FASTNET

Focus on Actions in Social Talk:
Network-Enabling Technology

1 Overview

1.1 The Second Channel in Human Communication

People are almost perfect communication devices. We have evolved language and speech, and later writing systems, to command and control, inform and entertain, and to generally socialise with each other [1, 2, 3]. Speech is perhaps the oldest form of human communication, yet speech is still only partly understood from the point of view of information technology [4]. There is a strong linguistic component in speech which is well-understood, but there is also a second channel – expressed by tone of voice and manner of speaking – which conveys very important but subtle information about the speaker, the discourse, and the hearer, that is still little understood [5, 6, 7, 8, 9, 10]. This project aims to provide tools and technologies for the processing of second-channel information in speech.

Human speech communication differs from written communication in several ways, the most important differences being in the use of intonation, speaking rate, and phonation style to indicate speaker states, attitudes, and intentions, both towards the listener and with respect to the discourse [11, 12, 13]. It sucessfully integrates the two channels, linguistic and interpersonal, of speech information. Current speech technology, on the other hand, is still largely based on more formal styles of speech that are closer to the written mode than to interactive conversational speech [14, 15], yet domestic users of the technology expect it to be able to respond to their normal modes of everyday conversational speech interaction [16, 17].

Recent research managed by the Lead Applicant and sponsored by the Japan Science & Technology Agency (JST) under the CREST initiative [18] has enabled the collection of a very large 5-year in-situ corpus providing 1500 hours of manually-annotated *natural* conversational speech [19, 20], and a realisation of how ordinary people use their voices in everyday social interaction [21]. These data have yielded some surprising results, one of which is the large amount of *non-verbal* speech that is unobtrusively present in normal everyday conversation. The often fragmented and ‘broken’ nature of spoken language was long thought to be simply due to ‘performance errors’ while the underlying ‘competence’ of the speaker was supposedly better represented by a system such as that used for the written language [22]. It now appears that this view of speech is incomplete, and that the frequent, repetitive, fragmentation of spoken language actually serves to carry a second channel of non-verbal information, essential to a proper understanding of the speaker’s intentions.

1.2 Social Interaction through Voice

In order to process this separate channel of social information, the existing speech technologies will need to be adapted and new modules added [23, 24, 25]. Specifically, new algorithms will be required for the detection and recognition of voice quality and speaking style [26, 27], with particular attention focussed on short backchannel utterances, frequent or idiomatic phrases, laughs, affect bursts, and other non-verbal speech events [28, 29, 30]. These developments will eventually enable a speaker-independent estimation of speaker states and discourse intentions by detection of the long-term and local variations in prosodic and voice-quality attributes carried by conversational utterances [31, 32, 33]. From the small changes across time in the prosodic characteristics of the speaker, which are particularly significant in the frequent and unobtrusive nonverbal speech sounds, the various speaker states can be estimated in a speaker-independent manner. Similarly, by mimicking these small dynamic changes in speech prosody, the equivalent affective states and intentions can be signaled in speech synthesis [34, 35].

A key component of the proposed research will be to produce a dictionary and an ontology of nonverbal speech mannerisms, and to determine the degree, if any, of speaker-dependence in such conversational utterances. The previous work cited above has indicated that many of these mannerisms are language-independent and can be recognised even by speakers of a different language, but there are of course also cultural and sub-cultural dependencies, as well as socially-determined constraints on their use. These remain to be determined as part of the proposed research

The research has practical application in several areas, including robotics, entertainment, call-centres, and other service technologies which use speech as an interface. The main technological focus of the research is to develop sensor devices for the detection and recognition of non-verbal components of speech, including backchannel responses, laughter, and other discourse markers. A central application will be in speech synthesis, since the knowledge obtained about changes in the long-term prosodic parameter settings can be applied in the waveform generation stage of expressive speech synthesis to mimic these speaking-style effects.

1.3 Knowledge & Understanding of the Voice

This work will provide not only knowledge and understanding of how speech communication works *in situ* but also an advanced understanding of how the voice works in communication and how the different strands of information in the voice interact [6, 13, 29, 36]. Voice research is by its nature very interdisciplinary and outputs will have impact on the fields of psychology, linguistics, speech production, speech pathology, and language training as well as speech technology.

1.4 Background – Why this Research Team?

The Lead Applicant has long worked in speech synthesis and recognition technology, but acknowledges that while corpus-based methods are undoubtedly the most efficient, there is still a need to develop and extend fundamental scientific knowledge in speech processing.

