STARTPAGE

PEOPLE
MARIE CURIE ACTIONS

Marie Curie Initial Training Networks (ITN)
Call: FP7-PEOPLE-2007-1-1-ITN

PART B

STAGE 2 — FULL PROPOSAL

SCALE

Coordinator contact information:
Dietrich Klakow
Spoken Language Systems
Saarland University
Email: dietrich.klakow@lsv.uni-saarland.de
Tel: +49–681–302–58–122

Closure Date: September 25, 2007
Date of Preparation: September 25, 2007

Marie Curie Initial Training Networks (ITN)

SCALE : Speech Communication with Adaptive LEarning

Coordinator: Dietrich Klakow
dietrich.klakow@lsv.uni-saarland.de
Date of Preparation: September 25, 2007

Contents

1 List of Participants **4**

2 Project Overview **5**

3 S&T Quality **6**
3.1 Introduction . . . 6
3.1.1 Speech Technology: the Promise . . . 6
3.1.2 Speech Technology: the Problems . . . 6
3.2 Scientific Objectives . . . 7
3.3 Proposed Research . . . 9
3.3.1 Research Theme 1: Bridging the Gap between Recognition and Synthesis (RS) . . . 9
3.3.2 Research Theme 2: Bridging the Gap between ASR and HSR (AHSR) . . . 10
3.3.3 Research Theme 3: Bridging the Gap between Signal Processing and Learning (SPL) . . . 11
3.4 Original, Innovative, and Interdisciplinary Aspects of SCALE . . . 12
3.5 Research Collaboration and Exploitation of Synergies . . . 13

4 Training **16**
4.1 Importance and Timeliness of the Training . . . 16
4.2 Overview and Structure of the Training Activities . . . 17
4.3 Local Training . . . 17
4.3.1 Individual Research Projects . . . 17
4.3.2 Bridging the Gap between Recognition and Synthesis (RS) . . . 18
4.3.3 Bridging the Gap between ASR and HSR (AHSR) . . . 21
4.3.4 Bridging the Gap between signal processing and learning (SPL) . . . 22
4.3.5 Local Technical and Complementary Skills Training . . . 23
4.4 Secondments and Industrial Experience . . . 24
4.5 Network Wide Training . . . 24
4.5.1 Summer Schools . . . 24
4.5.2 Workshops . . . 26
4.6 SCALE Conference . . . 27
4.7 General Aspects of the SCALE Training . . . 27

4.7.1 Supervision and Mentoring . . . 27
4.7.2 Personal Career Development Plans (CDPs) . . . 28
4.7.3 Early Stage vs. Experienced Researchers . . . 28
4.7.4 Networking and Exchange of Best Practice . . . 29
4.8 Size and Balance of the Training Program . . . 29

5 Implementation 30
5.1 Capacities of Each Partner . . . 30
5.2 Overall Workplan . . . 36
5.3 The Role of Industry and External Partners . . . 36
5.4 Complementarities and Exploitation of Synergies . . . 37
5.5 Overall Management of the Research Training . . . 38
5.5.1 Overall Scientific Co-ordinator (CO) . . . 39
5.5.2 Project Management Office (PMO) . . . 39
5.5.3 Supervisory Board (SB) . . . 39
5.5.4 Communication and Networking . . . 40
5.5.5 Decision-making and Conflict Resolution . . . 40
5.5.6 Monitoring, Reporting & Risk Situations . . . 41
5.5.7 Management of Knowledge and IPR . . . 41
5.5.8 Dissemination and Exploitation . . . 42
5.5.9 Financial Management . . . 43
5.6 Recruitment Strategy . . . 44

6 Impact 45
6.1 Benefits for Participating ESRs and ERs . . . 45
6.2 Benefits for Participating Institutions . . . 46
6.3 Further Benefits to the Community and the European Research Area . . . 48

7 Ethical Aspects 49

A Letter of Intent: Motorola 51

B Letter of Intent: Philips 52

C Letter of Intent: Eurice 53

1 List of Participants

Network participants

UdS University of Saarland, DE

Departments:
(1) Spoken Language Systems (http://www.lsv.uni-saarland.de/) (Overall Scientific Co-ordinator)
(2) European Project Office EPO (Administrative and Financial Coordinator)

Persons-in-charge:
(1) Prof Dietrich Klakow (Overall Scientific Coordination)
(2) Ms. Corinna Hahn (Administrative and Financial Coordination)

IDIAP Fondation Dalle Molle, CH

Department: IDIAP Research Institute, (http://www.idiap.ch)

Person-in-charge: Prof Hervé Bourlard

RUN Radboud University Nijmegen, NL

Department: Department of Language and Speech (http://lands.let.ru.nl/)

Person-in-charge: Prof Lou Boves

RWTH RWTH Aachen, DE

Department: Lehrstuhl für Informatik 6 (Human Language Technology and Pattern Recognition, http://www-i6.informatik.rwth-aachen.de/i6.html)

Person-in-charge: Prof Hermann Ney

UEDIN University of Edinburgh, UK

Department: Centre for Speech Technology Research (http://www.cstr.ed.ac.uk)

Person-in-charge: Prof Steve Renals

USFD University of Sheffield, UK

Department: Department of Computer Science (http://www.shef.ac.uk/dcs)

Person-in-charge: Prof Roger K. Moore

Associate partners

Motorola Motorola Ltd, UK

Department: Motorola European Research Laboratory (http://www.motorola.com)

Person-in-charge: Dr David Pearce

Philips Philips Speech Recognition Systems, A

Department: Technology Development (http://www.philips.com)

Person-in-charge: Dr Heinrich Bartosik

EURICE European Research & Project Office GmbH, DE

Department: Training (http://www.eurice.de)

Person-in-charge: Jörg Scherer

2 Project Overview

In order to realize the advantages of speech-based multimodal interaction, a new elite of researchers are required who are well-versed in the cross-disciplinary aspects of speech processing and spoken language communication, in contrast with the highly specialized researchers who are currently the norm. Educating such an elite class of researchers is the primary goal of the *Speech Communication with Adaptive LEarning*, or SCALE, Initial Training Network. We intend to achieve that goal by bringing together a number of *Early Stage Researchers* (ESRs) in a research and training programme that addresses the most pressing fundamental problems of the current technology in an application oriented environment.

Since it is now generally agreed that future speech technology can only be made sufficiently robust by making it more flexible and adaptive, SCALE will be concerned with adaptive learning approaches to all areas of speech processing, with particular focuses on ways in which machine learning techniques can be exploited for making signal processing, automatic speech recognition and synthesis adaptive. To determine the most promising direction into which adaptive processing should develop, the research in SCALE will be informed by existing and newly acquired knowledge about human speech processing. This holds even for microphone arrays, which bring obvious super-human capabilities. Yet, we are convinced that automatic speech processing can only reap the benefits of the advantages of multi-aural signal processing over binaural processing if we are able to use microphone arrays for focusing on the speech features in acoustic signals originating from multiple simultaneous sources.

The training to be delivered to the ESRs falls into four main categories. Firstly, each researcher will have a home institution, and a supervisor there who oversees daily activities. The assignment of researchers to home institutions will be made primarily on the basis of research interest. Secondly, each ESR will engage in research in one of the three interdisciplinary themes described in Section 3.3. Thirdly, each ESR will spend up to six months in secondments at one or more guest institutions. The selection of the guest institutions will be based on the personal education and development plan established for each ESR. Wherever possible, the guest institution will be one of the industrial partners. In this way we will ensure that the secondments enhance the career perspectives of the ESRs. The personal career development plans, and therefore also the secondments, will emphasize the need for cross-disciplinary and cross-sectoral (academia and industry) training and education. Fourthly, the ESRs will participate in workshops and summer schools organized by SCALE partners. In addition to technical training, the ESRs will also receive instruction in complementary skills: During the summer schools, network-wide specialized complementary skills training will be provided on intellectual property management, proposal writing and funding opportunities, project management, negotiation skills and research ethics. In addition, general complementary skills training courses (i.e. on entrepreneurship, gender and migration, presentation skills, language courses) will be available on a local level at each host institution.

The network will also hire two Experienced Researchers (ERs), each for 24 months. The goal of ER training will be to prepare them to positions as assistant or junior professors, lecturers, or comparable positions in academia or industry. Thus, while having the access to all training activities within SCALE that ESRs have, the focus of the ER training will be on performing independent research and learning the skills involved in supervising students and managing projects. As with the ESRs, a personal career development plan (CDP) will be developed for each ER, and each ER will have an assigned research area.

Finally, all fellows will participate in the organization of a SCALE conference in the first half of the final year of the project. In these ways, SCALE researchers will profit from the combined expertise of the consortium (comprising higher education institutions and two major IT companies) and receive a range of training and experience that is available neither from any single consortium partner nor in any single country.

3 S&T Quality

3.1 Introduction

3.1.1 Speech Technology: the Promise

All aspects of speech processing are rapidly growing in commercial and industrial importance. A leading market analyst reports that 129.8 million USD were generated by speech products in North America in 2004 from licensing and maintenance services, but predicts these revenues will increase to nearly one billion USD by 2011. Another analyst predicts revenues from speech-related applications in Germany will rise from 90 million Euro in 2004 to almost 200 million Euro by 2008. At the moment, speech-driven systems are deployed in customer-facing business areas such as online information for cinemas and railways, customer assistance for telecommunication services and telephone banking. These services are all based on the telephone with speech recognition, dialogue management and speech synthesis performed in a central network server.

A second type of application for which acceptable levels of performance have been reached is dictation on PC using a close-talking microphone. A closely related application is transcription services for professionals, for example to assist clinicians in processing patient records and consultant reports. The market for this in North America is estimated to be 20 billion USD. To date, only a small portion of this market has been captured by firms using ASR technology, but the market share of ASR-supported software is rapidly growing. Use of speech transcription is also growing for accessing audio and video recordings (and films) in media archives. Another application of speech transcription, as well as speaker identification, is in filtering telephone traffic in order to combat terrorism and organized crime. The social and economic value of applications of speech technology is difficult to quantify in Euros, but it is also difficult to overestimate.

A third application area for speech technology is in mobile devices that do not provide a fully-fledged keyboard for entering information. Today, the best known applications are name dialling in mobile telephones and entering departure and destination locations in navigation systems. Recently, interesting applications have emerged for accessing music on small devices such as an MP3 player, an iPod or an iPhone. In the near future we will certainly see services that use speech interfaces to offer mobile access to multimedia programming via the Internet. The link with audio/video indexing is obvious. Yet another application that relies on hand-held devices is exemplified by speech-to-speech translation, for example in contacts between medical personnel and patients in countries where few persons speak a western language. Last but not least, the vision of an ambient intelligence environment and ubiquitous computing seems impossible to realize without fully transparent multimodal interfaces in which speech plays a crucial role. Again, the social and economic return of ambient intelligence environments which will enable elderly people to prolong independent living is difficult to overestimate. For this to happen we must be able to create a keyboard-free society, in which human users will interact with and through ubiquitous computers in a completely unobtrusive manner by using speech in combination with gestures in a context-aware intelligent environment.

3.1.2 Speech Technology: the Problems

Much of the potential market impact for voice interfaces remains untapped because the performance of the core speech technologies is inadequate and does not provide for simple and unobtrusive verbal interaction with a device, let alone for interaction that is as natural as that with another human being. Application developers and academic researchers alike have identified several key problems that stand in the way of pervasive applications of speech technology. Virtually all the problems are in one way or another related to the lack of flexibility and adaptability of the technology. In its turn, this is directly related to the methods and algorithms on which all existing recognition technology is based: we rely on statistical pattern processing techniques, predominantly Hidden Markov Models (HMMs) which require training on large bodies of data, perhaps thousands of hours

of speech. This approach produces excellent performance when the speech to be recognised or synthesised is sufficiently similar to the training data, but performance deteriorates rapidly as the statistical distance between training examples and new input or output becomes larger. This drop in machine performance is much faster than in human performance on the same tasks, most probably because listeners and speakers are able to adapt to previously unobserved inputs and environmental conditions, arguably because they have a fundamentally different representation of speech and language that does not rely exclusively on the law of large numbers. Thus, there is an urgent need to research methods and algorithms that have adaptation and learning as inherent properties, and that can learn from relatively small amounts of data. Developing such novel methods and algorithms, however, requires the education of a new generation of researchers who are aware of techniques and algorithms under development in more than one of the disciplines that are active in the large and diverse field of spoken language communication.

3.2 Scientific Objectives

In recent years considerable technological progress has been made in all sub-fields of speech processing, and these advances have contributed to the rapidly increasing market acceptance of speech-driven products. While the increasing market penetration of speech products demonstrates the commercial viability of this technology, it is clear that if *automatic speech recognition* (ASR) is ever to expand beyond its primarily telephone-based and office dictation markets to become the truly ubiquitous human-machine interface, a challenge of fundamental importance will be to achieve the same performance with far-field sensors that is currently attainable only with close-talking microphones. In addition, it is necessary to move from carefully pronounced dictation in a specific domain to a wide range of conversational speech styles and almost any conceivable topic or domain. So called "found speech" produced for human listeners is much more challenging than speech generated in order to be recognised by a machine. Such breakthroughs would enable ASR to be used in applications such as the command and control of humanoid robots, for the automatic transcription of radio and television broadcasts, podcasts, meetings and conferences without head-mounted microphones, for interaction with automobile navigation systems and for hand-held speech-to-speech translation devices.

Mobile phones are growing in their computational resources and capabilities as well as providing a mobile interface to the internet. Speech input and output as part of a multimodal interface can greatly enhance the ease of use of such devices, provided that understanding the device's output and being understood by the device is as easy and transparent as in human-human communication. As users want to look at the device screen, however, the technology needs to improve to cope with the noise introduced due to the speaker's greater distance from the microphone. In some applications this problem may be tackled by using lightweight earpiece-microphone accessories with a Bluetooth connection. However, such a solution would be unacceptable in the home, for interfacing to future generations of entertainment and the web. For this purpose speech-based multimodal interaction would be a great asset, but users want to use the system hands-free and without wearing a headset. The rapidly growing elderly population who want to keep living independently in their familiar environment will use multimodal speech driven interfaces for communicating with appliances and services in smart homes, as well as with remote caregivers. It is well known that the speech production and perception of this population differs substantially from what is observed in the conventional training corpora, and that it (rapidly) changes over time, necessitating yet another level of adaptation.

Speech *synthesis* systems have reached levels of high intelligibility, yet exhibit a rather limited range of speaking styles and a general lack of expressiveness. In particular, unlike a human talker, no contemporary synthesis system makes intelligent adjustments to its spoken output (e.g. speaking louder or articulating more clearly) in response to the difficulties that may be imposed on its listener by the prevailing acoustic environment or communicative situation. In addition to a lack of expressiveness, current speech synthesis systems have a limited set of voices, which typically require substantial speaker-specific training data. For applications such as providing voices for characters or avatars in virtual environments it would be of great interest to create a new

voice using a minimum of training data.

Over the past decades considerable advances have been made in the area of computational models of *human speech processing* (HSP), which are able to reproduce key effects observed in psycholinguistic experiments. However, few if any of the psycholinguistic studies, address conversational speech in a range of domains and styles. Also, most probably existing psycholinguistic models do not accurately emulate the neural processing underlying human speech production and perception. Still, it is widely believed that fundamental aspects of computational models of human speech processing will elucidate essential differences between human processing and the processing currently implemented in speech technology. Therefore, we believe that further development of computational models of human speech processing will be able to inform the development of novel methods and algorithms for speech-driven human-machine interaction that are necessary to diminish the gap between human and machine performance levels. Specifically, enhanced modelling of human production and perception of casual, conversational speech will provide guidance for improving those aspects of speech technology that are most important for the applications discussed above. This will require a particular focus on research into the ways in which humans adapt to speakers and dialects they have not previously met, to dynamically changing acoustic environments and to changing speech styles.

So, despite the substantial improvements which have been made in each of the technologies mentioned above, most leading researchers in the field are convinced that the performance breakthroughs needed for enabling a much wider range of speech technology applications can only be obtained by innovative methods and algorithms rather than by incremental development of existing techniques. Moreover, it is generally believed that this breakthrough technology is crucially dependent on our ability to exploit the synergies between what were previously viewed as unrelated research areas. For example:

- Using notions from *human speech recognition* (HSR) to inform developments in ASR;
- Using the machine learning approaches that dominate ASR to improve speech synthesis;
- Using the notion of propagation of uncertainty that underlies ASR to better inform models of HSR, and speech synthesis;
- Replacing the relatively crude models of the speech signal used in ASR with more realistic models informed by speech production;
- Using machine learning techniques to develop adaptive speech signal processing.

The primary objective of SCALE is to create a environment that will foster synergistic cooperation between researchers in each of the aforementioned disciplines. SCALE has three principal scientific objectives:

1. To bridge the gap between speech recognition and speech synthesis;
2. To bridge the gap between human and automatic speech recognition;
3. To bridge the gap between signal processing and adaptive learning.

Bridging those gaps will require breakthroughs in theories, models and algorithms. By aiming at such ground breaking change, we intend to bring the field forward and, at the same time, educate a new generation of researchers with the knowledge, capabilities and experience that will enable them to continue this development.

To make the fundamental research objectives sketched above measurable, we plan to assess the improvements in the core technologies through regular evaluations on standard benchmark tasks. Improvements in automatic speech recognition technology will be gaged through the participation by consortium partners in the annual Rich Transcription / Meeting benchmark held by the US National Institute of Standards and Technologies

(NIST), as well as through specialised evaluations such as the PASCAL Speech Separation Challenge. Improvements in speech synthesis will be measured through participation in the Blizzard Challenge. We also plan to collect one additional speech corpus oriented around far-field sensors on mobile, hand-held devices such as PDAs or cell phones. The presence of two major industrial partners will ensure the research proposed here has a direct impact on the next generation of speech-driven products. Thus, in terms of measurable and industrially relevant performance, our technical objectives are:

1. To improve the performance of automatic speech recognition measured on a selected state-of-the-art ASR tasks. In particular focusing on (i) reducing the gap in performance between the use of close-talking microphones and that with far-field sensors, (ii) improving recognition in non-stationary background noise such as overlapping speech, (iii) achieving greater robustness for the speech of talkers from a wide demographic cross-section.

2. To improve the performance of text-to-speech synthesis. In particular focusing on the intelligibility and acceptability of speech synthesis as measured in listening tests where the listener is in a natural acoustic environments with background noise and reverberation.

3.3 Proposed Research

In this section, we discuss the primary science and technology themes of the SCALE research plan.

