pollard and I. Sag. *Head-Driven Phrase Structure Grammar*. University of Chicago Press and CSLI Publications, Chicago, IL, 1994.

R. Porzel and M. Baudis. The Tao of CHI: Towards Felicitous Human-Computer Interaction. In: *Proc. Human Language Technology Conference / North American Chapter of the Association for Computational Linguistics Annual Meeting 2004*, Boston, MA, 2004.

R. Porzel and J. Bryant. Employing the Embodied Construction Grammar Formalism for Knowledge Represenation: The Case of Construal Resolution. In: *Proc. 8th Int. Cognitive Linguistics Conference*, Logrono, Spain, 2003.

R. Porzel, I. Gurevych, and R. Malaka. In Context: Integrating Domain- and Situation-Specific Knowledge, 2006. In this volume.

S. Rapp and M. Strube. An Iterative Data Collection Approach for Multimodal Dialogue Systems. In: *Proc. 3rd Int. Conf. on Language Resources and Evaluation (LREC 2002)*, pp. 661–665, Las Palmas, Spain, 2002.

S. Renals and M. Hochberg. Efficient Evaluation of the LVCSR Search Space Using the NOWAY Decoder. In: *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP-96)*, Atlanta, GA, 1996.

A. Stolcke. SRILM — An Extensible Language Modeling Toolkit. In: *Proc. ICSLP-2002*, Denver, CO, 2002.

M.A. Walker, C.A. Kamm, and D.J. Litman. Towards Developing General Model of Usability with PARADISE. *Natural Language Engineering*, 6, 2000.

M.A. Walker, D.J. Litman, C.A. Kamm, and A. Abella. PARADISE: A Framework for Evaluating Spoken Dialogue Agents. In: *Proc. 35th ACL*, Madrid, Spain, 1997.

R. Wooffitt, N. Gilbert, N. Fraser, and S. McGlashan. *Humans, Computers and Wizards: Conversation Analysis and Human (Simulated) Computer Interaction*. Brunner-Routledge, London, UK, 1997.

V. Yngve. On Getting a Word in Edgewise. In: *Papers From the 6th Regional Meeting of the Chicago Linguistic Society*, Chicago, IL, 1970.

M. Zoeppritz. Computer Talk? Technical Report 85.05, IBM Scientific Center Heidelberg, 1985.