The two Co-Applicants have long been working towards an understanding and a model of voice quality and speech prosody [26, 31, 43, 44], and have produced tools and techniques for the manipulation of these parameters [24, 28, 30, 34]. The three of us have frequently exchanged ideas at international scientific conferences, and during visits to the ATR labs in Japan, and we realise that the combination of our skills might be sufficient to produce another paradigm shift in current speech technology. The use of corpus-based methods in speech synthesis, developed by the Lead Applicant, revolutionised the industry ten years ago by enabling the actual voice of a speaker to be used *without degrading manipulation* in concatenative synthesis [37, 38]. This is now the most widely used method for computer voice creation. However, the requirements of a very large corpus of voice samples implies that this system can not be easily miniaturised for use in cell-phones or other portable devices.

By combining our knowledge and applying the voice-quality parameterisation and manipulation techniques to a purpose-built corpus, we will be able to greatly reduce the size of the data that is required, while at the same time improving the naturalness of the synthesised speech that is produced. Because the parameterisation technology can be applied for categorisation of the various voice qualities as well as for their manipulation, the technology can be made two-way, and also used for recognition as well as synthesis. This will enable its effective use as part of a spoken-language interface for a wide range of applications.

The two Co-Applicants have already produced the world's first computer-based speech synthesis system for Irish [39, 40], so this work will extend their parallel research in this field and enable a more conversational style of speech to be synthesised. The Lead Applicant has recently been appointed Stokes Research Professor at Trinity in order to work more closely with the two Co-Applicants, and will be able to port the techniques developed in Japan and to use the new corpus both as confirmation of the generality of his previous findings, and also as part of an enhanced speech processing module for incorporation into interactive spoken dialogue systems for Irish and Irish-English speakers.

Furthermore, a fortuitous by-product of the work will be a greater understanding of crucial dimensions of the Linguistics of Irish, adding greatly to the prior research of the Co-Applicants in this area [41, 42, 45, 46].

2 State of the Art

There is much current international research devoted to social aspects of human interaction, but little is yet focussed on applications in speech technology. The Lead Applicant was present (and perhaps influential) during the initial discussions leading to the instigation of many of them, but was excluded (by residence outside Europe) from active participation in the EU initiatives. We anticipate close future relations, but very little overlap, with each. Nationally funded, current and recent research projects involving the Co-Applicants provide a synthesis test-bed and resources for Irish and will offer many points of interaction with this proposed research project.

2.1 Related International Research Projects

A large EU Network of Excellence, Humaine, has been studying human emotions (the Lead Applicant was called as an external examiner for this project by the EU) and the Co-Applicants have been participants from the outset. This can perhaps be considered the focal point of psychological aspects of related research, but our project is less 'emotion' than 'interaction'. Two new speech synthesis initiatives are related, but taking different approaches, and a 'Social Signal Processing' network that is a likely future partner has just been formed. Previous research projects AMI, CHIL, and MIT have been looking at human conversational interactions, but not in the context of speech synthesis.

2.1.1 HTS *HMM-based Speech Synthesis System* (http://hts.sp.nitech.ac.jp/)

The latest Japanese and US initiatives in speech synthesis research have been devoted to tackling the issue of reducing database size by parametric modelling of the entire speech system. HTS is perhaps the best known of these, but in our opinion (and experience) the resulting degradation of the output speech is unacceptable for use in high-quality systems.

2.1.2 ECESS: *European Centre of Excellence in Speech Synthesis* (www.ecess.eu)

ECESS is a European research consortium, led by Nokia and Siemens, devoted to improving the quality of expressive speech synthesis:

> "A central research question in this field involves what different kinds of emotions can be expressedby systematically controlling the prosodically determined modifications of words in fluent speech. A major topic in this new research field will be to investigate the prosodic variation of local speech tempo in connection with local voice quality, local pitch and local excitation energy. The restricted and strongly correlated combination of these locally varying parameters determines the individual voice characteristic of any naturally given or artificially defined speaker".

This is very close to our own interests and we anticipate eventually becoming an independent (own language) member of the consortium.

2.1.3 Humaine: (http://emotion-research.net/)

Emotion-oriented computing is a broad research area involving many disciplines. The EU-funded network of excellence HUMAINE is currently making a co-ordinated effort to come to a shared understanding of the issues involved, and to propose exemplary research methods in the various areas. The following quote is from their White Paper:

> The obvious application of emotion-oriented computing is as part of the general drive to let machines interface with humans as richly as humans interface with each other. Since emotion pervades natural human interaction, emotion-oriented computing is as fundamental to the drive as speech recognition and synthesis. Its importance is likely to become clearer as allied technologies mature, and the

> lack of emotional intelligence in otherwise humanlike systems becomes seriously anomalous. People already experience the effect to some extent in call centres that give the same bright message however long the user has been waiting.

This philosophy is very close to out own but, rather than assume that social interaction is primarily dependent upon emotions, we prefer instead to tackle this as an issue of language and communication.