3.3.1 Research Theme 1: Bridging the Gap between Recognition and Synthesis (RS)

An increasing number of current text-to-speech techniques are borrowed from speech recognition, particularly in areas such as the automatic segmentation of corpora for speech synthesis, and the development of the trajectory HMMs for speech synthesis. Indeed, the trajectory HMM is a good example of the research we have in mind, since the standard HMM as used in ASR was modified specifically for speech synthesis, and the modified form is now being reapplied to ASR. At the same time there is an increasing interest in links between basic concepts underlying unit selection synthesis and the way in which speech and language are represented and processed in the human brain. There is mounting evidence that we do not produce speech by arranging neural commands for individual movement of individual articulators in the right order, and that we do not parse acoustic input into phonemes, only to assemble them into meaningful words and phrases. Rather, there is evidence that speech production and perception involves the activation of longer stretches of speech that are represented in the brain on a hierarchy of levels of abstraction. Thus, both in speech synthesis and recognition research there is an interest in efficient storage and retrieval of units with the size of words and phrases.

In SCALE we aim to further develop these commonalities to improve both recognition and synthesis, and to provide richer, more sophisticated models of the speech signal, without sacrificing trainability and the ability to estimate the model parameters from (potentially very small amounts of) data. We focus on trajectory models, reactive speech synthesis, and more sophisticated signal models. In order to be able to test novel methods and algorithms and to make these available for medium-term commercial applications we will use state-of-the-art systems for both recognition and synthesis that the SCALE partners have built in-house and that we therefore can adapt to the results of our R&D efforts.

As described in Section 4.3.2, the individual research projects falling within the first research theme are:

1. RS-1: Trajectory HMMs for Reactive Speech Synthesis;
2. RS-2: Towards Speaker Invariance: Compensating for Coarticulation;
3. RS-3: Hierarchical Trajectory Models for Speech Recognition;
4. RS-4: Speech Synthesis by Analysis.

3.3.2 Research Theme 2: Bridging the Gap between ASR and HSR (AHSR)

Current ASR systems have been shown to be somewhat robust to various distortions of the input signal relative to the training data, but even the most powerful systems fall short dramatically compared with human performance. It is generally agreed that the gap between human and machine performance can only be bridged by introducing those aspects of human processing which make the difference. We believe that human speech processing is a process on several interacting hierarchical levels [11]. In this research theme we aim to improve the performance of automatic speech processing by developing and testing a hierarchical architecture. At the same time, we intend to advance our understanding of human speech processing by analysing the improvements in automatic processing obtained by introducing concepts from human processing.

Both human and automatic speech recognition deal with the transformation of a speech signal into a sequence of lexical tokens (words), and use concepts such as lexicon, search, word representations, context dependency and word activation [12, 14]. Computational models in both ASR and HSR must support adaptation to previously unheard patterns at the level of articulatory features, phonemes, words and phrases. It is well known that adults find it very difficult to learn distinctions between articulatory features [13] and phonemes that do not occur in their native language(s), while they have little difficulty in learning and processing new words and phrases.

While ASR for languages such as English, which has relatively simple morphological structure, can be approached by assuming that there is a fixed lexicon that contains all possible words, such an approach is doomed to fail with languages such as Dutch, German, Turkish and Arabic, which have a much more complex morphology. For languages with a complex morphology it is necessary to add an extra level to the decoder, where subword units will be combined to words in the lexicon or to new words that can be created from the lexicon and the morphological rules of the language. Obviously, this additional component of the decoder needs to integrate hypotheses about subword units resulting from bottom-up signal processing and detection of out-of-vocabulary words which relies very much on top-down verification [5, 2].

The efficiency and robustness of speech communications is to a large extent due to the fact that speech signals are highly redundant. All speech sounds are characterised by a large number of articulatory and acoustic features. Consequently, humans can recognise the sounds even if part of the features are distorted due to coarticulation or background noise. However, little is known about the ways in which humans detect and process the relevant features, especially in adverse acoustic conditions. Yet, it is reasonable to assume that the processes which allow us to separate speech from background signals are similar to the processes investigated in computational acoustic scene analysis [4, 1]. Therefore, we will develop rich and redundant acoustic representations of speech signals, including such features as harmonicity, onset time and location in conjunction with adaptive machine learning techniques to discover which features are relevant under specific environmental conditions. It goes without saying that the back-end decoder must be adapted to be able to process the novel features.

Although adults may not be able to learn to speak new languages without some remaining foreign accent, they are able to acquire representations of the acoustic and articulatory features of a new language at level that is sufficient for effective communication [17]. We shall investigate new approaches towards representing the acoustic signal via context- and language-invariant sub-word representations. In bridging the gap between human and automatic processing we will focus on the computational processes that both humans and machines must perform [9]. These representations will be derived from an acoustic stream without using a language model or phonotactic and grammatical rules of a particular language, but may employ universal (language-independent) constraints on speech production. The goal will be achieved through defining and extracting a new powerful physiologically and psychophysically motivated set of language- and task-independent acoustic descriptors (features), reflecting posterior probabilities of speech-specific (and language- and talker-independent) sub-word classes and features [15].

When trying to discover 'natural' units for representing speech, we expect to find a complex hierarchically

organised set of units, some with the size of full syllables, others with the size of sub-phonemic phenomena. Words and phrases can be composed of these basic units in many different manners, depending on speaker, speaking style, prosody, etc. While the highest levels of the hierarchical representations encode the invariant linguistic features of words and phrases, the lower levels in the hierarchy encode speaker, style and context effects. To be able to process such a complex hierarchical representation efficiently we will need some kind of associative memory, in which partial representations of the input signals on the lower levels activate representations of larger linguistic units, which can then be verified with limited computational power [7, 6]. Interestingly, a representation that models speech in the form of multi-level units will evidently affect both speech synthesis and recognition.

To reach the targets sketched in the previous paragraphs we will conduct the following research projects:

1. AHSR-1: Towards Open Vocabulary Speech Recognition;
2. AHSR-2: Data Association Multisource Acoustic Models;
3. AHSR-3: Sounds and Spoken Language;
4. AHSR-4: Associative Memories for Learning and Decoding Speech.

3.3.3 Research Theme 3: Bridging the Gap between Signal Processing and Learning (SPL)

The acoustic models of current ASR systems are based on trainable statistical models, usually HMMs. These model parameters are estimated using automatic machine learning techniques, such as the EM algorithm [8], whereas the signal processing component of most systems is essentially non-adaptive. We propose to investigate adaptive learning approaches to signal processing and its interface to statistical acoustic models. By dividing the primary signal processing level into multiple, possibly heterogeneous parts, through parallel recording channels, different feature extraction methods, or specific signal oriented modelling, learning could be introduced to signal processing. Learning then would be concerned with separating sources of variability (e.g. localization using multiple microphone systems), or with integration of signal properties from different features (feature and/or model combination), or by providing models with the potential to capture information for which HMMs are not well suited (conditional random fields, multistream approaches).

Multiple microphone sensors provide a way to localize and enhance acoustic sources. We are particularly interested in algorithms for improved acoustic signal capture from uncalibrated arrays of microphones to deliver improved automatic recognition. This goal requires minimizing the detrimental effects of reverberation and background noise. Most current approaches to this challenge, such as superdirective beamforming [3], do not frame the problem in an adaptive learning setting. We plan to investigate two basic approaches to adaptive signal processing in this context. Firstly, it is well-known that human speech, viewed in either the time or frequency domain, is a *non-Gaussian* signal. Conventional beamforming, however, treats all signals, both sources and interferences, as Gaussian [16]. We plan to develop beamforming techniques specifically tailored to the automatic recognition of far-field speech data that explicitly exploit this non-Gaussian nature to distinguish desired speech sources from noise and other interference, such as reverberation. Secondly, most current approaches to microphone array processing assume some knowledge of the array geometry; we propose to investigate approaches to microphone array processing based on ad-hoc collections of sensors with unknown positions, both relative and absolute.

With current technology, ASR systems using far-field sensors typically have an error rate 2-3 times as high as that provided by the close-talking microphone. The final goal of our research on speech processing with multiple microphones will be to bridge the gap in performance attainable with such far-field sensors and with close-talking microphones.

ASR systems usually have different strengths and deficiencies leading to complementary results, meaning that systems usually do not produce the same errors, or errors at the same positions. Therefore, system combina-

tion in principle can take advantage of the individual strengths of different systems. It has been shown that combinations of different systems with methods like ROVER, Confusion Network Combination (CNC), Discriminative Model Combination (DMC), or cross-system speaker adaptation consistently leads to improved performance. In SCALE, we propose a systematic study of a variety of approaches to system combination. The aim is to find further approaches with improved automatic selection capabilities to reduce the empirical optimization effort during system combination.

The individual research projects falling within the first research theme are:

1. SPL-1: Non-Gaussian Beamforming for Far-Field ASR;
2. SPL-2: Particle Filters for Robust Far-Field ASR;
3. SPL-3: Multi-Channel Modelling for Automatic Speech Recognition;
4. SPL-4: Investigations on Feature and System Combination Methods;
5. SPL-ER-1: Multiple Microphone Techniques for Dereverberation;
6. SPL-ER-2: Multi-Microphone Beamforming and Noise Reduction Using Auditory Processing.

3.4 Original, Innovative, and Interdisciplinary Aspects of SCALE

Spoken language has been described as "the most sophisticated behaviour of the most complex organism in the known universe" [11]. Scientific understanding of the speech communication process has long been of interest in several disciplines, principally linguistics, phonetics, acoustics and audition. The engineering goal of replicating aspects of speech communication (principally coding, recognition and synthesis) has been of interest to mathematicians, electrical engineers and computer scientists for more than 50 years. Each of these subjects, however, has a different culture and trains its disciples in different ways, making effective cross-subject communication difficult. It is only in recent years that it has become possible to exploit the multi-faceted nature of speech research to make genuine progress. The advances which we have described above, linking Human with Machine Recognition, Recognition with Synthesis and Signal Processing with Learning, are what now makes multidisciplinary work fruitful. At the same time, they point to a need for more scientists specifically trained to work across traditional boundaries, which is exactly the object of SCALE.

As particular examples of multidisciplinary innovation we highlight:

- Use of array processing techniques to improve the quality of a speech signal captured with far-field sensors.
- The use of concepts from auditory scene analysis to separate speech from simultaneous acoustic sources if microphone arrays cannot be used.
- The development of speech synthesis systems that use techniques derived from speech recognition in order to optimise the way they speak in different environments.
- The development of speech synthesis systems based on trajectory HMMs.
- The use of multistream statistical models for speech recognition and synthesis.
- Improvement of automatic speech recognition by introducing concepts of human speech processing.

In order to be successful in the future, scientists must bring together specific knowledge and understanding which is sufficiently broad and comprehensive, but at the same time sufficiently profound in all areas relevant to human language technologies. This includes, but is not limited to, an understanding of how sound propagates,

particularly in bounded acoustic environments such as rooms, how audio data may be effectively captured with far-field sensors, how digital signal processing may be used for speaker tracking and beamforming, how speech may be modelled for accurate recognition, and how the best word hypothesis or hypotheses may be efficiently extracted from a waveform based on one or more acoustic models. Presently, each of these topics is viewed as a largely independent discipline and researchers in, for example, array processing have only the vaguest of notions as to how word lattices are generated in the back end of an ASR engine. Similarly, many mainstream ASR researchers know as good as nothing about array processing techniques, and have little interest in the speech signal itself before it has been transformed to cepstral coefficients. It is this undesirable state of affairs that the SCALE project will change.

The combination of array processing techniques with automatic speech recognition is a newly emerging field, and one that holds great promise to make ASR the human-machine interface of first choice. One of the significant unsolved problems with conventional ASR is that it is largely dependent on the use of close-talking microphones (CTMs) for the capture of speech signals. Many studies indicate that the necessity of donning a CTM is viewed by users as a major annoyance, and drastically hinders user acceptance of ASR in many applications. As soon as a single CTM is placed at even a distance of one meter from the user, however, performance of conventional ASR systems degrades radically. This degradation is due to capture of environmental noise, reverberation, and the voices of other speakers along with the desired speech of the user. Recent research has revealed that the capture of far-field speech data with multiple microphones arranged with a known geometry (i.e., a *microphone array*) and the application of array processing techniques to combine the outputs of the several microphones can lead to large reductions in word error rate [10]. The SCALE project is well-situated to play a leading role in the development of array processing techniques for ASR, inasmuch as the SCALE consortium brings together experts in the respective research areas to provide the multidisciplinary environment which is indispensable for realizing the advantages of speech-based multimodal interaction and for educating the new generation of researchers required in both academia and industry.

Speech synthesis has been transformed in recent years through the use of data-driven techniques, underlying approaches such as unit selection, and (more recently) the development of statistical parametric modelling approaches similar to those used in speech recognition. Within SCALE we shall foster cross-fertilization between recognition and synthesis research—areas that have developed separately over the past three decades, as well as employing methods from modern machine learning. For example, speech synthesis systems need to model multiple asynchronous streams of data (pitch and MFCCs, in the simplest case), and speech synthesis researchers in SCALE will be in a position to learn from the multistream modelling approaches developed in machine learning.

3.5 Research Collaboration and Exploitation of Synergies

The consortium combines the expertise of eight major academic and industrial sites in speech processing in Europe. Together, they provide the multidisciplinary and intersectorial environment essential for the cross-disciplinary research within SCALE and the corresponding training of the new elite of researchers. The SCALE project has the significant advantage in that there are already close linkages between the partners, e.g. through collaboration in existing projects as well as student exchanges. For instance, IDIAP and UEDIN are currently co-coordinators of the AMIDA integrated project, in which USFD is also a partner. As part of the AMIDA project, UEDIN and IDIAP have collected overlapping speech data with two eight-channel microphone arrays. This data was intended to provide a realistic test bed for beamforming and blind source separation (BSS) algorithms.

UEDIN and IDIAP also organized an open competition, the Speech Separation Challenge (SSC), based on this data. One team led by IDIAP and another led by UDS met this challenge and developed complete and independent solutions for the SSC, which included speaker tracking, speech segmentation, beamforming and recognition modules. This collaboration has been further enhanced in that a doctoral student who had worked

Individual Research Project	UEDIN	Philips	IDAP	Motorola	RUN	RWTH	UDS	USFD
RS-1	•				•			
RS-2			•		•			
RS-3			•		•			
RS-4	•							•
AHSR-1		•				•		
AHSR-2	•			•				
AHSR-3	•		•					
AHSR-4					•			•
SPL-1				•			•	
SPL-2		•					•	
SPL-3				•				•
SPL-4		•				•		
SPL-ER-1				•			•	
SPL-ER-2		•					•	

Table 1: Research collaborations within the SCALE project for each of the projects. Further details about each project can be found in Section 4.3.1.

on the SSC team led by UDS is currently spending one year as a visiting researcher at IDIAP. IDIAP and RWTH currently collaborate in the US-DARPA funded GALE project, concentrating on new acoustic features and system combination. As a by-product, the collaboration also contributed to the RWTH ASR system that scored first in the final 2007 TC-STAR ASR evaluation in both the restricted and public conditions. UDS and USFD have already begun planning to collaborate for the 2008 NIST RT meeting evaluation. In exchange for help from Sheffield in building a suitable language model, Saarland University will make available beamforming algorithms developed during other projects. RUN, USFD and UEDIN have collaborated via visiting researchers in the area of ASR architectures inspired by HSR, and on specific technical issues such as the development of dynamic Bayesian network models for speech recognition.

The partners will extend, strengthen, and exploit these fruitful existing linkages in the context of SCALE and in the future. Table 1 provides an overview of the collaborations within the research themes envisioned for SCALE. One of the tasks undertaken by the SCALE consortium will be the development of a common set of software tools for speaker tracking, beamforming and speech recognition. The functionalities included in this toolbox will reflect the combined knowledge of all consortium partners with experience in far-field ASR. While the toolbox will be initially developed internally by members of the consortium, it will be made available to the larger research community by download from a common website. The availability of such a toolbox will have a significant impact on the state-of-the-art, in that researchers in acoustic array processing who previously were confined to working with plain vanilla ASR systems on trivial tasks such as TI Digits will have access to a state-of-the-art ASR engine along with the acoustic models and recognition networks needed to run experiments and conduct research on far-field ASR tasks of current interest. Mainstream ASR researchers, on the other hand, will gain access to state-of-the-art algorithms for speaker tracking and beamforming, which they most likely would not have taken the time to implement from scratch.

As to speech synthesis, SCALE will be instrumental in upgrading the Festival platform for text-to-speech conversion developed by UEDIN by including procedures for adapting the output of the system to the requirements of the listener and the acoustic environment. Festival will also be enhanced by virtue of the integration of results of the research in hierarchical acoustical models.

SCALE has a strong focus on machine learning. Therefore, toolkits such as the TORCH machine learning library developed by IDIAP will be used intensively in most, if not all individual projects. We anticipate that

several projects will encounter the need to add novel learning algorithms to the existing toolkit. Therefore, SCALE will result in a vastly upgraded version of TORCH, which will be maintained and made available to the research community by IDIAP.

In addition to developing the software tools described above, the SCALE project will also undertake one common data collection campaign. This data collection will be organized around the concept of speech dictation on mobile, hand-held devices such as PDAs and cell phones. The collection itself will be undertaken through a collaborative effort between Saarland University, and the two industrial partners, Motorola and Philips. Moreover, each host institution will offer fellows complementary skills training courses on a local level (i.e. entrepreneurship, presentation skills, gender and migration issues, language courses). The associated SME EURICE will contribute to the overall training programme by providing the fellows with the network-wide complementary skills training during the summer schools, and as such preparing the fellows adequately for their career in industry and/or academia in terms of (i) project and finance management, (ii) IPR, patenting and licensing, (iii) proposal writing and funding opportunities, and (iv) communication, negotiation, and research ethics.

References

[1] J. Barker, M. Cooke, and D. Ellis. Decoding speech in the presence of other sources. *Speech Communication*, 45:5–25, 2005.

[2] M. Bisani and H. Ney. Open vocabulary speech recognition with flat hybrid models. In *Proc Interspeech'05*, pages 725–728, 2005.

[3] M. Brandstein and D. Ward. Microphone arrays. *Springer Verlag*, 2000.

[4] A.S. Bregman. *Auditory scene analysis: The perceptual organisation of sound*. Bradford books, MIT-press, Cambridge (MA), 1990.

[5] K. Demuynck, D. Van Compernolle, and H. Van hamme. Robust phone lattice decoding. In *Proc Interspeech'06*, pages 1622–1625, 2006.

[6] Y. Han and L. Boves. Hierarchical acoustic modeling based on random-effects regression for automatic speech recognition. In *Proc Interspeech'07*, pages 878–881, 2007.