2.1.4 SSPNET: *Social Signal Processing* (http://www.sspnet.eu/)

SSPNET is a new EU Network of Excellence with funding running from 1 Feb 2009 for 60 months. The following quote is from their website:

> The ability to understand and manage social signals of a person we are communicating with is the core of social intelligence. Social intelligence is a facet of human intelligence that has been argued to be indispensable and perhaps the most important for success in life.
>
> Although each one of us understands the importance of social signals in everyday life situations, and in spite of recent advances in machine analysis and synthesis of relevant behavioural cues like blinks, smiles, crossed arms, laughter, etc., the research efforts in machine analysis and synthesis of human social signals like empathy, politeness, and (dis)agreement, are few and tentative. The main reasons for this are the absence of a research agenda and the lack of suitable resources for experimentation.

The SSPNet research effort will be directed towards integration of existing SSP theories and technologies, and towards identification and exploration of potentials and limitations in SSP. In that they focus on human-human interaction models and tools for human behaviour sensing and synthesis within socially-adept multimodal interfaces, they are very close to the proposed research, but we perhaps have a head start on them in that our component technologies are already largely well developed. We anticipate closer integration with this project as our own project matures.

2.1.5 COST Action 2103: (http://www.cost2103.eu/)

The Co-Applicants are members of this consortium, whose main objective is to combine previously unexploited techniques with new theoretical developments to improve the assessment of voice for as many European languages as possible, while acquiring in parallel data with a view to elaborating better voice production models.While the main emphasis is on clinical assessment and enhancement of voice quality, the interaction will be important for the development of the voice parameterisation envisaged in the present project.

2.1.6 COST Action 2102: (http://www.cost2102.eu/)

The main objective of this COST Action is “to develop an advanced acoustical, perceptual and psychological analysis of verbal and non-verbal communication signals originating in spontaneous face-to-face interaction, in order to identify algorithms and automatic procedures capable of identifying human emotional states. Several key aspects will be considered, such as the integration of the developed algorithms and procedures for application in telecommunication, and for the recognition of emotional states, gestures, speech and facial expressions, in anticipation of the implementation of intelligent avatars and interactive dialogue systems that could be exploited to improve user access to future telecommunication services". The Action profits from two former COST Actions (COST 277 and COST 278) that identified new appropriate mathematical models and algorithms to drive the implementation of the next generation of telecommunication services such as remote health monitoring systems, interactive dialogue systems, and intelligent avatars. The Lead Applicant has recently been elected a member of the Management Committee, and we foresee closer complementary relationships in the future.

2.1.7 ACII: *Affective Computing & Intelligent Interaction* (www.acii.org)

ACII is an international conference series of which the Lead Applicant is a founder and member of the International Advisory Committee: "Affective computing is a promising area in providing solutions to the many problems in detection, interpretation, inclusion and expression of emotions in future human-robot and human-computer interfaces. Realizing the growth in the field of affective computing in recent years, this conference series aims to provide a forum for scientists and emerging researchers to discuss problems, present their solutions and exchange ideas. ACII aims to provide an environment for the research community of affective computing and human machine interaction to strengthen possibilities for collaborations providing for improved solutions for the existing problems". This community will become a key forum for the presentation of results and for the wider integration of our work.

2.1.8 AMI: *Augmented Multi-party Interaction* (http://www.amiproject.org)

AMI targets computer enhanced multi-modal interaction in the context of meetings. The project aims at substantially advancing the state-of-the-art, within important underpinning technologies (such as human-human communication modeling, speech recognition, computer vision, multimedia indexing and retrieval).

> "The AMI Meeting Corpus contains 100 hours of meetings captured using many synchronized recording devices, and is designed to support work in speech and video processing, language engineering, corpus linguistics, and organizational psychology. It has been transcribed orthographically, with annotated subsets for everything from named entities, dialogue acts, and summaries to simple gaze and head movement."

This community is performing related analytical work, but with different and complementary goals. We will continue a close contact with their work, incorporating recognition developments, using similar technology, and extending their findings for use in conversational speech synthesis.

2.1.9 CHIL: *Computers in the Human Interaction Loop* (http://.server.de/)

This Integrated Project (IP 506909) under the European Commission's Sixth Framework Programme was jointly coordinated by Karlsruhe University (TH) and the Fraunhofer Institute IITB. The project was launched in 2004 and lasted 36 months. The project costs amounted to more than 24 million euros.

> "The objective of this project is to create environments in which computers serve humans who focus on interacting with other humans as opposed to having to attend to and being preoccupied with the machines themselves. Instead of computers operating in an isolated manner, and Humans [thrust] in the loop [of computers] we will put Computers in the Human Interaction Loop. We design Computer Services that model humans and the state of their activities and intentions. Based on the understanding of the human perceptual context, computers are enabled to provide helpful assistance implicitly, requiring a minimum of human attention or interruptions."

We will build on the foundation of this research by enabling computers to make a better and more efficient estimate of the human interlocutor's states & intentions within multimodal dialogue systems. Our philosophy is very similar.

2.1.10 AISB:*Affective Smart Environments* : (http://www.di.uniba.it/intint/ase07.html)

The Society for the Study of Artificial Intelligence and Simulation of Behaviour holds many conferences related to the proposed work.

> "Ambient Intelligence (AmI) is an emerging and popular research field with the goal to create 'smart' environments that react in an attentive, adaptive and proactive way to the presence and activities of humans, in order to provide the services that inhabitants of these environments request or are presumed to need. AmI is increasingly affecting our everyday lives: computers are already embedded in numerous everyday objects like TV sets, kitchen appliances, or central heating, and soon they will be networked, with each other as well as with personal ICT devices like organizers or cell phones.

Although there is not yet any direct link with this work, we share many of their assumptions and beliefs, and anticipate incorporating ideas from this community in the application-development aspects of the proposed project.

2.1.11 Active Listening & Synchrony

The research community is moving in the direction of Active Listening & Synchrony, with a COST 2102 International School being held this year on the theme [?] and an Interspeech (the major international speech processing conference [?]) Special Session devoted to it [?]:

> Traditional approaches to Multimodal Interface design have tended to assume a 'ping-pong' or 'push-to-talk' approach to speech interaction wherein either the

> system or the interlocuting human is active at any one time. This is contrary to many recent findings in conversation and discourse analysis, where the definition of a 'turn', or even an 'utterance' is found to be very complex; people don't 'take turns' to talk in a typical conversational interaction, but they each contribute actively to the joint emergence of a 'common understanding'.

The Lead PI is a leader of both these initiatives which show that a new conversational speech interface is needed, and we see them as directly relevant to the proposed project.

2.2 Related National Research Projects

The Co-Applicants' ongoing and recent research projects provide important resources and points of contact with the research proposed in the present application.

2.2.1 Voice analysis

Recent projects directed by the Co-Applicants, Emovoice, funded by the Irish Research Council for Science, Engineering and Technology, and VOCON, funded by Hitachi, Japan, have involved the development of tools for voice analysis, as well as exploratory analyses of emotive voice. These feed directly into the research proposed here.

2.2.2 SFI funded CNGL

The Co-Applicants are members of the SFI funded C-Set project: CNGL Centre for Next Generation Localisation, for which the Lead Applicant has also acted as international advisor. It involves collaboration between Irish universities and industry, and aims to develop a new model of localisation which allows (among other things) for the personalisation of information. For example, it would ideally aspire "to automatically extract information (such as gender, age, emotion) from speech input relevant to personalisation and generate personalised speech output".

2.2.3 Irish synthesis and related Irish language processing resources

The Co-Applicants have been involved for a number of years in the development of Irish text-to-speech synthesis and related resources. CABÓIGÍN, funded by Forás na Gaeilge lead to the development of the first full synthesis system for the Irish language, available at www.abair.ie. The follow-on project, CABÓGAÍ II, is now extending this facility to the other main dialects of Irish. These projects in turn built on the corpus collection and annotation work carried out in WISPR: *Welsh and Irish speech processing resources,* a joint Wales-Ireland venture funded by EU-Interreg programme, and involving collaboration with the University of Bangor, Wales, TCD, UCD and DCU. The synthesis and the associated resources provide a platform for the novel synthesis prototypes envisaged here.

2.2.4 Prosody of Irish Dialects: funded by the Irish Council for Research in the Humanities and Social Sciences

This project, and follow on research by the Co-Applicants at the TCD lab has served to provide a first account of Irish intonation. This will be a useful background to the prosodic analysis of the Irish dialogue corpora envisaged here.

3 Objectives

The overall objective of the research, as stated in the introduction, is to enable technology for the processing of conversational speech. However, in addition to collection of a conversational speech corpus and the development of models and techniques for the synthesis and recognition of conversational speech mannerisms, we will also incorporate visual aspects of human social interaction in order to produce more robust and general-purpose technology.

3.1 Sensor Devices for Speech Processing

Whereas the main focus of the proposed research explicitly concerns speech processing, it will be clear from the above that the visual dimension of spoken interaction can not be ignored. Systems for future speech processing are now able to make use of all channels of communication information and a camera or video sensor is proving to be almost as useful as a microphone for processing information about spoken interactions [47]. Human interlocutors typically make frequent use of their eyes as well as their ears when talking, to check that the other person is attentive, listening, understanding, agreeing, etc., and devices are now becoming capable of similar processing [48, 49]. Modelling these global aspects of conversations, as well as development of the sensor devices, will form a necessary part of the proposed research.