[7] Jeff Hawkins and Sandra Blakeslee. *On Intelligence*. Times Books, New York, 2004.

[8] Frederick Jelinek. *Statistical Methods for Speech Recognition*. MIT Press, Cambridge, 1997.

[9] David Marr. *Vision. A computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman, San Francisco, 1982.

[10] J. McDonough, K. Kumatani, T. Gehrig, E. Stoimenov, U. Mayer, S. Schacht, M. Wölfel, and D. Klakow. To separate speech!: A system for recognizing simultaneous speech. In *Proceedings Machine Learning and Multimodal Interaction*, 2007.

[11] R. K. Moore. PRESENCE: A human-inspired architecture for speech-based human-machine interaction. *IEEE Trans. Computers, Special Issue on Emergent Systems, Algorithms an Architectures for Speech-Based Human-Machine Interaction*, 56(9), September 2007.

[12] D. Norris. Shortlist: A connectionist model of continuous speech recognition. *Cognition*, 52(3):189–234, 1994.

[13] O. Scharenborg, V. Wan, and R. K Moore. Towards capturing fine phonetic variation in speech using articulatory features. *J. Speech Communication, Special Issue on Speech Recognition and Intrinsic Variation*, 49:811–826, 2007.

[14] O. E. Scharenborg, D. G. Norris, L. F. M. ten Bosch, and J. M. McQueen. How should a speech recognizer work? *Cognitive Science*, 29(6):867–918, 2005.

[15] Tanja Schultz and Katrin Kirchhoff (Ed.). *Multilingual Speech Processing*. Elsevier, Academic Press, New York, 2006.

[16] H. L. Van Trees. *Optimum Array Processing*. Wiley-Interscience, New York, 2002.

[17] J.F. Werker and S. Curtin. Primir, a developmental framework of infant speech processing. *Language Learning and Development*, 1:197–234, 2005.

4 Training

4.1 Importance and Timeliness of the Training

As indicated in Section 3.1 all aspects of speech processing are rapidly growing in commercial and industrial importance. However, much of the potential market impact remains untapped because existing technology lacks flexibility and adaptability: The performance of the core speech technologies is inadequate for many applications. It does not provide for simple and unobtrusive verbal interaction with a device, let alone for interaction that is as natural as that with another human being. Moreover, as shown in Section 3.3, the existing problems are due to the fragmentation of research in speech processing which can only be overcome by bridging the gaps between currently little connected research fields. Thus, by bridging existing gaps between various areas like ASR, machine learning, signal processing, HSR and speech synthesis SCALE will make a significant and urgently needed contribution to move the research of speech processing forward and enable the development of high performance speech technology in the European market.

The central goal of the SCALE training program is to train ESRs and ERs in such a way that they can benefit from progress in different fields and, beyond that, get a broader background on which they can build their early careers and also adjust more easily to inevitable changes in their specialisms. The training program will therefore not only be multi-disciplinary but also inter-sectoral, meeting the requirements of both industry and academic research. As the objective of research in academia and in industry is different, it is of utmost importance for young researchers to know how to function well in both environments. Here secondments play the key role, where ESRs and ERs alike can experience the research environment in two major companies. By training young scientists in a way that they will be well-poised to make contributions in both industry and academia, SCALE expects to make a major contribution by securing the leading position of Europe in fundamental research in speech processing and putting European companies in a position to reap the fruits of the advances in speech science and technology.

To maximize the training benefits for participating ESRs and ERs, the training will reflect the current demands of academia and industry, such as knowledge and application of the latest research techniques, cross-disciplinary thinking and communication. The specific objectives of the SCALE training programme are

1. to provide each researcher with opportunities for inter-European networking and information exchange, cross-disciplinary and intersectorial expert knowledge in the field of speech processing and complementary skills (such as research and project management, IPR, proposal writing etc.);

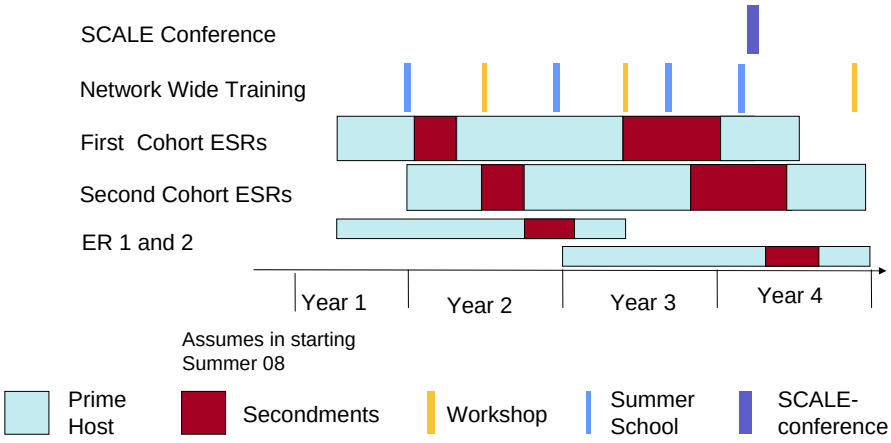


Figure 1: Time plan for SCALE training activities.

2. the development of a cadre of broadly knowledgeable top-level European scientists who are able to flourish in both academic and industrial environments;

3. the development of strong cooperation and close links between academia and industry across Europe;

4. to strengthen Europe's leadership and competitiveness in the field of speech processing through the comprehensive education and qualification of human potential.

4.2 Overview and Structure of the Training Activities

Figure 1 provides an overview of the SCALE training programme. It combines local research training (i.e. the individual research projects, local technical and local complementary training), network wide activities (i.e. workshops and summer schools), industrial experience (i.e. secondments) and the international SCALE conference.

4.3 Local Training

The local training consists of the individual research training projects, local technical training and local complementary skills training. The major part will be the individual research training projects in which the fellows co-operate with two sites in the network.

4.3.1 Individual Research Projects

The research projects which will coincide with the ESRs Ph.D. topics planned within the SCALE are project are listed in Table 2 along with the responsible supervisors.

Table 3 provides an overview of the milestones in the first, second, and third year of each Ph.D. topic.

For all individual research projects, the fellow's personal career development plan (CDP) will list specific milestones in addition to those in Table 3. Common to all projects is the goal to publish. We expect both ESRs and ERs to publish one journal article or reviewed conference paper per year. Journals could be, for example, Speech Communication, Computer Speech and Language, Journal of the Acoustical Society of America, IEEE Transactions on Audio, Speech and Language Processing and IEEE Transactions on Signal Processing. The best conferences are IEEE ICASSP and Interspeech. The following subsections (see subsection 4.3.2, 4.3.3, 4.3.4) will provide a detailed overview of the planned individual research topics for each research theme.

Table 2: *Planned individual research projects for the ESRs and ERs.*

	Topic	First Supervisor	Second Supervisor
RS-1	Trajectory HMMs for Reactive Speech Synthesis	Simon King (UEDIN)	Bert Cranen (RUN)
RS-2	Towards Speaker Invariance & Compensation for Coarticulation	Hynek Hermansky (IDIAP)	Odette Scharenborg (RUN)
RS-3	Hierarchical Trajectory Models for Speech Recognition	Lou Boves (RUN)	Hynek Hermansky (IDIAP)
RS-4	Speech Synthesis by Analysis	Roger Moore (USFD)	Korin Richmond (UEDIN)
AHSR-1	Towards Open Vocabulary Speech Recognition	Hermann Ney (RWTH)	Erhard Rank (Philips)
AHSR-2	Data Association Multisource Acoustic Models	Steve Renals (UEDIN)	Holly Francois (Motorola)
AHSR-3	Sounds and Spoken Language	Hynek Hermansky (IDIAP)	Joe Frankel (UEDIN)
AHSR-4	Associative Memories for Learning and Decoding Speech	Lou Boves (RUN)	Roger Moore (USFD)
SPL-1	Non-Gaussian Beamforming for Far-Field ASR	John McDonough (UDS)	James Rex (Motorola)
SPL-2	Particle Filters for Robust Far-Field ASR	Dietrich Klakow (UDS)	Heinrich Bartosik (Philips)
SPL-3	Multi-Channel Modelling for Automatic Speech Recognition	Thomas Hain (UFSD)	David Pearce (Motorola)
SPL-4	Investigations on Feature and System Combination Methods	Hermann Ney (RWTH)	Heinrich Bartosik (Philips)
SPL-ER-1	Multiple Microphone Techniques for Dereverberation	John McDonough (UDS)	David Pearce (Motorola)
SPL-ER-2	Beamforming and Noise Reduction Using Auditory Processing	Dietrich Klakow (UDS)	Heinrich Bartosik (Philips)

4.3.2 Bridging the Gap between Recognition and Synthesis (RS)

Topic RS-1 Trajectory HMMs for Reactive Speech Synthesis: Conventional speech synthesis approaches do not react or adapt to the dialogue context. In this project we aim to automatically adapt the behaviour of a statistical speech synthesiser based on factors obtained from processing the speech of the dialogue partner, such as speech recognition confidence, speaking rate and loudness.

- Goal: We seek to improve the naturalness of speech synthesis in the context of a natural dialogue, in terms of subjective evaluation in a spoken dialogue system
- Relevance: This research aims to make the speech generation component of a human-computer dialogue system sound more natural by adapting appropriately to the user (e.g. in case of recognition errors) simplification of ASR schemes while at the same time reducing the demand for large quantities of training data. On the other side, this should also lead towards better integration of coarticulation knowledge inside speech synthesis system.

Topic RS-2 Towards Speaker Invariance Compensation for Coarticulation: When viewed as just sequences of cepstral vectors, speech seems to exhibit an enormous degree of random variation. Yet, experiments with

Table 3: *Milestones for research projects (relative to start of the individual research project)*

	Ind. Research Projects	Year 1	Year 2	Year 3
RS-1	Trajectory HMMs for Reactive Speech Synthesis	Baseline system; construct mathematical models	Implementation of models and testing	Empirical studies on synthesis in the context dialogue system
RS-2	Towards Speaker Invariance Compensation for Coarticulation	Implementation of extractors of detailed phonetic features	Mixed-effects models for predicting phonetic features	Measure effect on quality of unit selection synthesis
RS-3	Hierarchical Trajectory Models for Speech Recognition	Mathematical formulation of hierarchical model	Development of efficient search algorithms	Comparisons studies with current techniques
RS-4	Speech Synthesis by Analysis	Model for quantitatively assessing speaking environment	Implementation of environmental and test	Further refinement based on empirical studies
AHSR-1	Towards Open Vocabulary Speech Recognition	Modeling and recognition architecture for unseen words	Data-driven modeling and language specific aspects	Comparing automatic techniques with expert/oracle knowledge
AHSR-2	Data Association Multisource Acoustic Models	Mathematical formulation of common sound sources	Implementation of source separation algorithms	Empirical studies: ASR performance on overlapped speech using single mic.
AHSR-3	Sounds and Spoken Language	Definition of features and their representations	Implementation of feature extractors and training procedures	Performance comparisons of conventional and universal features
AHSR-4	Associative Memories for Learning and Decoding Speech	Development of linguistic and para-linguistic models	Implementation of assoc. memory architecture	Empirical studies on ASR performance improvements
SPL-1	Non-Gaussian Beamforming for Far-Field ASR	Mathematical formulation of non-Gaussian pdfs	Implementation of beamforming algorithms	Development of online algorithms
SPL-2	Particle Filters for Robust Far-Field ASR	Selection of suitable form of particle filter	Incorporation of particle filter and beamforming	Empirical comparisons with other post-filter techniques
SPL-3	Multi-Channel Modelling for Automatic Speech Recognition	Novel methods for channel modelling	Incorporation of channel model into ASR engine	Empirical comparisons with array processing techniques
SPL-4	Investigations on Feature and System Combination Methods	Mathematical formulation; e.g., Bayes' error on confusion nets	Implementation of optimal combination algorithms	Empirical studies of different combination levels.
SPL-ER-1	Multiple Microphone Techniques for Dereverberation	Mathematical formulation: effect of reverberation	Experimental comparisons with array processing techniques	—
SPL-ER-2	Beamforming & Noise Reduction Based on Auditory Processing	Mathematical formulation: effect of auditory models	Experimental comparisons with array processing techniques	—

speech synthesis and speech perception have made it abundantly clear that there is essential structure underlying this deceptive superficial randomness: adding random changes to parameter tracks or minor discontinuities in tracks at the borders between successive units in concatenative synthesis result in audible and undesirable perceptual effects. At the same time it appears that listeners can decode subtle details in speech signals, for example to distinguish between the monosyllabic word "ham and the first syllable of the word "hamster . In this project we will look for representations of speech signals that make it easier to account for superficial variation in terms underlying invariant linguistic representations. A pivotal factor in this endeavor is coarticulation; we will not only investigate the effects of immediate neighbouring sounds, but also the effects of the syllable, word and prosodic unit of which speech sounds are a part.

- Goal: To develop novel feature representations for conversational speech and mixed effects linear statistical models comprising a rich set of linguistic and phonetic predictors to discover the structure underlying superficial variation in speech signals.
- Relevance: For building and annotating databases for high quality concatenative speech synthesis more insight is needed in the relation between underlying structure and superficial variation in speech signals. This insight is also a necessary for improving the selection of units for concatenation and for implementing the smoothing across unit borders. Insight in the structure underlying superficial variation will also enhance decoding accuracy in ASR.

Topic RS-3 Hierarchical Trajectory Models for Speech Recognition: Template Models (e.g. episodic models) of speech units make it easy to account for detailed aspects of speech signals that make speech synthesis natural and speech recognition difficult. For Structural Models (e.g. HMMs) this is precisely the opposite. Hierarchical models, inspired by emerging knowledge about the structure and functioning of the human neocortex, combine the advantages of template and structural models. In this project we investigate ways of making hierarchical models computationally feasible, and research the use of these new models to bring to bear context dependence in long time windows.

- Goal: We intend to use concepts from unit selection speech synthesis and human speech processing to improve the performance of automatic speech recognition with hierarchical acoustic models.
- Relevance: This project intends to develop a principled way for dealing with the trajectory folding problem that complicates automatic speech recognition, especially for casual speech with many articulatory reductions and for speech in noisy environments. The project will profit from synergy with ongoing research in unit selection speech synthesis, while at the same time reaping the benefits from emerging knowledge about human speech processing.

Topic RS-4 Speech Synthesis by Analysis: Human speakers make constant adjustments to their vocal output in response to a range of different factors. For example, local and global changes are made as a function of the level of interference and noise in the acoustic environment, their familiarity with the listener(s), the formality of the situation, and the phonetic content of the material that is being spoken. Such behaviour is not exhibited by state-of-the-art speech synthesisers, and hence their expressive and communicative abilities are fundamentally limited by the static nature of their underlying generation process. This project aims to research the possibilities for introducing a feedback path into automatic speech generation such that adjustments can be made in real-time as a function of its perceived effectiveness.

- Goal: To develop a novel form of speech generation that will automatically optimise its communicative effectiveness as a function of its speaking environment.

- Relevance: This research will extend contemporary research in speech synthesis by drawing on models of human speech production that involve established on-line adjustments (such as the Lombard effect) as well as borrowing techniques from automatic speech recognition to be placed in the feedback path. It is expected that the knowledge gained will not only inform computational models of human speech production, but will also lead to a greater understanding of some of the hidden variables that need to be taken into account by the next generation of automatic speech recognition.

4.3.3 Bridging the Gap between ASR and HSR (AHSR)

Topic AHSR-1 Towards Open Vocabulary Speech Recognition: Conventional state-of-the-art speech recognisers use a closed and word-based dictionary. This approach has proven to be feasible for languages like English and for domains where the vocabulary is known in advance. Very much unlike human listeners, ASR based on such dictionaries cannot recognise new words and reaches its limits in the case of morphologically rich languages. So alternative approaches are required. As an example, this could be the recognition of subword units which are then composed to words using various knowledge sources.

- Goal: Development and implementation of algorithms to close this gap between ASR and HSR. Validation and comparison with existing approaches on a suitable task.
- Relevance: A closed and (purely) word based dictionary is insufficient for many important applications like for example the recognition of Broadcast News or morphologically rich languages.

Topic AHSR-2 Data Association Multisource Acoustic Models: In the auditory scene analysis approach to processing multiple acoustic sources, features such as harmonicity, onset time and location are used to drive speech separation processes. Is it possible to learn approaches for processing multiple acoustic sources from data? This project will explore the use of probabilistic generative models for data association to address this problem.

- Goal: We seek to reduce the word error rate by 25% in the context of the recognition of overlapped speech, using a single far-field microphone.
- Relevance: The separation and recognition of multiple acoustic sources, from a single acoustic channel is core to many realistic speech recognition tasks, e.g. recognition of telephone speech and speech from broadcast content.

Topic AHSR-3 Sounds and Spoken Language: This project addresses the issue of language independent subword units introduced in Section 3.3.2. In this project we intend to develop techniques for representing speech signals in the form of universal articulatory and acoustic features, and we will develop representations of the signals in terms of the novel features that are suitable for integration in probabilistic decoders.

- Goal: The definition of a schema in which acoustic events can be described in ways which are language and context-invariant, and the subsequent automatic extraction of such features. In the terms introduced by Marr [9], the representation definition is a level 1 problem of computational theory, its extraction is a level 2 problem of algorithm development.
- Relevance: The approach to be investigated in this project addresses an important gap between HSR and ASR. Human listeners have the capacity of representing speech sounds in a largely language independent manner, and current automatic recognisers do not. Accounting for the computational aspects of the human representations and processing in automatic systems will yield substantial improvements of the accuracy of the decoder, especially under adverse acoustic conditions.

Topic AHSR-4 Associative Memories for Learning and Decoding Speech: It is reasonable to assume that human speech processing is more robust than automatic processing because humans are able to bring to bear multiple sources of context information. Different sources of context information are linked in densely connected networks. In this project we will focus on linguistic and para-linguistic constraints on pronunciation variation. To learn the connections between the individual sources and to be able to harness the learned connections during recognition of new input speech we will explore associative memory architectures, informed by knowledge about human speech processing.

- Goal: This project aims to improve automatic speech recognition performance by harnessing multiple and densely interconnected context constraints. At the same time, the research will put tentative models of human speech processing to the test.
- Relevance: Maximal use of complex context constraints is a necessary prerequisite for closing the gap between automatic and human speech recognition. Associative memories seem to be a promising architecture for learning the context constraints and for using these in actual recognition tasks in real-time with affordable computational resources.

4.3.4 Bridging the Gap between signal processing and learning (SPL)

Topic SPL-1 Non-Gaussian Beamforming for Far-Field ASR: Conventional beamforming algorithms assume that all sources, both desired and undesired, emit stochastic signals are Gaussian-distributed. Recent research has revealed, however, that neither speech signals nor complex subband samples thereof are Gaussian, and that superior beamforming performance can be obtained by exploiting this fact.

- Goal: We seek to reduce the word error rate by a factor of 25% on a far-field speech recognition task such as the speech separation challenge through the use of non-Gaussian beamforming algorithms.
- Relevance: This research will be highly relevant to the challenge of making ASR without a close-talking microphone feasible, and thereby making speech the man-machine interface of first choice.

Topic SPL-2 Particle Filters for Robust Far-Field ASR: Recent research has revealed that particle filters are effective as a means of post-filtering a speech signal previously enhanced through beamforming. Through further research on such post-filters and their combination with the non-Gaussian techniques described above, we will obtain significant improvements in far-field ASR performance.

- Goal: We seek to reduce the word error rate by a factor of 20% on a far-field speech recognition task such as the speech separation challenge through the use of particle filters applied to the output of a beamformer.
- Relevance: This research will be highly relevant to the challenge of making ASR without a close-talking microphone feasible, and thereby making speech the man-machine interface of first choice.

Topic SPL-3 Multi-Channel Modelling for Automatic Speech Recognition: This project addresses the problems explained under the heading of 'Exploiting Multiple Microphones' in subsection 3.3.3.

- Goal: It is standard to include multiple microphone recordings into automatic speech recognition by adopting an enhancement based approach which is mostly independent of the construction of the speech recognition system. Here the objective is to investigate adaptive methods that optimise the front-end feature extraction for all channels in conjunction with the acoustic models for speech recognition. This will include:

  - Transformational schemes on signal and/or feature level;
  - Integrated noise compensation for all channels;
  - Temporal variation of processing using adaptation techniques;
  - Multi- and single-channel modelling and combination.

- Relevance: The use of multiple microphones is important for many applications. The project attempts to investigate generic approaches to using these sources in an integrated fashion without the requirement for exact knowledge of geometry.

Topic SPL-4 Investigations of Feature and System Combination Methods: In practice different knowledge sources (e.g. features or systems) are available for ASR, but they are all imperfect and often complementary. So combining them can lead to improvements in performance. Which are good candidates to combine and which is the optimal way to do it?

- Goal: A better comprehension of the above questions, e.g. exact Bayes error on word level vs. approximations like for example frame error and confusion networks.
- Relevance: Recent projects (e.g. TC-STAR) have shown that system combination is a simple and effective method to achieve significant improvements over the best single system. Existing combination techniques use various heuristics and approximations which is due to the complexity of the problem.

Topic SPL-ER-1 Multiple Microphone Techniques for Dereverberation:

- Goal: We seek to reduce the word error rate by a factor of 25% on a far-field speech recognition task through the use of multimicrophone dereverberation algorithms.
- Relevance: This research will be highly relevant to the challenge of making ASR without a close-talking microphone feasible, which will enable a host of applications on mobile devices.

Topic SPL-ER-2 Multi-Microphone Beamforming and Noise Reduction Using Auditory Processing:

- Goal: We aim for the combinination of beamforming and noise reduction based on auditory processing for multiple microphone signals, i.e. mimicking the human aural perception process. Improvements in ASR rate as compared to separate processing steps are expected.
- The enhancement of speech signals for the improvement of ASR is of vital importance for the application in noisy environments and for far-field speech recognition.

4.3.5 Local Technical and Complementary Skills Training

In addition to the individual research project, fellows will have full access to local technical training courses in the field of speech communication available at their host institution. The choice of courses suitable for each fellow will be discussed with the fellow and defined in his/her personal career development plan (CDP), as described in Section 4.7.2. Furthermore, complementary skills training will form another important component of the SCALE local training training. Similar to the local technical training, it will be tailored to the demands and challenges of the fellows and will take into account the latest educational insights into the learning process of students and researchers. As a result, ESRs and ERs will receive a unique and comprehensive training which will not only include scientific knowledge but also important meta-scientific aspects. On a local level, the following complementary training opportunities are available at each host institution:

1. *Entrepreneurship:* to give fellows an insight into scientific enterprise including preparation of a business plan, marketing, and legal aspects;

2. *Presentation skills:* to provide fellows with complementary skills in free speech, making presentations, using corresponding software etc.

3. *Gender and migration issues:* to support fellows in pursuing their individual careers in harmony with individual life plans;

4. *Language courses:* to help fellows acquire the relevant command of scientific English or other European languages relevant for their research training and industrial placements.

Additional complementary skills training courses which are available at the host institution may be selected in accordance with the fellow's personal CDP.

4.4 Secondments and Industrial Experience

In accordance with their individual research topics, the vast majority of the fellows (approx. 90%) will be required to spend some time working with one of the industrial partners. The expectation is that ERs will have a "home" academic lab, taking at least one (and preferably two) internships in one of the industrial laboratories. One internship (about three to six months) will take place in the first year to give fellows an insight into the industrial working condition and dimension of their topic from the start. A second internship will take place in the latter portion of the second or early in the third year, so that the researchers gain the experience of transitioning theoretical knowledge into marketable products. This would ensure a pragmatic aspect to the solution proposed by the ESR, as well as the opportunity to assess the exploitation potential of the research. Overall, SCALE fellows working with the industrial partners will thus benefit from the experience of working in an industrial R&D environment, bringing a more commercial and applied focus to their research. They will develop and practice additional complementary skills arising from the industrial setting, for example project management and commercial exploitation of results. The remaining fellows (approx. 10%) whose individual reserach topic is more suitable to be seconded in another academic institution will benefit from the comprehensive specialist expertise available at their secondment institution. This will not only include the scientific expertise, but also the opportunity to get to know different working conditions (e.g. in Germany or in the U.K.), which again contributes to the fellows' overall international experience and excellence, benefits and strengthens his position in the international research community, and allows for the forging of close ties on a personal level.

4.5 Network Wide Training

The network-wide training will not only be available to the SCALE fellows, but also be open to researchers beyond the community of SCALE fellows, for example PhD students and post-docs in the labs of network partners. Network-wide training activities available within SCALE include both summer schools and workshops, each of which will be described in further detail below.

4.5.1 Summer Schools

Generally, we expect summer school participants to come from a variety of backgrounds, including speech perception, machine learning, computer science, signal processing, and speech technology. As such, the annual summer schools will provide an ideal forum for getting to know other researchers in one's own or related

Month 11	**First Summer School (UDS)**	
Theme Training	Distant Speech Recognition	UDS
Complementary Training	IPR, Patenting and Licensing	EURICE
Other courses	Finite-State Transducer Methods in ASR System Combination Techniques	UDS RWTH
Month 20	**Second Summer School (USFD)**	
Theme Training	Spoken Language Processing by Mind and Machine	USFD
Complementary Training	Communication, Negotiation and Research Ethics	EURICE
Other courses	Computational Models of Human Speech Processing Digital Processing of Speech and Images	RUN RWTH
Month 31	**Third Summer School (UEDIN)**	
Theme Training	Adaptive Speech Synthesis	UEDIN
Complementary Training	Project and Finance Management	EURICE
Other courses	Statistical ML for Speech Synthesis Acoustic Array Processing Techniques	UEDIN UDS
Month 39	**Fourth Summer School (RUN)**	
Theme Training	Beyond HMMs	RUN
Complementary Training	Proposal Writing and Funding Opportunities	EURICE
Other courses	Brain Imaging in Psycholinguistics Pattern Recognition and Neural Networks	RUN RWTH

Table 4: Schedule of SCALE summer schools.

areas, the exchange of best practice and ideas, and networking activities. They will consist of short presentations by the fellows, complementary skills training, in-depth courses on a particular research theme under which the summer school is headed and additional technical courses. Overall, we plan to start with one day of short student presentations, intended at updating the network on their progress and giving an opportunity to discuss new ideas. The second day will largely be devoted to complementary skills training (including training on IPR, project management, etc.). The major part of the summer schools will comprise the specialised courses that will be of interest across the network. Each summer school will have a specific research theme, for which in-depth training courses are provided. In addition to these thematic "gap-bridging" courses, additional courses will be offered including tutorials on basic techniques (e.g. the EM algorithm, finite state transducer models, computational models of human speech processing), hands-on laboratory sessions (e.g. Festival speech synthesis, the TORCH machine learning toolkit — both of which were developed by SCALE partners), and advanced courses (e.g. dynamic Bayesian network inference, articulatory feature extraction). Table 4 provides an overview of the planned contents of each summer school in terms of training on the particular reserach theme, complementary training, and additional courses available to the fellows - each of which will be described in greater detail below.

The detailed description of each summer school theme and the resulting technical training courses is as follows:

1. *Distant Speech Recognition:* The first summer school will cover the current state-of-the-art in automatic speech recognition. This will include presentations on HMMs, feature extraction and search techniques based on weighted finite-state transducers. In addition, there will be modules on speech recognition with far-field microphones and related techniques such as speaker tracking and beamforming.

2. *Spoken Language Processing by Mind and Machine:* The main focus will be the link between human and machine speech recognition. The courses will compare and contrast spoken language processing as performed by machines with the corresponding processes performed by human beings. Theories of

human speech perception, production, cognition and discourse will be discussed alongside algorithms for automatic speech recognition, synthesis, understanding and dialogue. Attention will be given to the latest research attempts to unify these areas within a common framework.

3. *Adaptive Speech Synthesis:* Adaptive machine learning approaches are making a major impact in speech synthesis, for example in the development of statistical parametric models, such as the trajectory HMM. This summer school will make SCALE researchers aware of recent results of statistical machine learning approaches in speech synthesis. The course will include extensive laboratory work using the FESTIVAL and HTS systems, with an emphasis on systems that can adapt in an unsupervised manner.

4. *Beyond HMMs:* For many good reasons, hidden Markov models (HMMs) have dominated the ASR field for the past 20 years. Thus it is no surprise that it is often the only technology that new researchers coming to the field consider to be viable. The main goal of this particular school is to make young researchers aware of possible alternatives to HMM-based ASR. The school will discuss the history of ASR and (some good and some perhaps insufficient) reasons for abandoning particular historical approaches, to critically evaluate the positive and negative aspects of HMM-based ASR, and discuss possible alternatives, (such as template-based/episodic models, hierarchical subword unit based recognisers, discriminative keyword-based approaches, various parallel multi-stream schemes etc). This summer school is planned to happen just before Interspeech 2011.

As indicated above, the summer schools will also provide network-wide complementary skills training courses to all fellows. The network-wide complementary training courses within SCALE were deemed indispensable by both the academic and industry partners for the future career of ESRs and ERs and as such, will give all fellows a comprehensive preparation for their future career in both industry and academia. The courses will be provided by the associated partner EURICE (a member of both the European Association of Research Managers and Administrators and of the IPR-Helpdesk) as part of the summer schools. It will be similar to the successful skills programme "Young Leaders in Science" and include the following four modules:

1. *IPR, Patenting and Licensing:* the fellows will learn to understand the exploitation-related issues of their research, issues related to knowledge transfer and sharing, and to the practical use of research and its products;

2. *Communication, Negotiation and Research Ethics:* the fellows will not only acquire relevant negotiation and moderation techniques as well as knowledge of ethical isses specific to their research area, but also learn how to deal with conflicts in organisations and how to build and maintain networks.

3. *Project and Finance Management:* the fellows will acquire basic techniques of project and finance management including planning, controlling, budgeting, co-ordination and reporting;

4. *Proposal Writing and Funding Opportunities:* the fellows will learn where and how to apply for research funding both on a national and on a European level.

4.5.2 Workshops

SCALE workshops will take place once a year in the winter. They will be of 2-3 days duration and include presentations of research results, progress on benchmark tasks, discussion of future directions, and supervisory team meetings for the fellows. A highlight of the SCALE workshops will be invited keynote speakers, from outside the network. We also plan to use the meeting recording system developed by IDIAP during the AMIDA project to make indexed video of the workshop presentations together with projected data available for download over the internet. Overall, the workshops will thus serve as opportunity for fellows to present themselves

and their research on a European platform, to get in touch with fellow researchers, to exchange ideas and discuss problems within and outside one's research area, to obtain more personalized feedback on their reserach and progress than would be possible at the summer shool, and to get a glimpse of what is happening outside the network.

4.6 SCALE Conference

In addition to the local and network-wide training acitivites, we plan to hold a SCALE conference in 2011, as we expect the SCALE programme to reach its highest level of productivity in that year. We plan to hold it as a satellite conference preferably to Interspeech, in which case it would be in September/October 2011, or alternatively jointly with ICASSP, depending on which of the two conferences is held in Europe. This allows us to reach a critical mass of experts from outside the network and to make the brand SCALE, its research activities, and its fellows known internationally. For the SCALE conference we will solicit papers, both from members of the consortium and the larger research community, in each of the subject areas addressed by the project: automatic and human speech recognition and synthesis, signal processing and adaptive learning. We expect about 100 participants from outside the network. In organising the conference, we can draw on the experience of members of the network who organised Interspeech in 2007 (RUN), the MLMI conference series (UEDIN and IDIAP, 2004-2007) and ICPhS (UDS, 2007). USFD will chair Interspeech in 2009. The staging of such a conference will also provide a further practical educational experience, inasmuch as the project ER and the ESRs will be involved in all aspects of the conference, from the call for papers to the selection of the venue, and the conference organisation. Along with the leading authorities in each area from the consortium and the larger research community, the fellows will also participate in the review of papers submitted to the conference and the organisation of the technical programme.

4.7 General Aspects of the SCALE Training

4.7.1 Supervision and Mentoring

The personnel involved in SCALE includes the 14 fellows (12 ESRs and two ERs, one ER at any point in time except for a short hand-over period) and the senior scientists presented below in section 5.1. Following the practice which has proven successful in other EC-funded training networks in which the SCALE partners have participated (e.g. PIRE, SPHERE, SPHEAR, HOARSE, S2S), each fellow will be assigned a supervisory team charged with mentoring, monitoring and guiding her/his progress. The supervisory team will consist of (i) the first supervisor in the host institution, with whom the fellow will work on a daily basis, (ii) the second supervisor from a different partner institution, typically one in which the fellow will take up a secondment (see Section 4.3 for details on supervisors and individual research projects). Both first and second supervisor will be selected based on their expertise in this particular area, thus ensuring that each fellow is being supervised by the two most knowledgable senior scientist his/her research area. The supervisory team will formally meet on a 6-monthly basis (independent from the meetings at workshops and summer schools) by telephone/video conferencing. For each of these supervisory board meetings, the fellow will produce a progress report. This report will give details about the research progress and training and include plans for the next 6-monthly period. Most of the business of the supervisory team will be to consider and and provide detailed feedback on the reports. The reports will also be put on the SCALE internal wiki to make the progress transparent and to give the supervisory board an overview of the state of the project.

4.7.2 Personal Career Development Plans (CDPs)

The diversity of backgrounds of the fellows, combined with the wide range of training offered within SCALE makes it essential that each fewllow has a well worked out and personalized Career Development Plan (CDP) which will take full advantage of the available training (including both local and network-wide as well as technical and complementary skills training activities). Thus, for each fellow a personal CDP will be defined at the beginning of her/his research training (see Figure 2). The procedure for the identification of the fellows' research and training needs will be as follows: Each fellow will undergo a brief evaluation of knowledge and skills to take stock of the status quo and to define training needs. This process will be performed openly and with mutual input, i.e. the ESR and ER in question will actively raise her/his own wishes and requirements for training, based on personal preference and career plans for the future. Based on this assessment, the ESR/ER and the first and second supervisor will develop the individual CDP with defined learning outcomes, detailed milestones and a timetable for implementation. To ensure coherence within the network and the training needs of both industry and academia, each CDP will be evaluated by the supervisory board. As such, the personal CDPs are expected to have a clear interdisciplinary and inter-sectoral orientation to industry and academia. Every six months the fellow, together with her/his supervisors, will review and update the personal CDP at the supervisory board meetings.

4.7.3 Early Stage vs. Experienced Researchers

Each ESR will be offered one of the twelve individual research topics presented in Section 4.3, Table 2 (see Topics RS-1 to SPL-4) as topic for their PhD. The local research training for ESRs will include both their individual reserach projects, local technical training (e.g. graduate courses in the particular research topic that are offered locally by the host institution) and, if necessary and in accordance with the personal CDP, local complementary skills training courses offered locally by the host institution (see section 4.3). In addition, the vast majority of ESRs will spend 3 to 6 months of secondment with an industrial partner as required by their individual research project. In the remaining cases, where a secondment with a different academic institution is more suitable to the individual research topic, the ESRs will have their secondments with this particular institution.

Furthermore, all ESRs will be required to participate in the network-wide training activities offered within SCALE (i.e. both summer schools and workshops) as well as in the international SCALE conference. This will facilitate the networking among fellows on a multi-disciplinary and pan-European level. While the workshops give fellows the opportunity to present their research and as such, serve as a measure of progress monitoring and control, the summer schools will provide a wide range of complementary and technical training. Last but not least, the SCALE conference will not only allow for networking beyond the SCALE ITN, but will also provide a practical educational experience, inasmuch as the ESRs will be involved in all aspects of the conference, from the call for papers to the selection of the venue, and the conference organisation.

Each ER will be assigned an individual research project (see Topics SPL-ER-1 and SPL-ER-2 in Table 2, Section 4.3). Principally, ERs will have the same access to both local research training available at the host institution and the network-wide training opportunities offered within SCALE. However, given that the two ER positions which will be offered to early post-docs, the ERs will have a much greater research expertise and specialist background than ESRs. As such, they will to a much larger extent be expected to develop their own interdisciplinary research programme, and actively extend their research outlook. Their training will therefore include expanded teaching opportunities, research supervision (typically of Masters students), the opportunity to develop and submit research proposals, and a much greater degree of research and project management so as to adequately prepare them for lectureship positions, assistant professorship, or positions in industry. In particular, ERs will be expected to actively contribute to the network-wide training courses in terms of teaching, to organise one of the workshops, and to contribute to the organisation of the summer schools. Through this targeted training, we expect the SCALE ERs to develop into independent scientists, and

Table 5: *Overview of Researchers to be Financed by the Contract in Person Month*

Network Team	ESR (A)	ER (B)	VS (C)	Total (A+B+C)
IDIAP	72	0	0	72
UEDIN	72	0	0	72
RUN	72	0	0	72
RWTH	72	0	0	72
UDS	72	48	0	120
USFD	72	0	0	72

they will form a significant route for the transfer of knowledge and expertise within the network.