3.2 Tracking Movement & Synchrony

In meetings-related research [50] similar to AMI, the Lead Applicant has supervised the development of software for detecting faces in a scene and for tracking face-related movements and body-related movements in parallel with the processing of speech in a conversation [47]. By tracking the synchrony of movements between speaker and listener or between members of a meeting, it is proving possible to make informed estimates of their mental processes with respect to the dialogue and discourse engagement [51, 52]. Considerable research has been carried out into the detection of humans in an image and this technology can be considered quite robust [48]. Machines can now see people and easily track their behaviour and movements. The remaining challenge is to employ such technology in speech processing devices and to integrate the visual information with that from the audio stream for a higher-level integration of discourse-related information.

3.3 Eliciting Natural Conversational Speech

Much of the research into multimodal conversational interaction cited above makes use of sophisticated sensor technology including head-mounted microphones, wall-mounted camera arrays, table-centre eye-tracking devices, etc., which can sometimes impede onto the situation and make participants aware of the artificiality of the environment. Considerable research is therefore needed into designing collection contexts in which devices can capture useful and necessary information without being immediately obvious to the participants.

To maximise the usefulness of the captured speech as a resource for further research, it will be necessary to influence the mental states of the participants, in order to elicit specific voice qualities and speaking styles. To this end, we consider it essential that a special-purpose studio be prepared and equipped, in which elements of the ambiance can be controlled. By influencing the mood of the setting and by control over the types of interlocutor, both local and remote, that a subject engages with, the voice quality and speaking styles can be unobtrusively manipulated

In every case, informed consent will be required beforehand and subjects will be offered the right to review their own data after recordings to prevent embarrassment or abuse of personal information, but the recording process for the data collection itself should be as comfortable and 'un-invasive' as possible in order to elicit the full range of voice qualities and interaction mannerisms efficiently.

4 Methodology

The research will necessitate a targeted collection of a large amount of interactive speech data (replicating the scope of previous corpus collection for Japanese but greatly compacted through the benefit of specific techniques and understanding gained with hindsight from the previous experience) and the annotation of parts which particularly indicate speakers' conversational moves, before the development of new algorithms for analysis and recognition of voice quality and other dynamic aspects of prosody and an evaluation of their performance on both Irish and Irish-English speech data.

This will be followed by a study of the paralinguistic and acoustic/auditory characteristics of the conversational speech features, leading to the creation of an ontology of interactive speech moves. Finally, a statistical model of the relation between speech characteristics and speakers affective states and discourse intentions will be developed. The fruits of this stage of the research will consist of the annotated database, the ontology of nonverbal speech, and the statistical model for mapping between interactive speech features and a discourse interpretation of the speakers states and intentions.

Data design and collection will predominate as research issues in the first years, in parallel with development of recording and control techniques. Algorithms for the detection of voice-quality dynamics and prosodic cues will be simultaneously developed using appropriate parts of a preliminary corpus during the first year, and these algorithms will be further extended and refined in subsequent years. From years 2 and 3, the initial version of these algorithms will be adapted to include the video information for application in a multi-modal dialogue interface. From year 3, the algorithms will be implemented in a prototype recognition system for testing the automatic detection of speaker states and intentions from paralinguistic features.

Protection of the inventions and evaluation of the prototype implementations will become increasingly important in the latter part of the project. Evaluation of the component algorithms and of the integrated system will take place in years 3 and 5. A symposium for the presentation of results and comparison with related research from other laboratories will be organised in Dublin in the third and final years.

This research does not specifically require any large equipment or major resources other than what can already be made available at Trinity. We will require access to a space of approximately 75 to 100 sq. metres for a recording studio, access to computer networking facilities, workspace etc., which will be provided. We will require additional office space for approximately 10 people. Project-specific recorders (audio and video), computers, microphones, cameras, and memory storage will need to be purchased. There will be considerable labour-related expenses, however, as the manual work involved in corpus data preparation is considerable. This does not require particularly skilled personnel and training will be provided for all temporary staff such as labellers, annotaters, filing staff, and research assistants. Programmers will be hired in the third and fourth years of the project to provide professional software development skills which PostDocs and PhD students cannot be expected to have mastered.

5 Project Management

A series of workpackages has been devised, according to the principal themes of the research: (i) Infrastructure Development, (ii) Corpus Creation and Analysis, (iii) Voice Processing, (iv) Technology Development, and (v) Evaluation & Outreach.