4.7.4 Networking and Exchange of Best Practice

SCALE will provide various different opportunities for networking and the exchange of best practice. The workshops, summer schools, and the SCALE conference present the most intensive forms of information exchange and networking opportunities within SCALE. The SCALE partners expect to collaborate on a pan-European, inter-sectorial and cross-disciplinary level through the organisation of, and participation in, these training events. In addition to this, the SCALE partners will directly involve external experts from both industry and academia as members of the supervisory board and guest/keynote speakers at training events. The secondments will provide another means of networking. They will have a typical duration of 3 to 6 months, and may focus on forging new research links, or to fully exploit the network's available resources (such as the instrumented meeting rooms and proprietary speech corpora available at some of the partners). Day-to-day communication within the network will be by email, internal project web sites and wikis. In addition, telephone/video conferencing will be used where appropriate, particularly for the supervisory board, which is only likely to meet physically at the six-monthly workshops/summer schools. Last but not least, it is important to note that the networking activities and the exchange of best practice will even go beyond SCALE as the network partners are and will be collaborating within other national and international projects (e.g. EdSST, AMI, PASCAL, TC-STAR, DIRAC, PRE, S2S, (IM)2, PIRE).

4.8 Size and Balance of the Training Program

Within SCALE each of the academic network partners will host two ESRs. In addition, the co-ordinator UDS will host two ERs. Table 5 displays the number of ESRs and ERs to be financed by the contract. As such, the proportion between ESRs and ERs reflects the typical balance prevalent in most research groups participating in SCALE.

In the past, all SCALE partners have already handled projects much larger than SCALE and hence the expectation is that also SCALE matches the capacities of the partners: Examples of past projects larger than SCALE are the joint International Research Training Group by UDS and UEDIN involving more than 30 ESRs from both sites. The Partnership for International Research and Education (PIRE), an ITN-like co-operation of UDS with Johns Hopkins University and Brown University involves 21 ESRs and 6 ERs. Moreover, USFD coordinates the training part of the FP6 AMIDA project. RUN has similarly shown its capacity to host fellows already in the S2S RTN. IDIAP and UEDIN jointly coordinate AMI and AMIDA. 50 ESRs and ERs have undergone internships on the AMI and AMIDA training programmes. RWTH is part of GALE. Complementary skills training provider EURICE has not only extensive experience in the provision of complementary research training, but also been involved in various European research projects training networks such as the IPR Helpdesk or Galenos.

5 Implementation

5.1 Capacities of Each Partner

Saarland University (UdS): Saarland University and its department of computational linguistics and phonetics is very active in speech and language related fields with overall six research groups and more than forty researchers. UdS is active in speech recognition, speech synthesis and machine learning for human-computer interaction and language processing. The department of computational linguistics and phonetics has participated in large national and international projects like VerbMobil and TALK. Presently it is coordinating the German part of PIRE, an NSF/DFG funded international research training group. The campus of UdS is also the home of the German research center for artificial intelligence (DFKI) with its 180 researchers working in eight departments. Moreover, to facilitate the conduct of international research projects, UdS has established the European Project Office EPO which actively supports researchers in the day-to-day management of such projects by taking care of non-scientific tasks such as the financial and administrative co-ordination and since 2000 has managed all projects and training networks from FP4-FP7 co-ordinated by the UdS.

Prof. Dietrich Klakow has extensive experience in the construction of large vocabulary conversational speech recognition (LVCSR) systems, and participated in the Broadcast News evaluations sponsored by the US National Institute of Standards and Technologies in 1997-1999. He was a manager of the dictation research group at Philips Research in Aachen, Germany, managing projects worth 1.5 million Euros per year. Since 2003 he has been professor of electrical engineering at Saarland University and supervises approximately nine graduate students. Presently he coordinates one of the subprojects of SmartWeb as well as the German part of PIRE.

Dr. John McDonough assembled and supervised the team responsible for collecting multimodal data at the University of Karlsruhe (UKA) in connection with the EU project CHIL, *Computers in the Human Interaction Loop*. Dr. McDonough also led the UKA research effort for developing far-field ASR technology, and supervised UKA's participation in the audio technologies portion of the CHIL evaluation campaigns. From August 2006 until June 2007, Dr. McDonough led the team consisting of members from UKA and Saarland University that developed a system for recognizing *simultaneous* or *overlapping* speech captured with eight-channel circular microphone arrays. This system achieved the lowest word error rate in the PASCAL Speech Separation Challenge, Part II.

Ms. Corinna Hahn, has been involved in the administration and management of European RTD projects and research training networks for more than 10 years. She joined EPO in 2000 and since then has successfully managed various international research projects and training networks, e.g. Calculemus, MESEMA, MESA, NEPSA, and, more recently, DEXMART and PREDATOR.

Key personnel: Prof Dietrich Klakow (overall scientific co-ordinator) (10%), Dr John McDonough (50%), Ms. Corinna Hahn (administrative and financial co-ordination) (30%).

Key publications/patents:

1. J. Kneissler and D. Klakow,"Speech Recognition for Huge Vocabularies by Using Optimized Sub-word Units", Proc. EUROSPEECH, (2001).
2. K. Kumatani, T. Gehrig, U. Mayer, E. Stoimenov, J. McDonough, and M. Wölfel, "Adaptive Beamforming with a Minimum Mutual Information Criterion", to appear *IEEE Transactions on Audio, Speech, and Language Processing*.
3. J. McDonough, K. Kumatani, T. Gehrig, E. Stoimenov, U. Mayer, S. Schacht, M. Wölfel, and D. Klakow, "To Separate Speech!: A System for Recognizing Simultaneous Speech", in *Proceedings on Machine Learning and Multimodal Interaction, 2007*.

IDIAP Research Institute (IDIAP): The IDIAP Research Institute (www.idiap.ch) is an independent, not-for-profit, research institute located in Martigny, Switzerland, and affiliated with the Swiss Federal Institute

of Technology at Lausanne (EPFL), and the University of Geneva. With a research staff of more than 75 scientists (including EPFL professors, senior scientists, postdoctoral researchers, PhD students and developers), the primary missions of IDIAP are research, education, and technology transfer in the areas machine learning, speech and audio processing, computer vision, information retrieval, biometric authentication, multimodal interaction, and multiple multimodal research activities across these disciplines. IDIAP is involved in numerous national and international (EU and US) projects, as well as in multiple collaborative projects with the industry.

Professors Bourlard and Hermansky are leading researchers in speech processing and will bring their experience to SCALE. Researchers working in the framework of SCALE will benefit from the ability to interact within the larger community of researchers at IDIAP. In speech processing alone there are eight post-doctoral and senior RTD personal as well as many PhD students working in related topics. New IDIAP collaborators will be allocated personal workstations/laptops which provides access to IDIAP's centralised computing facilities. In addition, IDIAP is equipped with cluster and distributed computing facilities totalling around 170 cores, plus centralised storage.

Key personnel: Prof Hervé Bourlard (7.5%) and Prof Hynek Hermansky (7.5%)

Key publications/patents:

1. G. Aradilla, J. Vepa, and H. Bourlard, "Using Posterior-Based Features in Template Matching for Speech Recognition", in International Conference on Spoken Language Processing, 2006.
2. H. K. Maganti and D. Gatica-Perez, "Speaker Localization for Microphone Array-Based ASR: The Effects of Accuracy on Overlapping Speech", in Proc. Int. Conf. on Multimodal Interfaces (ICMI), 2006.
3. J. Pinto, A. Lovitt, and H. Hermansky, "Exploiting Phoneme Similarities in Hybrid HMM-ANN Keyword Spotting", in Proceedings of Interspeech, 2007.

Motorola European Research Laboratory (Motorola): The Motorola European Research Laboratory is part of the Motorola Labs, the central research facility for Motorola, conducting industrial research relevant to Motorola's future business, primarily in the areas of mobile communications, automotive and broadband. The team, located in Basingstoke, UK, conducts research in the areas of speech coding, speech and multimodal interfaces, noise robust speech recognition, and microphone arrays for improved acoustic interfaces. In particular, Motorola has led the process of development of international standards for distributed speech recognition (ETSI Aurora and 3GPP) and the associated databases and evaluation criteria widely used in the speech research community for robustness evaluations.

Key personnel: Dr David Pearce (7.5%)

Key publications/patents:

1. D. Pearce, J. Ferrans, J. Engelsma, J. Johnson, "An Architecture for Seamless Access to Distributed Multimodal Services", Interspeech 2005, Lisbon, 4-8 Sept 2005
2. D. Pearce, "Robustness to Transmission Channel - the DSR Approach," Keynote paper, COST278 & ICSA Research Workshop on Robustness Issues in Conversational Interaction, Aug 2004.
3. D. Macho, L. Mauuary, Bernhard Noé, Y.M. Cheng, D. Ealey, D. Jouvet, H. Kelleher, D. Pearce, F. Saadoun, "Evauation of a Noise-Robust DSR Front-end on Aurora Databases, ICSLP 2002, Sept 2002

Philips Speech Recognition Systems (Philips): Philips Speech Recognition Systems develops and markets speech recognition technology for professional users. Integrated into medical, legal and financial IT systems, its technology is used to increase efficiency in institutions with a large dictation volume. Philips Speech Recognition Systems (PSRS) became one of the first company to release a commercially available continuous speech recognition engine that allowed the spoken word to be converted to text on a PC. As early as 1994 Philips launched the first client/server-based continuous speech recognition system for the radiology market. Today, PSRS as the world's largest supplier of dictation devices, is a leading provider of speech recognition

technology for professional applications. More than 20 recognition languages, a portfolio of 150 ConTexts and over 8000 installations of SpeechMagic in hospitals, practices and law firms bear testimony to the leading position of PSRS in professional dictation and speech recognition markets.

The PSRS technology development team is continuously improving the recognition systems at the cutting edge of technology. The team is composed of experts in speech technologies from the fields of natural language processing, acoustics, signal processing, and machine learning. Research and development is carried out in close contact with academic and industry partners, for example in the Austrian Competence Network for Advanced Speech Technologies COAST (www.coast.at).

Key personnel: Dr Heinrich Bartosik (5%), Dr Erhard Rank (5%), Dr Walter Müller (5%)

Key publications/patents:

1. J. Peters, Ch. Drexel, "Transformation-Based Error Correction for Speech-to-Text Systems," in Proc. ICSLP, Jeju Island, Korea, pp. 1449–1452, 2004.
2. H. Bartosik, W. Müller, and M. Schatz, Adaptation of a speech recognizer from corrected text, Patent WO 01/04874 A1 = US 6.725.194 (ET: 20.04.2004) = EP 1.110.204 (OT; 27.06.2001)
3. H. Bartosik and K. Rajic, Speech recognition device to mark parts of a recognized text, Patent WO 03/034404 A1 = EP 1.438.710 (OT: 21.07.2004)

Radboud University Nijmegen (RUN): The Department of Language and Speech at RU Nijmegen, led by Prof L. Boves, has an outstanding track record in research and teaching in the area of automatic speech recognition, speaker recognition and language technology. Thanks to close collaboration with the Max Planck Institute for Psycholinguistics on the university campus RU Nijmegen also has a strong focus on commonalities and differences between human and automatic speech processing and computational modelling of human speech processing. Specific topics in automatic speech recognition include pronunciation modelling, noise robustness, phone decoding, and automatic phonetic transcription. In collaboration with the Nijmegen Institute for Cognition and Information (NICI) the Dept. of Language and Speech has been conducting research on cognitive aspects of multimodal dialogue systems. The on-campus F.C. Donders Institute for cognitive brain imaging offers access to advanced measurement facilities for fundamental research in human speech processing. The Dept. of Language and Speech has a formal link with the T.N.O. Human Factors Institute in Soesterberg.
The Dept. of Language and Speech has been involved in numerous national, international and EU-funded research projects. Currently, Lou Boves is managing the FP6-FET project ACORNS.

Key personnel: Prof. Lou Boves (7.5%), Dr Louis ten Bosch (7.5%), Dr-ir Bert Cranen (7.5%), and Dr O. Scharenborg (7.5%)

Key publications/patents:

1. Scharenborg, O., ten Bosch, L., Boves, L., Norris, D. (2003) "Bridging automatic speech recognition and psycholinguistics: Extending Shortlist to an end-to-end model of human speech recognition", *Journal of the Acoustical Society of America*, 114 (6), 3032-3035.
2. de Wet, F., Weber, K., Boves, L., Cranen, B., Bengio, S. en Bourlard, H., (2004). "Evaluation of formant-like features for automatic speech recognition", *Journal of the Acoustical Society of America*, Vol. 116, No. 3, pp. 1781-1792.
3. Scharenborg, O., Seneff, S., Boves, L. (2007) "A two-pass approach for handling out-of-vocabulary words in a large vocabulary recognition task", *Computer Speech and Language*, 21 (1), 206-218.

RWTH Aachen University (RWTH) RWTH Aachen University, Lehrstuhl fuer Informatik 6 is concerned with research and development of advanced stochastic and automatic learning based algorithms for speech and pattern recognition and for language processing. The chair, Prof. Dr.-Ing. Hermann Ney, has had a proven record of active research in this area over many years, both at Philips Research until 1993 and, since then, at

RWTH Aachen University. His scientific work focuses on statistical techniques for decision-making in context and statistical pattern recognition. Prof. Ney's interests cover many aspects of pattern recognition, such as signal processing, search strategies, image recognition, sign language and gesture recognition, speech recognition, natural language understanding and translation of both written and spoken language. He has been and continues to be involved in a number of successful national or European projects, such as the German project Verbmobil and the European projects CORETEX, EUTRANS, TRANSTYPE 2, LC-STAR, PF-STAR, and TC-STAR. The scientific achievements have been documented in a large number of contributions to international journals and conferences.

Key personnel: Prof. Dr.-Ing. Hermann Ney (7.5%), and Dr. Ralf Schlüter (7.5%).

Key publications/patents:

1. M. Bisani and H. Ney: Open Vocabulary Speech Recognition with Flat Hybrid Models. Proceedings of the European Conference on Speech Communication and Technology, Interspeech, pp. 725-728, Lisbon, Portugal, September 2005.
2. B. Hoffmeister, T. Klein, R. Schlüter, and H. Ney: Frame Based System Combination and a Comparison with Weighted ROVER and CNC. In Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP), pp. 537-540, Pittsburgh, PA, September 2006.
3. R. Schlüter, W. Macherey, B. Müller, H. Ney. "Comparison of Discriminative Training Criteria and Optimization Methods for Speech Recognition". In Speech Communication. Vol. 34, pp. 287-310, May 2001.

University of Edinburgh (UEDIN): The Centre for Speech Technology Research (CSTR) at UEDIN is an interdisciplinary research centre spanning the schools of Informatics and Linguistics. CSTR currently has about 40 researchers with expertise in speech recognition, speech synthesis, machine learning, acoustic phonetics, speech perception and multimodal interaction.

UEDIN has profound experience in speech synthesis which includes the ongoing development of the open source FESTIVAL system, based on unit selection, as well as expertise in trajectory-HMM speech synthesis. It also brings extensive speech recognition know-how in both acoustic and language modelling, the development of large-scale systems, and microphone array processing.

Within Edinburgh, there is close collaboration with the Human Communication Research Centre, the Institute for Adaptive and Neural Computation, and the Speech Science Research Centre at Queen Margaret University. CSTR coordinates the FP6 Integrated Projects AMI and AMIDA (jointly with IDIAP), the Marie Curie EST Host Fellowship project Edinburgh Speech Science and Technology (EdSST), and the Festival open source speech synthesis system.

Key personnel: Prof Steve Renals (7.5%), Dr Simon King (7.5%), Dr Korin Richmond (7.5%), Dr Joe Frankel (7.5%)

Key publications/patents:

1. A. Dielmann and S. Renals (2007). "Automatic meeting segmentation using dynamic Bayesian networks". *IEEE Trans. on Multimedia*, 9(1):25-36.
2. S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester (2007). "Speech production knowledge in automatic speech recognition". *Journal of the Acoustical Society of America*, 121(2):723-742.
3. J. Cabral, S. Renals, K. Richmond, and J. Yamagishi (2007). "Towards an improved modeling of the glottal source in statistical parametric speech synthesis". In *Proc.of the 6th ISCA Workshop on Speech Synthesis*.

University of Sheffield (USFD): The Speech and Hearing group (SPandH) at the Dept of Computer Science, USFD is one of the most prominent in this field worldwide, with seven faculty, and about 30 active researchers

in total. Facilities include a speech perception lab as well as state of the art computing facilities, such as Grid computing. SPandH has established a reputation for work which crosses the boundaries between the communities involved in speech research, a feature which is central to this proposal. It has been instrumental in bridging the gap between speech science, hearing and speech technology and brings this experience to the consortium. In particular, SPandH has pioneered the glimpsing theory of human speech perception in noise and its engineering counterpart, missing data techniques for robust ASR. Recent work on episodic models links ASR system building to models of HSR. SPandH works closely with the Machine Learning (ML) group in the same Department. SPandH coordinates the training programme for the FP6 IP AMI and coordinated the FP5 RTN HOARSE.

Prof. Roger K. Moore is chair of Spoken Language Processing in SPandH group. He has over 30 years experience in speech technology R&D and much of his work has been inspired by insights derived from human speech perception and production. Prof. Moore is currently developing a unified theory of spoken language processing that combines accounts from a wide variety of different disciplines concerned with the behaviour of living systems - many of them outside the normal realms of spoken language - and compiles them into a new framework in the general area of 'Cognitive Informatics' called 'PRESENCE' (PREdictive SENsorimotor Control and Emulation).

Prof Phil Green founded the SPandH group and has made significant contributions to robust speech processing by humans and by machines, for instance the 'missing data' paradigm. He is an experienced coordinator of EC projects, including the training networks SPHERE (FP3), SPHEAR (FP4) and HOARSE (FP5). He currently coordinates the training programme of the FP6 Integrated Project AMIDA.

Thomas Hain is a lecturer in SPandH group at USFD and a world-leader in Automatic Speech Recognition, heading the ASR team in the AMI-AMIDA Integrated Projects. The AMI recognition system was shown to yield the best performance on meeting transcription in the most recent international competition (NIST evaluations) for such systems. He is also a member of the IEEE Speech Technical Committee.

Guido Sanguinetti is a lecturer in the Machine Learning (ML) group. His scientific expertise lies in the development and application of approximate inference techniques to non-linear, non-Gaussian problems.