The initial stages of the research will involve collection and transcription of a representative corpus including coverage of Irish and of Irish English, and will link closely into present (separately-funded) research into speech synthesis for Irish[**?**]. The design of the corpus will combine know-how from the Co-Applicants re matters of content, with the previous experience of the Lead Applicant re matters of style and format. In parallel with this corpus collection and annotation, we will develop novel signal processing tools for the analysis of voice quality and speech prosody. Processing will be tested both for the recognition of different voice qualities and speech mannerisms, and for the generation or replication of these effects using speech synthesis devices.

The second stage of the research involves testing the signal processing results in the context of speech synthesis (in close collaboration with the concurrent Irish synthesis project), for the provision of an interactive "chatty" style of speech that will be required for conversational interfaces. The prototype interface thus developed will initially be evaluated through teaching use in Irish-language classrooms, but we envisage its incorporation into more sophisticated commercial applications, such as machine interpretation, robotics, and customer-services. Ongoing testing will go hand in hand with further development and refinement of the analysis/recognition and of synthesis modules. A key innovative element of the research will be to develop methods that allow the efficient collection of conversational speech data without the need for extensive recordings. This will require development of both capture devices (cameras and recorders) and capture environments (equivalent to a recording studio) that encourage participants to relax, interact informally, and maximise the range of speaking styles and formats.

5.1 Workpackages

The package identifier indicates one of the five main strands underlying the proposed research: (i) Infrastructure Development, (ii) Corpus Creation and Analysis, (iii) Voice Processing, (iv) Technology Development, and (v) Evaluation & Outreach, and year of start:

5.1.1 Infrastructure Development

- WP 1.1a : Research Staff Ramp-up
  Recruiting PhD students, Installing Post-docs, Hiring RA & assistants. Equipping office space with photocopier, personal computers, web facilities for data sharing.

- WP 1.1b : Recording Equipment & Environment
  Design of the studio space, purchase of the equipment, installation and testing, calibration of the recordings, installation of filestore & servers.

- WP 1.3 : IP Protection Team
  initial patenting of inventions, discovery of pre-existing techniques, determination of protectable IP.

- WP 1.4 : Graduating PhD team
  final reports (including theses) to be prepared and reviewed, PhD students to be graduated and their future employment considered.

- WP 1.5 : Preparing Follow-on Research
  PostDocs & PhD potential not to be wasted - follow-on research planning based on experiences over previous 4 years, EU funding proposals.

5.1.2 Corpus Creation and Analysis

- WP 2.1 : Conversational Speech Corpus
  Design of the material - what and how to collect. Preliminary recordings using local and remote interlocutors. Selection of suitable informants.

- WP 2.2a : Main Corpus Recording
  Testing elicitation techniques & teleconferencing methods for speech data collection. Initial corpus construction.

- WP 2.2b : Voice and Prosody Analysis:
  Modelling interaction of voice and prosody parameters and relation to communicative intent. Will include perceptual labelling of parts of the corpus.

- WP 2.3a : Annotation Software Services & Data Distribution Tools
  providing web services for remote annotation and processing of the corpus data

- WP 2.3b : Modelling & Recognition Software
  extending parameter extraction techniques to automate detection of stylistic features, evaluating same in context of style recognition.

- WP 2.3c : Ontology Polishing & Interface Development
  completion of the ontology, including reports, web-based publication, and inclusion in the synthesis interface.

- WP 2.4 : Corpus Completion and Full Annotation
  subsidiary recordings for complete coverage in the corpus for each voice, ontology-based targeting, manual & automatic annotation of all speech & video data.

5.1.3 Voice Processing

- WP 3.1 : Design of Voice Parameterisation System:
  Initial design and plan of voice parameterisation system.

- WP 3.2 : Implementation of Voice Parameterisation System:
  Full implementation of the system. Some initial testing from corpus data.

- WP 3.3 : Parameter Extraction Software Refinement
  Complete feature analysis of the entire corpus to produce database of perceptual & acoustic features, final refinement of software.

- WP 3.4 : Robustification and Testing of Voice Analysis System:
  Enhancing robustness of voice parameterisation system for analysis of live recordings. Extensive testing on corpus.

- WP 3.5 : Final Evaluation of Voice and Prosody Modelling System:
  A model of tone of voice in dialogue interactions and integrated voice analysis system.

5.1.4 Technology Development

- WP 4.1 : Fundamental Technology Design
  Design, installation and testing of data capture and statistical modelling software, development of analytical/feature-extraction software and annotation tools.

- WP 4.2 : Multimodal Data Streaming
  linking video and audio processing, jointly modelling the separate streams of information, installing ECA software.

- WP 4.3 : Synthesising Conversational Speech
  special-purpose interface for the concatenative synthesis of phrasal chunks of speech, leading to better understanding of ontology through usable software.

- WP 4.4a : Synthesising Multimodal Interaction
  incorporating voice & movement features in ECA development,

- WP 4.4b : System Development & Testing

5.1.5 Evaluation & Outreach

- WP 5.1 : Publicity
  Ongoing reporting of results and advertising of the project, will include national and international conference and symposium attendance.
- WP 5.3 : Evaluation of the Results
  to include International Symposia at Trinity. Audiences will include both academic & industrial specialists. Web-page services, integration with outputs of other projects.
- WP 5.4 : Marketing of Research Products
  developing closer links with industrial partners for the inclusion of voice-processing modules in practical applications internationally.

5.2 Local Team

We are requesting funding for two postdocs (or equivalent), one with major responsibility for technology development, and one responsible for voice system development. They will help supervise the four PhD students, and the three research assistants employed on this project. The Co-Applicants will each devote 10% of their time. Christer Gobl will have primary responsibility for voice processing, and Ailbhe Ní Chasaide will be involved in corpus design and in voice and prosody modelling. Nick Campbell, as Lead Applicant, will devote 50% of his time to the work and will be responsible for the management of the project and development of specific corpus analysis techniques, statistical modelling, and technology development. The roles of the external collaborators will be as explained below.

An initial ramp-up period is planned, with one postdoc and two PhD students starting from year one, and the rest joining in year two. Since the postdocs will only be employed for four years each, we envisage one PhD student continuing on postdoc status in year five.

5.3 The Contribution of the Collaborators

Professor John Laver (Queen Margaret College, Edinburgh) and *Dr. Janet Beck* (Queen Margaret College, Edinburgh) will contribute crucially to WP2.2 where voice source parameters are correlated with auditory/perceptual ratings of voice quality on the one hand and with affective labelling on the other. Professor Laver and Dr. Beck are renowned experts on the auditory classification of voice quality and may be involved in the training of team members in Dublin.
Professor Julie Berndsen (UCD) will collaborate on WP3.1, where her phonetic feature recognition tools may serve as a preprocessing step in voice parameterisation system. The developing system will enable a two-way refinement which should mutually beneficial. There is already a close collaboration through the SFI CNGL project (see above).
Professor James Mahshie (George Washington University) will collaborate on the testing of the voice parameterisation system, working on WP3.2 WP3.3, and WP3.4.
Professor Francis Nolan (Cambridge University) will collaborate on WP2.2 on voice and prosody modelling.
Professor Hideki Kashioka NAIST (Nara Institute of Science & Technology) Department of

Applied Linguistics in Japan will assist with WP2.3c and WP4.4a.
Dr. John McKenna (Dublin City University) will assist with WP3.1, WP3.2 and WP3.3.
Dr. Rónán Scaife (Dublin City University) will assist with WP3.4 and WP3.5.
Professor Toshiyuki Sadanobu Kobe University, Japan) will assist with WP2.3a WP3.5.
Professor Petra Wagner Bielefeld, Germany will assist with WP4.3 and 4.4a.
Professor Fred Cummings UCD will assist with WP2.1a and WP4.4a

5.4 Gannt Chart

```
         |    yr1    |    yr2    |    yr3    |    yr4    |    yr5        |

wp1.1a   ------------------
wp1.1b    ---    ---------       -----
wp1.3                            ---          -----------------
wp1.4                                                 ----           ----
wp1.5                                                    ---   ---   ----

wp2.1         ------
wp2.2a               -----------       -----
wp2.2b                      -----------------------------
wp2.3a                           -------          ------
wp2.3b                              ---------        -----
wp2.3c                           ----        ----      ----
wp2.4                                          ------------------

wp3.1   ------------
wp3.2               ------------
wp3.3                           ------------
wp3.4                         ------------
wp3.5                                                    ---------------

wp4.1   ------                  ------
wp4.2                ------------------------------------
wp4.3                           -------------------------------
wp4.4a                                       ------------------
wp4.4b                                       ------            ---------

wp5.1  -----------------------------------------------------------------
wp5.3                                    ---                       ---
wp5.4                                         --------------------------
```

References

[1] Hurford, J. “The evolution of language and languages” pp.173-193 In R.Dunbar, C.Knight, & C.Power (eds) *The evolution of culture*, Edinburgh University Press, 1999.

[2] Sampson, G., “Educating Eve: the ‘Language Instinct’ Debate”, London, Cassell, 1997.

[3] Condon, W., S., Communication: Rhythm and Structure. Rhythm in Psychological, Linguistic and Musical Processes, J. R. Evans and M. Clynes. Springfield, Illinois, Charles C Thomas Publisher: 55-78. 1986.