Key personnel: Prof Roger Moore (7.5%), Prof Phil Green (7.5%), Dr Thomas Hain (SPandH) (7.5%), Dr Guido Sanguinetti (ML) (7.5%)

Key publications/patents:

1. Scharenborg O., Wan V and Moore R K. "Towards capturing fine phonetic variation in speech using articulatory features", J. Speech Communication, Special Issue on Speech Recognition and Intrinsic Variation, Vol.49, pp.811-826, (2007).
2. Ma, N., Green, P., Barker, J. and Coy, A. (2007) "Exploiting correlogram structure for robust speech recognition with multiple speech sources", Speech Communication, in press, available from http://www.sciencedirect.com/science/journal/01676393
3. Hain, T et al (2007) "The AMI System for Transciption of Speech in meetings", PROC IEEE ICASSP 2007

Eurice GmbH (EURICE)

Eurice GmbH (EURICE) is a company that offers comprehensive and specialized support all around international research projects. Most importantly, EURICE works as a training partner of the European Commission and, in co-operation with European experts and the Commission, has developed a modularized system and related training contents to adequately prepare scientists for mastering the challenges of working in international (in particular European) research projects. EURICE also is a member of EARMA, the leading association of research managers and administrators across Europe, which sets the highest standards for research, management, and administration. Currently, EURICE is managing and providing its services in 27 EC-funded research projects worldwide (which, amongst others, include the IPR Helpdesk).

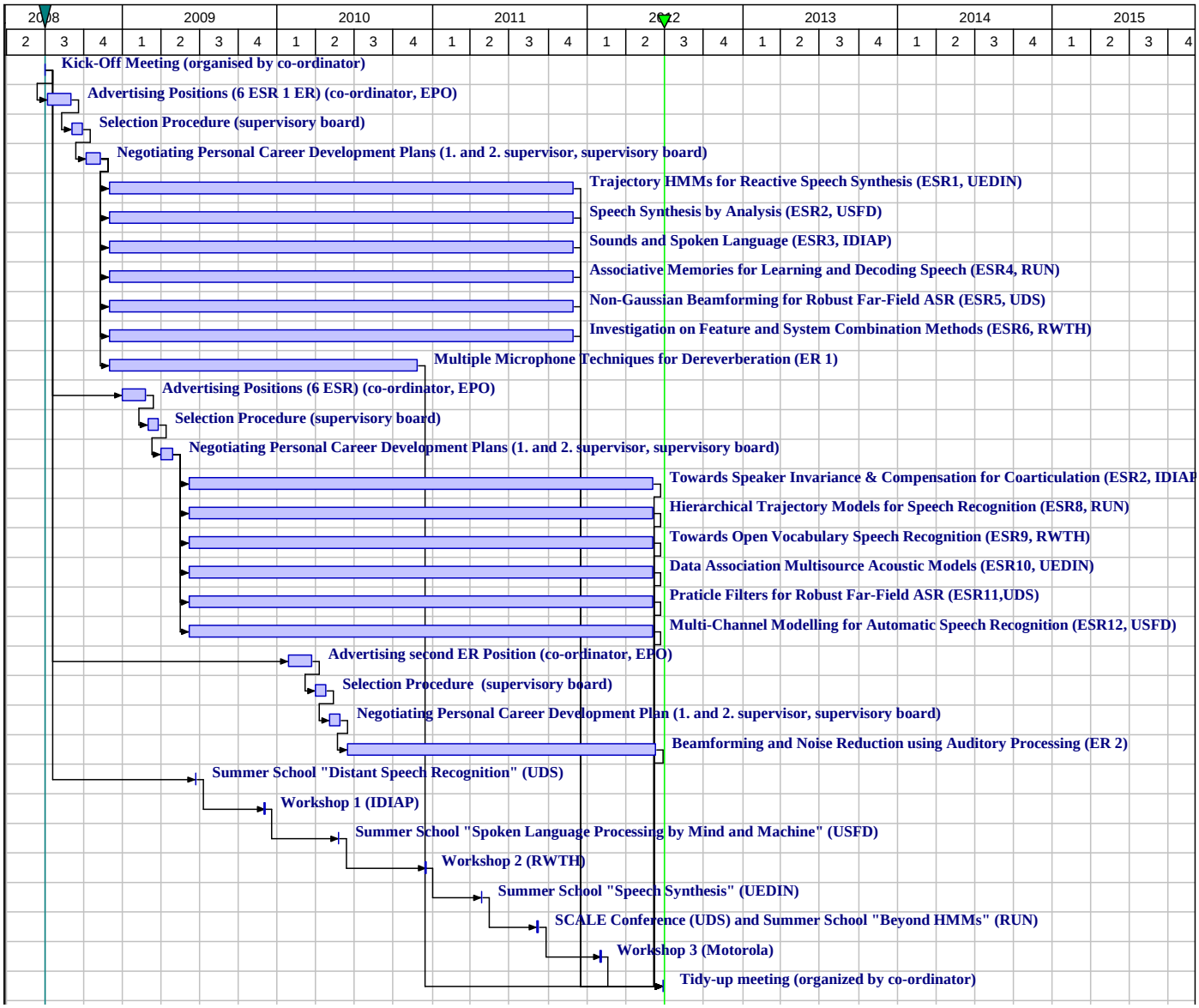


Figure 2: Overall Workplan for SCALE

Mr. Jörg Scherer, the Managing Director of EURICE, will be in charge of the complementary training provided within SCALE. He has been working as a research manager in both the academic and industrial sector for the past ten years and has an outstanding track record in the coordination, management and administration of European RTD projects, spanning the 4th-7th EU Framework Programmes. His portfolio comprises more than 100 grant applications and the management of more than 40 European research projects. Not only does he act as evaluator for the European Commission in different EU funding programmes in the field of innovation, research and education, but he also is a member of the IPR Helpdesk, consultant for the Saarland Ministry of Economy for the setup and implementation of a Regional Innovation Strategy. In 2006 alone, he provided complementary skills training in more than 50 training events and seminars.

Key personnel: Mr. Jörg Scherer (2%)

5.2 Overall Workplan

Fig. 2 provides and overview of the overall workplan for SCALE illustrating the major administrative tasks like recruiting the fellows, the planned research training as well as the planned schedule for workshops, summer schools and the SCALE conference. For illustration purposes it assumes a starting date of 1.7.2007 which, however, can easily be adjusted according to the given circumstances. Behind each task, the people or groups responsible for performing these tasks are named. While Fig. 2 represents the overall workplan, milestones, and deliverables of SCALE, the progress of each individual research project will be monitored in terms of the technical milestones presented in Section 4.3.

Overall, SCALE aims to achieve the following goals which at the same time represent quality indicators against which the successfulness of SCALE can be measured:

- Successfull negotation and implementation of all 14 CDPs.
- On average two reviewed conference papers (e.g. ICASSP or INTERSPEECH) and one journal paper (e.g. IEEE Transactions on Audio, Speech and Language Processing) per ESR/ER.
- A total of 10 invention disclosures (i.e. approximately one for every six months of industrial secondment).
- 70% of the ESRs will have completed their PhD at the end of SCALE.
- All of the ESRs will find a position within 2 months after completion of their PhD either in industry or in academia.

5.3 The Role of Industry and External Partners

Industry plays a major role in the SCALE training network. First of all, industry provides vital training opportunities for ESRs and ERs in form of secondments. As indicated above (see section 4), the vast majority of the SCALE fellows are expected to take their secondments in the industrial laboratories of Motorola and/or Philips, thus getting first-hand practical experience of industrial research. Second, industry will substantially be involved in the supervisory board. As indicated earlier (see above, section 3), much of the potential market impact currently is untapped because existing technology currently lacks flexibility and adaptability. As such, the involvement of partners from industry in the supervisory board (Motorola, Philips, German Telekom) ensures that training is as comprehensive and adequate to the existing industrial needs as possible. Third, as part of the network-wide training activities, external partners from industry (e.g. German Telekom) will act as keynote or guest speakers, again bridging the gap between industry and academia. Motorola will organize a workshop. The associated SME EURICE will provide the complementary skills training on the network level, which is indispensable to adequately prepare fellows for their future career in both industry and academia, e.g. as assistant professors, lecturers, project managers, etc. Last but not least, industry certainly represents a potential user of the scientific developments emanating from the network's activities. For both Motorola and Philips, exploitable knowledge (such as patents) is expected particularly in the area of research theme 2 (Bridging the Gap between ASR and HSR) and 3 (Bridging the Gap between Signal Processing and Learning).

The external partners which will be included in the supervisory board will not only strengthen the role of industry in this ITN, but also bring in outstanding expertise from academia as well as from the International Speech Communication Association (ISCA). Namely, the following three external members are included:

- Dr. Jin Liu (Jin.Liu@t-systems.com) from the German Telekom (http://www.telekom.de). Dr. Liu holds a senior position in the Advanced Voice Solutions Department for German Telekom. She coordinated German Telekoms participation in large projects like SmartWeb and Compass 2008. Through the participation of Dr. Liu, the German Telekom will contribute its invaluable expertise as a telekom operator.

	Philips	IDAP	Motorola	RUN	RWTH	UdS	USFD
UEDIN	•	•	•	•		•	•
	Philips	•		•	•	•	•
		IDIAP	•		•	•	•
			Motorola				•
				RUN	•		•
					RWTH	•	
						UdS	

Table 6: Existing collaborations within the SCALE consortium.

- Prof. Isabel Trancoso, the president of International Speech Communication Association (ISCA) (http://www.isca-speech.org/). In her position as ISCA president, Prof. Isabel Trancoso is brings in the overall perspective of the international speech community. ISCA is the organiser of Interspeech, the worlds largest conference dedicated exclusivly to speech processing with typically 1500 atendees per year coming both from industry and academia. As such, the participation of ISCA through Prof. Trancoso will contribute to the increased reach of SCALE beyond the network.
- Prof. Sadaoki Furui from Tokyo Instiute of Technology (http://www.furui.cs.titech.ac.jp/). Prof Furui is one of the best known researchers in speech technology. From 1991 to 1997 he was head of the NTT Human Interface Laboratories. Presently he is Head of the Department of Computer Science of Tokyo Institute of Technology. As such, he will not only contribute his excellent scientific knowledge, but also ensure that the research training within SCALE is on a top-international level.

5.4 Complementarities and Exploitation of Synergies

The research teams within SCALE complement each other very well as they bring expertise in automatic speech recognition, human speech recognition, text-to-speech synthesis, and machine learning (see Section 1 above for details). Moreover, there are many existing bilateral and trilateral collaborations between the partners which will be exploited for the benefit of the ESRs and ERs. These are summarized in Table 6 below.

Major existing collaborations include:

- USFD, IDIAP and UEDIN have a long history of fruitful collaboration, through various EU projects, including the FP6 integrated project AMI, which is jointly co-ordinated by UEDIN and IDIAP, and whose training programme is co-ordinated by USFD.
- IDIAP and USFD have collaborated in two previous ITNs, SPHEAR (FP4) and HOARSE (FP5).
- Mobility of researchers has resulted in collaborative links between RUN and UEDIN, RUN and USFD, RWTH and UdS, and UEDIN and IDIAP.
- RWTH, UdS and Philips have long standing links dating back to the mid nineties (e.g. Verbmobil).
- RUN has collaborated with Philips in Dutch national research programmes and in FP4 projects MAIS and ARISE. More recently, RU Nijmegen has collected and annotated speech corpora for Philips Research in Aachen. RUN has collaborated with USFD and UEFIN in the FP5 project COMIC and in the FP6 RTN S2S. Currently, RUN and USFD are partners in the FP6-FET project ACORNS.
- Motorola sponsored PhD students at USFD and Philips at RWTH.

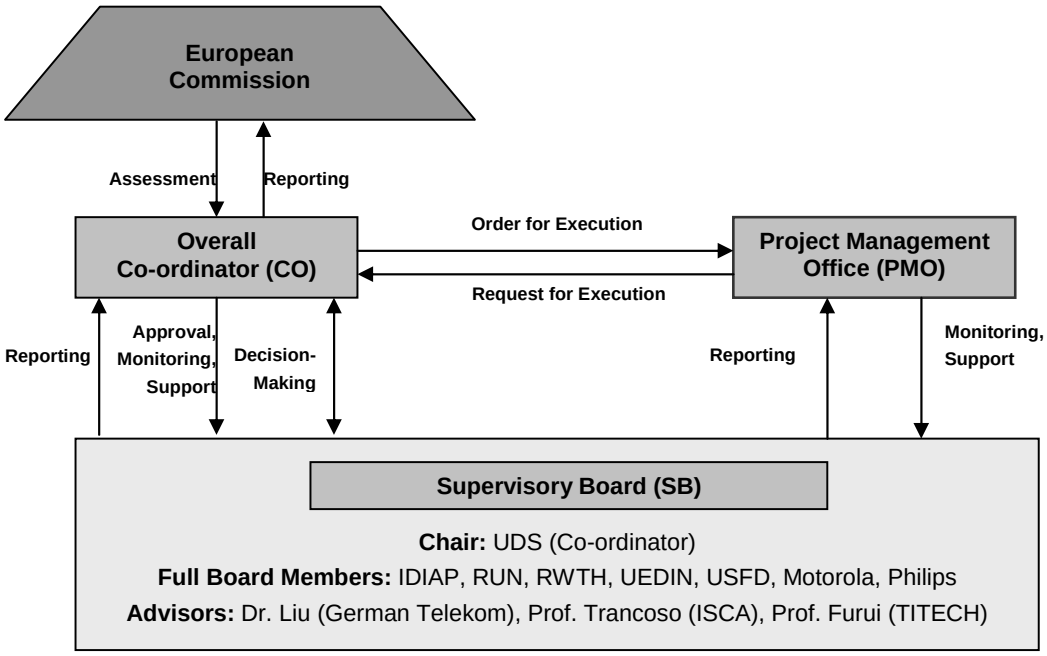


Figure 3: Proposed management structure of the SCALE project.

- Philips is like USFD, IDIAP and UEDIN part of AMI and AMIDA.

As such, the different network partners bring in different expertise. For example is UEDIN well known for its work in synthesis, RWTH has used statistical modeling techniques with extreme success and UdS as demonstrated its competence in Far Field Speech recognition in this year's speech separation challenge. RUN focuses on fundamental research in the relation between human and machine speech processing. These different areas of expertise are also reflected in the training program and greatly contribute to its multi-disciplinarity. Similar to the academic partners involved in SCALE, the two industrial partners have different profiles. Motorola uses speech in mobile devices whereas Philips dictation software for the professional market is the dominating product in Europe. These different profiles also require specific types or variants of new algorithms and methods to be researched during the secondments. These complementarities are further exploited by the inclusion of outstanding external partners from industry (Dr. Liu, German Telekom), the International Speech Communication Society (Prof. Trancoso), and academia (Prof. Furui, Tokyo Institute of Technology), in the supervisory board (see above Section 5.3).

5.5 Overall Management of the Research Training

The purpose of project management is to ensure the successful coordination and management of the different project activities. Within SCALE project management will comprise the scientific, administrative, and financial management of the ITN including the management of communication, knowledge, IPR, dissemination and exploitation activities, and recruitment. Taking into account the relatively small size of the project, the proposed management structure will be as simple as possible, but as comprehensive and specialized as necessary to successfully meet the requirements of the different aspects of project management. Figure 3 illustrates the management structure within SCALE. As shown in Figure 3, the proposed management structure includes the overall scientific co-ordinator (CO), the project management office (PMO), and the supervisory board (SB).

5.5.1 Overall Scientific Co-ordinator (CO)

The main task of the scientific co-ordinator (CO) is the overall monitoring and co-ordination of the project. In particular, the overall scientific co-ordinator will monitor the implementation of the training and research activities and be responsible for the management of the results. He will also assume overall responsibility for the assessment and reporting activities. He will act as the single point of contact between the participants and the Commission. Moreover, he will be the central contact point for members of the supervisory board in terms of scientific and training-related questions. He will closely collaborate with the project management office (PMO) as to ensure efficient and overall co-ordination and management of all remaining project-relevant questions.

In SCALE, the function of the scientific coordinator is carried out by UdS (namely Prof. Klakow). Prior to his appointment at UdS, Prof. Klakow was managing projects worth 1.4 million Euros per year. At UdS he currently is managing the SmartWeb subproject on "On- and Off-line Extraction of Information" (1.4 million Euros) and PIRE (0.8 million Euros). PIRE is an international research training program with Johns Hopkins University, Brown University and Charles University Prague). Prof. Klakow is also participating in the DFG-Internation research training program with Edinburgh University. He has more than 50 papers and eighteen patents.

5.5.2 Project Management Office (PMO)

In order to reduce the administrative and organisational burden for the scientific co-ordinator and the partners who will be heavily involved in the training and research activities, the project management office (PMO) will be in charge of the management of all administrative, financial, legal aspects of the project. The project management office thus is the contact point for all administrative, financial, and legal questions from both the scientific co-ordinator and members of the supervisory board. In SCALE, the function of the project management office (PMO) will be carried out by EPO, the European Project Office. EPO is part of the Saarland University administration which has specialized in the management and conduct of European research projects. In the 8 years of its existence, EPO has managed and is managing various projects under FP4, FP5, FP6 and FP7 with different scientific and non-scientific backgrounds including information and communication technologies, new material sciences, nanotechnologies, medicine, law, biology. As such, EPO has also been in charge of research training networks (e.g. Calculemus, NEURAD, Galenos) and EuroConferences. As indicated above (see Section 5.1), the responsible project officer for SCALE, Ms. Corinna Hahn, has extensive experience in the conduct and management of both European research and training projects.

5.5.3 Supervisory Board (SB)

The supervisory board will be chaired by the UdS (namely: the overall scientific co-ordinator) and consist of (i) full board members (i.e. one representative - the person listed as 'person in charge' in Section 1 - from each network participant and from the associated partners Philips and Motorola) and (ii) three advisors from outside the network (Figure 3). Both chair and full board members will have one vote with the chair having the casting vote. The advisors will, with their expertise, actively support the full board members and provide advice and guidance in the decision-making process. The differentiation between full board members and advisors was deemed necessary so as to keep the decision power within the network but to also be able to take into account invaluable external expertise from industry, professional organisations, and academia. The inclusion of the associated partners Philips and Motorola as full board members was necessary to acknowledge their importance as industrial partners in the field of speech processing and their active involvement in the training activities (e.g. as providers of secondments and second supervisors of the fellows). Overall, the involvement of associated industrial partners and several international experts from outside the network aims to ensure that the training is not only relevant to the professional community and top quality from a scientific and academic

point of view, but also satisfies existing needs and requirements of industry, thus giving the recruited ESRs and ERs adequate preparation for a career in both sectors.

Starting with the kick-off meeting in month 1, the supervisory board will meet approx. every 6 months (at the summer schools, workshops, and the SCALE conference) to monitor and evaluate the progress of the research, training, dissemination and exploitation activities as well as plan the further activities in these areas. As such, the SB will serve as the forum (i) for making strategic decisions concerning the overall training plan, research, technical objectives and project management, dissemination, exploitation, and IPR, (ii) for resolving administrative and organisational issues, and (iii) for the continued sharing and dissemination of best practice amongst the partners. Also, the supervisory board will take responsibility for the recruitment of the fellows (see subsection 5.6 for details). While on a day-to-day basis, each full member of the supervisory board will be the central contact and responsible person for all project-related activities within his/her institution, the supervisory board as a whole will be responsible for the co-ordination of the project from a strategic point of view.

5.5.4 Communication and Networking

Within SCALE, all partners will be kept fully informed about the status of the research and training activities, the planning and any other issues which are important to the partners in order to carry out their tasks and to maintain and increase the synergies of their collaboration. As indicated above (see description of CO and PMO), the overall scientific co-ordinator (namely Prof. Klakow) and the project management office (namely C. Hahn) act as central contacts and, within their assigned competences, ensure that partners are fully informed about all developments within the project at any time. Day-to-day communication as well as the distribution of interim results will primarily be carried out by email and via the project website. The project website will be organized by UdS and include both a public (which will be open to the public) and a restricted area (only available to the partners). The public area will provide general information about the project, a list of vacancies within the ITN and their description, a calendar plus description of upcoming events (workshops, summer schools, conferences) and any materials (e.g. research reports) appropriate for dissemination. The restricted area will be password-protected and only be accessible to the network partners. It will provide access to all project-related reports and materials (progress-, activity-, management reports and related information).

5.5.5 Decision-making and Conflict Resolution

The detailed rules for decision making will be laid down in the consortium agreement at the beginning of the project. Decisions will be made at the regular meetings of the supervisory board (SB) as the need arises. Significant decisions will be announced in the meeting agenda and all partner institutions need to be represented at these meetings. The overall scientific co-ordinator will chair the decision-making process and ensure that all decisions and the corresponding discussions are written down and saved for later reference. Each full board member will have one vote. Depending on the scope and importance of the decision to be taken, the votes will need to satisfy the requirements for either simple majority or unanimity. Significant decisions will require unanimity. Led by both the overall scientific co-ordinator the supervisory board will thus also be in charge of the resolution of conflicts that may arise during the course of the project.

Should a partner not fulfil his/her task within the project, he/she will formally be cautioned by the supervisory board. Where necessary, possible solutions will be elaborated through discussion among all partners concerned. The overall scientific co-ordinator will chair the discussions and set the schedule for solution-finding. In the case of critical deviations from the work plan, the European Commission will be informed and consulted by the overall scientific co-ordinator.

5.5.6 Monitoring, Reporting & Risk Situations

While it is the supervisors' task to control research and training progress of their fellows (see Section 4.7.1 on a local level, it is the scientific co-ordinator's responsibility to *monitor and control* overall project progress by contacting the persons-in-charge of the research training at each partner on a 3-monthly basis to check the progress made at each institution. In turn, within their own institution the persons-in-charge are required to (1) monitor the status of research, training, and financials of their respective work and (2) to inform the scientific co-ordinator and the project management office regularly on the status quo of the research and training activities carried out at their institution - according to the assigned competences of these two entities. Most importantly, each partner has the responsibility to *report immediately* to the overall scientific co-ordinator if any risk situations emerge that may conflict with SCALE's objectives or the successful completion of the assigned research and training tasks. As such, critical issues will become apparent very quickly in the context of the day-to-day communication between the partners.

Risk situations that may emerge within SCALE include changes in scheduling of deliverables and/or allocated funding, the loss of a fellow (e.g. because of maternity, head-hunting from competitors), the loss of a supervisor, or the loss of a partner. Loss of a supervisor will be compensated swiftly by the co-supervisor and by appointing adequate replacement. Loss of a partner will be compensated by the network partners and/or the intensification of the existing close relationship with suitable associated partners. Loss of a fellow, if early in her/his research, will be compensated by recruitment of a new fellow. In any case, as indicated above (see Decision making and conflict resolution), the scientific coordinator supported by EPO will direct the solution-finding process.

In addition to these monitoring and reporting activities, the *regular meetings* of the supervisory board present another form of monitoring progress. Starting with the kick-off meeting, they will be hosted at the different partners' sites in an approximate interval of six months. To maximize the efficiency of the network and minimize costs, these meetings will be held in the course of the summer schools 2. Additional meetings and/or video/telephone conferences will be held when the need arises, e.g. to co-ordinate the preparation of reports or to discuss any critical issues that may emerge in the course of the project.

5.5.7 Management of Knowledge and IPR

Particular attention will be paid to the management of knowledge and IPR, i.e. the rights of the parties involved to utilise the knowledge and inventions that may be made during such collaboration. This is particularly important in a project with industrial partners. Knowledge management will therefore address the description of background (i.e. pre-existing knowledge), the identification of scientific results and foreground (i.e. knowledge generated within SCALE), as well as a description of potential means of dissemination and exploitation (see below Dissemination and Exploitation). The details for the management of knowledge, IPR, and dissemination and exploitation-related activities will be provided in terms of specific rules in the consortium agreement.

At the beginning of the project, all partners will be required to declare their background and to specify if they intend to invoke the exclusion of any part of it from the obligation to grant access rights to the other partners. As for the foreground, not only will the scientific co-ordinator require all partners to identify the foreground that will be generated in the course of the project, but also make all partners aware of the importance of IPR issues. Generally, reports issued during the project will be treated as confidential as there is the possibility of securing patents on the results.

The proposed basic concept concerning the ownership of results and IPR for SCALE is that the partner institution generating the results and intellectual property owns these items and is responsible for their legal protection and transfer. Any intellectual property created by a visiting fellow during a period of industrial secondment will be owned by the host industrial partner. Results jointly developed shall be jointly owned, but any of the owners will be free to individually exploit the joint IPR. All partners involved in SCALE shall

have the right: (i) to express their interest in the protection of any results from SCALE; (ii) to protect results after negotiations with parties directly involved with the results; and (iii) to negotiate exploitation rights on the results and IPR coming out of SCALE even if they are not directly involved in the associated research efforts. The effective transfer of rights shall be done in specific agreements (licenses or transfer of ownership) to be negotiated between the parties directly involved.

If the use of background of a partner is necessary to exploit the foreground generated in SCALE, the owner of such background, if free to do so (i.e. having no previous incompatible commitments), shall negotiate in good faith, at favourable conditions, access rights to allow the use of the background. Generally, all partners shall have the right to publish scientific results deriving from SCALE. If these results involve joint IPR, then all concerned partners are: (i) allowed to review communications prior to publication and (ii) given sufficient provision to protect specific results within a prior agreed period before publication.

5.5.8 Dissemination and Exploitation

An initial plan for the dissemination of project results will be prepared during the first 4 months of SCALE, and will be updated throughout the lifetime of the project (including the review of IPR issues). Within SCALE, the project results will be disseminated and exploited via the following routes:

Project Website: The project website, which will be online in Month 2 of the project, will be the main public face of the project. It will provide a portal to project research (and also to research training opportunities), including publications, online and indexed multimodal recordings of workshops, state-of-the-art reviews, links to participant labs and researchers, and links to related research and projects.

Peer-reviewed papers: SCALE partners all have strong publication records in the major journals, conferences and workshops in the field, and this will continue and be strengthened. Within SCALE, we shall put particular emphasis on collaborative research, and hence papers jointly authored across partners, in particular industry-academia collaborations. As indicated above (see Section 5.2), each fellow is expected to produce at least one conference and a journal paper.

Special sessions and issues: A way to maximize the impact of project outputs is through the organization of coordinated routes of dissemination by the supervisory board (e.g. rather than through journals paper through the organization of special issues, or, in terms of training events special sessions at conferences and workshops). This also helps to establish a project ”brand”.

Summer schools: SCALE summer schools, featuring courses from scientists in SCALE (including ERs) are an excellent way of disseminating project results and methodologies to fellows (in particular) from both inside and outside the network.

SCALE book: Experience from past projects has shown that summarizing the results of a large project in a book is an very efficient way of dissemination. In the project ”Verbmobil” (co-ordinated by DFKI/UDS) the corresponding book written by participants was cited more than 200 times, which is by far more than any paper written in the course of this project. A book arising from SCALE, perhaps containing some reviews of the state-of-the-art of SCALE research areas, would provide a very condensed view of the results of the project but will also show how the different topics covered in the ITN are connected, and how the different gaps addressed in the ITN are bridged which is one of the objectives of the ITN. Such a book has the potential to become a standard reference and will certainly have a large impact on the unification of the existing fragmented specialized areas. A first draft version based on the papers published and 6-monthly research report written will be delivered at the end of the project.

Links with ongoing projects: All partners of the SCALE consortium are at the same time members of one or more other high profile projects including European IPs and other large scale projects funded directly by an EU member state. Results of the ITN can thus be easily transferred into and out of those projects. Very often researchers working on related topics at a specific site closely cooperate and in this way distribute and exploit the results of their own projects. A specific way of dissemination and exploitation is the transfer of ESRs who have completed their PhD in SCALE to continue as an ER in a different project. We can also imagine joint

publications on specific topics together with members of other projects.
Corpora and evaluation: In the course of the project we will create corpora as part of our research activity (data collection in Month 15 and Month 27). Corpora help to focus the community on important tasks (e.g. AURORA for noise robust distributed speech recognition). Increasing network bandwidth makes it feasible to distribute corpora across the web (this strategy has been adopted recently by some large projects (e.g. AMI)). Together with the corpora we will provide evaluation schemes and baseline results. Those baselines will be published in suitable papers and the tools necessary for automatic evaluation and reproduction of the baselines will be distributed together with the corpora. The existence of such evaluation protocols and baseline results makes a dramatic difference in the take-up and use of released corpora. In specific cases (depending on the effort and the expected impact on the community) we may organize a competitive evaluation in the style of DARPA benchmarks, or work closely with the development of existing evaluation programmes.
Technology transfer: A very important way to exploit the project results and at the same time to measure the success of the ITN is technology transfer. Therefore, we will identify to which extent practical applications can be derived from the results that will create value for the European ICT industry. On the one hand, the results can directly be exploited by the participating industry partners, Motorola and Philips, in two ways: First, they can directly use their own results. Second, the results developed by the partners can be transferred into these companies on the basis of (i) publications, patents, software generated within SCALE and (ii) ESRs and ERs trained in the network who may later be employed by those companies. On the other hand, companies outside the consortium too will benefit from the technologies developed in SCALE, e.g. by means of papers and patents (through licensing). We expect that a significant fraction of the fellows trained in the network will later be employed at other companies and transfer technology this way. As some of the academic partners in SCALE have close relations to large companies or SMEs (e.g. because of other projects or because they were employed themselves in a company before they joined a university) the transition of fellows into a company is expected to be smooth. A final possibility is founding new companies: Many of the participating universities like UDS or UEDIN support fellows that develop ideas for new products as part of their research in founding a company. This support typically includes the provision of (i) additional training for the founders of a new company, (ii) initial infrastructure for new companies and (iii) professional advice on how to run a start-up.
Invention disclosures: Another means by which the results of SCALE will be exploited is through invention disclosures. They present the first step to protecting inventions. The partners expect roughly one invention disclosure for every six months of industrial secondment.

5.5.9 Financial Management

The financial management within SCALE will be carried out by the project management office (EPO, see Figure 3 above) and involves (i) the set-up and maintenance of financial records, (ii) the co-ordination and control of cost claims and audit certificates submitted by the partners, (iii) the follow-up of EC payments, and (iv) the distribution of partner shares and the monitoring of payments. As indicated above (see description of the project management office), EPO has substantial experience in the administrative and financial management and conduct of European research projects. Once the funding is received, EPO will distribute it to the partners according to the number of fellows recruited and the associated research and training costs. A percentage (to be agreed at award stage) will be retained until the final cost statements are produced by the partners. The total estimated funding for SCALE is 2.749.219 Euros. The majority of this funding (an estimated 1.987.790 Euros) will be used for the activities carried out by the recruited ESRs and ERs (i.e. categories A-D). The remaining funding (an estimated 511.500 Euros) will be for the activities carried out by the host institutions (categories E-G) and overhead (category H, approx. 249.929 Euros). With an estimated 193.500 Euros, the management budget will not exceed the required limit of 7% of the total budget. Within SCALE, the management costs consist of personnel costs associated with the overall management and co-ordination of the project including, amongst others, the co-ordination and conduct of the monitoring and reporting activities and financial management, as well as costs for audit certificates and reporting activities at each partner. For the co-ordinator UdS,

this amounts to an estimated 121.000 Euros and for each remaining network partner to 14.500 Euros.

5.6 Recruitment Strategy

The recruitment of researchers will be conform with the requirements of the Code for Recruitment of Researchers. The 12 ESRs will be recruited in two cohorts - a practice which proved to be highly successful in previous research and training projects as it allows a better and wider access to the employment market. The first cohort will be recruited immediately after the start of the project (start of recruitment process at start of project, anticipating starting date for ESRs in month 4-6 of the project). Then the next recruitment process will start in month 6 for the second cohort of ESRs who are expected to start one year after the beginning of the project. Regarding the recruitment of the ERs, the first ER will be recruited together with the first cohort of ESRs at the beginning of the project. The recruitment process for the second ER will start in month 18, so that he/she can resume their work from month 24 at the latest. Each academic partner will be the primary host of one ESR during the first cohort, and one more during the second. The co-ordinator UdS will in addition be host to both ERs.

The recruitment procedure within SCALE is as follows: The positions will not only be announced on the SCALE web-site, web-sites of the individual groups and host universities, the Marie Curie recruitment web site, different national, European and international science job websites, but also be posted on the various mailing lists available in the community. Based on past experience (DFG-IRTG, S2S, and PIRE) this method can easily generate more then 50 applications and result in the hireing of up to 9 ESRs. After all fellowship candidates have submitted their applications, including CVs, university transcripts, letters of recommendation, and statements of research interest, these documents will made available for inspection to all partners in the consortium. The consortium partners will then independently evaluate the applications and form an opinion as to the suitability of each candidate for a SCALE fellowship. These evaluations will include a numerical rating of each candidate. The coordinator will pool together the recommendations of the individual partners and develop a short list of candidates whose applications are to receive further consideration. The short list will be discussed with all SCALE partner sites until a consensus has been reached.

The candidates on the final short list will be invited to a SCALE recruitment symposium, which will be held by the supervisory board and chaired by the overall scientific co-ordinator. After a series of interviews during the symposium, as well as presentations of research interests and career plans by each applicant, the partners will meet to take final decisions as to which candidates are to receive fellowship offers, which are to be placed on a waiting list, and which are to be removed from consideration. Considered criteria will include not only scientific excellence, but also soft criteria like their mobility experience, their capacity to work in a team and effectively communicate ideas developed by the team etc. Shortly after the recruitment symposium, the candidates will be informed of the status of their applications, and the first offers of fellowships will be made.

Equal Opportunities

The European Commission places a strong emphasis on equal opportunities, in particular on gender equality in science (see, e.g., the 1999 Framework put forward by the Helsinki Group on Women in Science or the Commission's 2001 Science and Society Action Plan). While in general the gender balance in engineering disciplines is substantially biased towards the male side, speech technology is rapidly becoming the exception. In part this is due to the emerging link between research in automatic and human speech processing. In Interspeech-2007 we have seen approximately 30% female registrants, and the proportion of female participants among the PhD students was even higher. In SCALE we will target the recruitment activities such that the proportion of females among the ESRs will exceed the 30% observed in Interspeech-2007 attendance.

This goal will be achieved by (i) actively encouraging women and foreign students to participate in this ITN and (ii) providing a flexible, family-friendly, and supportive work environment, (iii) facilitating the mobility

of women/researchers with families. Firstly, female fellows (and similarly foreign fellows and/or fellows with family) will explicitly be invited in the job description to apply for SCALE Ḃy advertising the vacant positions not only on the websites of the individual research groups (which, as indicated, are still somewhat male dominated), but also centrally on the university webpage and on different national, European and international science job websites, SCALE will try to maximize the reach of its advertisement both internationally and with regard to female recruitment. Secondly, there will be flexible working hours at each host institution and/or the opportunity to work from home if necessary. Thirdly, mentors at each host/company which do not only provide general support on gender issues but also organisational support related to arrangements needed if parents travel or if children accompany their parents (e.g. Kindergarten, nanny, school).

6 Impact

6.1 Benefits for Participating ESRs and ERs

High demand for researchers in speech
The growth of the penetration of speech technology enabled applications and services outlined in section 3 has been mirrored by a growing demand for scientists and engineers with a training in the field. Indeed, in the past it has been common for vacancies in the field of speech technology to be filled with students from adjacent disciplines, because of the lack of highly trained specialists in this area. Rather than abstract numbers, the best proof of the demand for speech technology scientists and engineers is the fact that all master and PhD students of the academic partners involved in SCALE have acquired jobs and positions in the field immediately after (and often already before) their graduation. As such, the first and most important benefit to fellows participating in SCALE will be their great career potential in the area of speech technology.

Demand for engineers in general
SCALE sets out to educate scientists who combine a deep knowledge of some aspects of adaptive signal processing and machine learning as they apply to speech processing together with well-founded operational knowledge of many other aspects of pattern recognition, artificial intelligence, and computer science. Moreover, SCALE fellows bring experience in academia as well as industry. Engineers and scientists with such a deep and broad education will have no difficulty in finding jobs as postdocs in tenure tracks in universities as well as in a wide range of commercial companies and government agencies even outside the field of engineering in general. For example, researcher who studied in HOARSE and SPHEAR (speech and hearing training networks in FP5 and FP4) comment:

“At present I have a post-doctoral research position in the Helsinki University of Technology in the Laboratory of Computer and Information Sciences. I am currently building my new research group, which at present consists of one PhD student and one undergraduate student. My current group is still enjoying the collaboration networks and connections I was able to build during my SPHEAR and HOARSE placements,” Dr Kale Palomaki, Helsinki University of Technology.

“I have participated in the HOARSE Marie-Curie Training Network, during the last three years of the project. My field involved mostly engineering for microphone array processing. This time has greatly benefited to my career, as I found my current position through connections made during the HOARSE project,” Dr Guillaume Lathoud, Alpstein.

Gender Balance
As indicated earlier (see section sec:recruitment), even though speech technology is not as male dominated as many other engineering disciplines, it remains a somewhat male-dominated research area. SCALE we will target the recruitment activities such that the proportion of females among the SCALE fellows will be increased. Moreover, we expect that the focus on medium-term applications, many of which are in fields related to health care and social support services and for which SCALE research provides the basis, will help to kindle the

interest of female scientists beyond SCALE. Last but not least, we expect that SCALE activities will support the recruitment of female master and PhD students in the future: By setting an example of best practice in the community, we hope to make this research area more attractive to female fellows which, as we expect, will lead to an improved gender balance in speech processing R&D in the medium-term future.

Mutual recognition of training and degree

The SCALE research and training programme also ensures that all fellows, regardless of their country of registration, will receive the same training. Training will therefore, of course, be mutually recognised. All training courses offered within SCALE will carry a specific ECTS credit, which individual universities will recognise within their PhD programme and count towards PhD studies (Please note: not all European countries operate a credit system for PhD studies, there is such system for ER training). Since all degrees will be awarded in the area of speech processing, mutual recognition of research performance and ultimately recognition of the PhD degree itself is ensured for each fellow and will, beyond SCALE contribute to the unification of research training activities and degrees.

Specific impact on career of ERs

One of the most important requirements for a tenure track position in European universities is international research experience, with strong emphasis on affiliations with leading research groups. Additional requirements include a large number of publications in peer reviewed journals and conference proceedings, and hands-on experience in supervising master and PhD students. Experience in writing successful research proposals is another requirements of increasing importance. Finally, tenure track candidates are supposed to have experience with working not only in academia, but also in industry. The training that the ERs in SCALE will receive includes all those aspects. Therefore, we are confident that participation in SCALE will put the ERs in the best possible position for pursuing a career in academia. Again, a SPHEAR trainee comments: "The interdisciplinary background as well as the international experience acquired throughout the SPHEAR project clearly helped me in getting my postdoctoral research position at MARCS Auditory Laboratories (University of Western Sydney, Australia; 3/2003-6/2006) and my current Assistant Professor position at CAHR (Technical University of Denmark; since 7/2006)," Dr Jorg Buchholz.

6.2 Benefits for Participating Institutions

Motorola

Motorola very much values its contacts and relationships with academia as a source of innovative ideas and research talent and has many different touch points across the range of technologies that the company is exploring. Specific to the speech research area Motorola has sponsored several PhD's students previously (notably with one of the SCALE partners, USFD). However, our ability to continue to do so has been constrained by changing budgetary circumstances. The SCALE programme provides a very effective framework within which we can renew those prior good relationships and as such greatly expands our contact with leading universities across Europe in the speech area. The framework and support of SCALE enables wide collaboration to happen at a totally new level. It is anticipated that the new relationships started within SCALE will lead to continued collaborations and spin-offs into new areas in the future.

The benefits of participation come at a number of levels. Firstly, there is the direct work conducted by the students while on industrial secondment within Motorola Labs. It is anticipated that the students will have a running start provided by a level of expertise obtained from the relevant training they receive within SCALE Ṫhis quality of student enables us to explore leading edge areas in projects that otherwise would not be possible. We expect that this period of industrial experience will also bring benefits to the students by giving them first hand experience of working in industrial research on relevant problems and applying the practical application of their skills. Second, there is the visibility of all the research work conducted within SCALE by all the ESRs on their individual projects. Third, there is secondary benefit from the associated contact with the wider

faculty within the academic departments. Fourth, by the end of the project there will be a pool of well trained researcher talent who we would be interested in recruiting from. Some of the ESRs will no doubt choose to progress careers in academia and also with other organisations, nevertheless this will still be of secondary benefit to Motorola in the bigger picture as the overall technology capability and ecosystem is enhanced, leading to more and better applications and services.

Mobile Devices is one of Motorola's main businesses so these technologies are of particular interest. The speech interface technologies in the SCALE project have particular relevance for mobile applications and the ease of use of mobile computing devices. The expected impact is described in more detail in Section 6.3 on *Benefits to the European Research Area*. As presented there, the particular issue of using a device in the noisy environments of mobile users is important and with the device held at arm's length so the user can view the device screen. Substantial research challenges remain to take the core technologies forward to a level that will deliver high user satisfaction. Technical progress in the speech area is often made in a series of incremental steps (some larger than others). We look forward to the research within SCALE providing worthwhile improvements of commercial value during the project itself but also to the work and people providing the groundwork and expertise for future work to make additional steps beyond this.

Philips

Philips Speech Recognition Systems engages a very active technology development group, mainly with academic background, that maintains close contact to academic and other research institutions. Several fruitful collaborations have emerged through this, for example the project SPARC and the Austrian Competence Network for Advanced Speech Technologies (COAST). The experiences from these and other collaborations have shown that the direct link between academic and industrial research is of great advantage for the improvement of ASR speech technologies at Philips, specifically if challenges from the industrial field are picked up by academia, and if results from academic research can be applied in industrial products.

The expected benefit for Philips from SCALE thus starts with the mutual knowledge transfer between academic research and our technology development group, particularly through the visits of ESRs during secondments. We plan to direct ESRs attention to industrial product specific challenges during the first secondment, and accordingly shape the subject of their theses, which should ensure the beneficial application of theses results in our products. Of course we will specifically benefit if researchers trained in SCALE choose to join Philips after finishing their PhD.

Indirect benefit is expected from the visibility of Philips Speech Recognition Systems as a participant of SCALE and the general focus of this training network on emerging speech technologies, which will foster the European standing in this field, and the expected further close relation with the network partners.

Benefits for Academia

In the experience of partners with prior involvement in international training networks, including USFD, UEDIN, UdS, RUN and RWTH, candidates who apply for positions in research training networks are generally of high quality and possess both the initiative and motivation necessary to succeed within such a programme. International cooperation within training networks fosters excellent science which would not have been possible without the complementary expertise which a network provides. Some very productive long-term collaborations began in training networks.

Through the synergies and cooperative efforts undertaken under the aegis of the SCALE project, it is envisioned that the profiles of all academic partners will be significantly enhanced. Judging from past experience, this is most readily achieved by demonstrating technical excellence in one of the large, open ASR evaluations such as have been staged by the US National Institute of Standards and Technologies for most of the last 20 years. Similar evaluation efforts of speech recognition and related technologies have been undertaken also in Europe,

particularly by the TC-STAR project, whose goal was to develop a speech-to-speech translation system, and by CHIL, which sought to develop the technologies required for intelligent, multimodal living and working spaces. UEDIN coordinated the 2007 Blizzard evaluation of speech synthesis, and will do again in 2008. Also, USFD's participation in RT was a joint AMI participation also involving SCALE partners IDIAP and UEDIN.

Members of the SCALE consortium are well-placed to participate in such future evaluations, inasmuch as the consortium contains members such as USFD and RWTH who have participated in such evaluations in the past, and can therefore help other members such as UdS and IDIAP to build the infrastructure needed to mount a serious evaluation campaign. Moreover, the principal contributors at UdS have significant experience in mounting such evaluation campaigns.

In addition to participating in such open evaluations, the members of SCALE are committed to making source code implementations of the algorithms used for speaker tracking, beamforming, and ASR available for download from a project web site. This is expected to further enhance the profiles of the consortium members, the impact of the project, as well as to have a very beneficial effect on the larger community, and the rate of progress in the several research fields with which SCALE is concerned.

6.3 Further Benefits to the Community and the European Research Area

It is currently estimated that there 2.5 Billion mobile phone users worldwide. While mobile voice calls remain the dominant application these devices continue to increase in functionality bringing ever increasing capabilities that improve communication, information access and entertainment(cf. the marketing of the iPhone by Apple and the rumours around the introduction of the gPhone by Google). At one level, there is the trend for integration of the capabilities of specialised products onto the mobile devices such as cameras and music players. At another level, there is the increased connectivity and bandwidth being provided by 3G and 4G wireless networks and the increased computation power and memory of the device which together mean that the mobile device will become a highly capable portable computing device with web access.

One of the challenges of compact device size is the user interface, particularly in the problem of text entry given the relatively small size of the screen for display. In particular, the alphanumeric keypad provides a barrier to the ease of text entry, whether that be for access to information (e.g. peoples names for directory assistance, place names and addresses for location based services), search (local to the device or on the web) or dictation of email or IM messages. Voice entry and voice output hold the potential of a much improved user interface for these types of application. Voice can be particularly effective when it is combined with visual output and integrated with other modalities to provide a multimodal interface.

However, one area of robustness that remains a challenge is background noise: While performance may be good in quiet conditions, it deteriorates rapidly in higher noise environments. Of course, users want the applications to function in all the environments that they find themselves in and, in general, for mobile device users this often includes a wide variety of tough background noise environments. Moreover, in these mobile information access applications, users normally want to look at the screen to read the information they have accessed. Unless a headset is used, the increased distance from the microphone on the device this implies that the signal to noise ratio and impact of reverberation will be substantially worse. While the use of bluetooth headsets is becoming more acceptable, there is still a large proportion of the population who do not like the inconvenience of wearing a headset.

What is needed to usher in the widespread adoption of ASR and speech synthesis as the user interface of first choice are technical solutions that can substantially improve performance in noisy environments and when a device is held at arm's length, as will be the case when a user looks at the screen of his PDA or cell phone. A large part of the research proposed within SCALE aims directly at this goal, the achievement of which will have a huge impact on the way voice interfaces are used for mobile services and portable computing devices. We can anticipate that the outcomes of the research conducted with SCALE will demonstrate measurable progress

in narrowing the gap between ASR performance obtained with close-talking and far-field microphones. It is realistic to assume, however, that to really understand the issues well enough and develop solutions that deliver the desired levels of performance will take the continuous contributions of ongoing research over many years beyond the project. Therefore, we see the impact of the SCALE project to be at two levels: the first being the contribution of the work within the SCALE project itself and the second being the training of a group or researchers with the depth of knowledge and experience who can continue to contribute to future generations of research building on the foundation of knowledge established from the training received from SCALE.

In addition to providing such human resources to the community, SCALE will also produce a complete suite of public domain software tools containing reference implementations of all algorithms and techniques required for far-field ASR and reactive speech synthesis. This includes person tracking, beamforming, as well as an ASR engine based on weighted finite-state transducers and enhanced versions of the Festival speech synthesis platform. Moreover, through participation in open technology evaluations such as the Speech Separation Challenge and the NIST RT evaluations, the SCALE consortium will demonstrate that these algorithms are in fact state-of-the-art. We anticipate that the availability of reference implementations of all algorithms developed during SCALE as well as a cadre of early stage researchers who are well-versed in the application of such techniques will have a galvanizing effect on the field of ASR research, and will provide the impetus that makes far-field speech to text transcription *the* problem of most interest to the ASR mainstream. We also hope to foster closer ties between the mainstream ASR and acoustic array processing communities through the SCALE project by training scientists who are well-acquainted with both fields.

The development of recognition based on far-field sensors and speech synthesis that adapts to the intelligibility of its output will be *the* decisive innovations that finally make speech the man-machine interface of first choice. Hence, this technology will have a large implication for those firms developing ASR and text-to-speech technology. Because SCALE is a European project, the development of this technology will greatly enhance both the prestige of Europe within the global human language technologies research community, as well as have significant economic impact for those firms and countries that can successfully exploit new product offerings based on it.

7 Ethical Aspects

As indicated in Section 3.5, a data collection effort organized around a dictation task for mobile, hand-held devices, such as cell phones and PDAs, will be undertaken jointly by Saarland University, Motorola, and Philips during the SCALE project. Three to six hours of such data is required for speaker tracking, beamforming, and far-field speech recognition experiments. The speaker tracking experiments will involve determining the locations of active speakers with respect to the microphones of a sensor array, as this information is required for subsequent beamforming. Beamforming does *not* require determining the absolute position of a speaker, nor does it require determining the speaker's identity.

Participation in the data collection activities will be completely voluntary. Before the start of these activities, participants will be comprehensively informed by staff of the partner in charge of the study, both orally and in writing, about the purpose and course of these activities. *Informed consent* will be obtained prior to the collection of any data by each participant in the presence of a witness unrelated to the study. Sensitivity and attention will be focused on the protection of the individual's freedom of choice and respect for individual's autonomy. Most importantly, the participant will explicitly be given the liberty to withdraw from the study at any time without any consequences. Each participant will need to sign an informed consent form. A copy of the consent form will be given to each participant for his/her records. The consent form will highlight that participation in the study is voluntary with a written assurance that confidentiality will be maintained. Overall, the informed consent will include the following information: the purpose of the research, the duration, procedures to be followed, a description of possible benefits to the participant, steps taken to maintain confidentiality, a

description of the policy to withdraw with no loss of benefits, a statement that participation to the study is voluntary, and contact details of the senior scientist in charge for more information.

The *confidentiality* of the data, as it is may lead to the identification of the individual participant, will be protected. As such, all data will be anonymized and treated strictly confidentially. The data will be encrypted using *anonymous* identification numbers (IDs) for use by the researchers. Only non-identifiable and anonymous data will be shared and used among the consortium members and especially for scientific purposes such as presentations or publications. The development of a code to encrypt the personal data of data collection participants is presently under development and will be implemented if and when the SCALE project is funded. The data will be processed electronically by password-protected computers not connected to the internet or external servers with different access right levels. Each researcher involved in the study will have the key code list for identifying participants to be able to follow up on them. All secondary data analyses will be conducted on data without personal identifiers. All data will be collected in accordance with the relevant EU *legislation*, such as (i) the EU Charter of Fundamental Rights, and (ii) Commission Decision on standard contractual clauses for the transfer of personal data to third countries, under Directive 95/46/EC - 15.06.01 (2001/497/EC).

	Yes	**Page**
Informed Consent		
Does the proposal involve children?	No	
Does the proposal involve patients or persons not able to give consent?	No	
Does the proposal involve adult healthy volunteers?	No	
Does the proposal involve Human Genetic Material?	No	
Does the proposal involve Human biological samples?	No	
Does the proposal involve Human data collection?	Yes	p.14
Research on Human embryo/foetus		
Does the proposal involve Human Embryos?	No	
Does the proposal involve Human Foetal Tissue / Cells?	No	
Does the proposal involve Human Embryonic Stem Cells?	No	
Privacy		
Does the proposal involve processing of genetic information or personal data (eg. health, sexual lifestyle, ethnicity, political opinion, religious or philosophical conviction)	No	
Does the proposal involve tracking the location or observation of people?	Yes	p.14
Research on Animals		
Does the proposal involve research on animals?	No	
Are those animals transgenic small laboratory animals?	No	
Are those animals transgenic farm animals?	No	
Are those animals cloning farm animals?	No	
Are those animals non-human primates?	No	
Research Involving Developing Countries		
Use of local resources (genetic, animal, plant etc)	No	
Benefit to local community (capacity building i.e. access to healthcare, education etc)	No	
Dual Use		
Research having potential military / terrorist application	No	
I CONFIRM THAT NONE OF THE ABOVE ISSUES APPLY TO MY PROPOSAL	No	

A Letter of Intent: Motorola

Richard Eden

Director
Motorola Ltd

24th September 2007

To whom it may concern

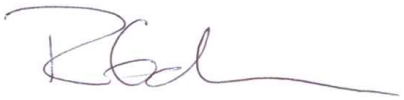
Marie Curie Initial Training Networks (ITN)
SCALE : Speech Communication with Adaptive Learning

This letter is to confirm Motorola's support and commitment to the SCALE project. Motorola will participate at the level of an Associated Partner and provide periods of industrial secondment for the ESR students within Motorola Labs UK, which is part of Motorola's corporate research capability.

Motorola very much values its contacts and relationships with academia as a source of innovative ideas and research talent. The SCALE ITN will provide a very effective and unique framework to expand our collaboration with universities across Europe in the speech area.

Speech interface technologies are important enablers for improved user interfaces to mobile devices, and can greatly impact the way people interact with the information and services facilitated by mobile devices. This has wider implications for a technologically enabled society and also for Motorola as a major player in the mobile communications market. The research on new approaches within the SCALE ITN, and the enhanced training of new talent will progress the technologies and help close the gap between automatic systems and the human levels of performance that users expect.

Yours faithfully



R Eden
Director

Motorola Ltd
Redwood, Crockford Lane, Chineham Business Park, Chineham, Basingstoke, Hampshire, RG24 8WQ
T: +44 1256 790790 T: +44 1256 790053 (Direct) F: +44 1256 790072 (Direct)
Registered Office: Jays Close, Viables Industrial Estate, Basingstoke, Hampshire, RG22 4PD (Registration Number 912182 England)

B Letter of Intent: Philips

PHILIPS

Philips Speech Recognition Systems GmbH

Triester Str. 64, A-1101 Wien

To whom it may concern
(to be included as a part of SCALE FP7 ITN proposal)

September 24th, 2007

Letter of Intent

The mission of Philips Speech Recognition Systems GmbH (PSRS) is to improve patient care by providing the most efficient, accurate and convenient information capture platform. One of the important components of this platform is large vocabulary speech recognition technology. In this area PSRS is one of the market leaders in the legal and healthcare market and its platform SpeechMagic has won several awards.

In order to maintain our strong position, PSRS invests a significant amount of its turn-over in R&D and we are continuously looking for opportunities to strengthen our capabilities and network in the global market. Therefore we intend to take part in the proposed FP7 Marie Curie Initial Training Network SCALE as an associated partner.

Philips Speech Recognition Systems GmbH intends to host a number of 4 Early Stage Researchers (ESR's) during their assignments, preferable 2 ESRs from the first cohort, and 2 from the second cohort. With the research on new approaches and the enhanced training of new talents we see a great value for staying at the leading edge of technology.

Best regards,

Philips Speech Recognition Systems GmbH
Triester Straße 64
A-1101 Wien
www.philips.com/speechrecognition

PHILIPS

Sitz: Wien, Österreich
Firmenbuchgericht: Handelsgericht Wien
Firmenbuchnummer: FN 271 890p

www.philips.at

C Letter of Intent: Eurice

Eurice GmbH
Sciene Park Saar
Stuhlsatzenhausweg 69
66123 Saarbrücken
Phone: +49-681-9592 3360
Email: contact-us@eurice.eu

Saarbrücken, 24th of September 2007

Letter of Intend for Marie Curie Initial Training Network (ITN) SCALE

To whom it may concern

Eurice GmbH is a training partner of the European Commission. Its mission is to provide the most comprehensive and specialized support for the planning and conduct of international EU-funded research projects which in particular includes the provision of the corresponding research training. The company also is a member of EARMA, the leading association of research managers and administrators across Europe.

In co-operation with European experts and the Commission, Eurice GmbH has developed a modularized system and related training contents to adequately prepare scientists for mastering the challenges of working in international (in particular European) research projects. As associated partner of the SCALE ITN, Eurice GmbH intends to provide the network-wide complementary skills training courses which are offered as part of the summerschools.

With best regards,

Jörg Scherer
(Managing Director Eurice GmbH)

ENDPAGE

PEOPLE
MARIE CURIE ACTIONS

Marie Curie Initial Training Networks (ITN)
Call: FP7-PEOPLE-2007-1-1-ITN

PART B

STAGE 2 — FULL PROPOSAL

SCALE