[4] Campbell, N., “Developments in Corpus-Based Speech Synthesis; Approaching Natural Conversational Speech”, IEICE Transactions on Information & Systems, pp.376-383, Vol E88-D, No.3, March 2005.

[5] Campbell, W. N., “Voice Quality; the 4th prosodic parameter”, in Proc 15th ICPhS, Barcelona, Spain, 2003.

[6] Gobl, C. and Ní Chasaide, A. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication,* 40, 189-212.

[7] George MS, Parekh PI, Rosinsky N, Ketter TA, Kimbrell TA, Heilman KM, Herscovitch P, Post RM. “Understanding emotional prosody activates right hemisphere regions”, Arch Neurol. 1996 Jul;53(7):665-70.

[8] Ross, ED, Edmondson and J, Seibert, GB: The effect of affect on various acoustic measures of prosody in tone and non-tone languages: A comparison based on computer analysis of voice. J Phonetics 14:283-302, 1986.

[9] R.Adolphs (2003). ”Is the Human Amygdala Specialized for Social Cognition?” In: The Amygdala in Brain Function. Annals of the New York Academy of Sciences 985: 326-340.

[10] Campbell, N., “How Speech Encodes Affect and Discourse Information - Conversational Gestures”, NATO Security through Science , Vol.18, IOS Press, pp.103-114, May 1, 2007.

[11] Campbell, N., “Getting to the heart of the matter; speech as the expression of affect; rather than just text or language”, Language Resources and Evaluation, Vol.39, N.1, 109-118, 2005.

[12] Chafe, W., “The Analysis of Discourse Flow”, In Deborah Schiffrin, Deborah Tannen, and Heidi E. Hamilton (eds.), The Handbook of Discourse Analysis, pp.673-687. Oxford: Blackwell. 2001

[13] Gobl, C. (2008). Exploring voice source dynamics and its signalling function in speech: techniques and data. *Proceedings of the 6th International Conference on Voice Physiology and Biomechanics ICVPB 2008,* Tampere, Finland, 27-49.

[14] Campbell, N., “Conversational Speech Synthesis and the need for some laughter”, IEEE Transactions on Audio, Speech, and Language Processing, Vol 14, No. 4,July 2006.

[15] “What do people hear? - a study of the perception of non-verbal affective information in conversational speech”, Campbell, N., Erickson, D., Journal of the Phonetic Society of Japan, ,V.8,N.1, pp.9-28, 2004.

[16] Iida, A., Campbell, N. and Yasumura, M. “Design and Evaluation of Synthesised Speech with Emotion”. Journal of Information Processing Society of Japan Vol. 40, 1998.

[17] Iida, A., Sakurada, Y., Campbell, N., Yasumura, M., “Communication aid for non-vocal people using corpus-based concatenative speech synthesis”, Eurospeech 2001.

[18] JST/CREST Expressive Speech Processing project, introductory web pages at: http://feast.his.atr.co.jp/

[19] Campbell, W.N., “Databases of Emotional Speech”, in Proc ISCA (International Speech Communication and Association) ITRW on Speech and Emotion, pp. 34-38, 2000.

[20] Campbell, W. N., “Recording Techniques for capturing natural everyday speech”, in Proc Language Resources and Evaluation Conference (LREC-2002), Las Palmas, Spain, 2002.

[21] Campbell, N., “ On the Use of Nonverbal Speech Sounds in Human Communication”, Verbal and Nonverbal Communication Behaviors V, LNAI Vol. 4775, pp.117-128, 06 Oct. 2007.

[22] Chomsky, et al ... (then and now) ...

[23] Campbell, N., “Expressive / Affective Speech Synthesis”, in Springer Handbook on Speech Processing and Speech Communication, Eds 2007

[24] Ní Chasaide, A. and Gobl, C. (2002). Voice Quality and the Synthesis of Affect. In E. Keller, G.Bailly, A. Monaghan, J. Terken and M. Huckvale (Eds.) *Improvements in Speech Synthesis,* Wiley and Sons, 252-263.

[25] Campbell, N., “Expressive Speech Processing & Prosody Engineering”, in New Trends in Speech Based Interactive Systems , Eds: Fang Chen & Kristiina Jokinen, Springer, 2007

[26] Gobl, C. and Ní Chasaide, A. (1999). Techniques for analysing the voice source. In W.J. Hardcastle and N. Hewlett (Eds.) *Coarticulation: Theory, Data and Techniques,* Cambridge University Press, Cambridge, 300-20.

[27] Mokhtari, P, & Campbell, W. N., “Automatic detection of acoustic centres of reliability for tagging paralinguistic information in expressive speech.”, in Proc LREC 2002.

[28] Ní Chasaide, A. and Gobl, C. (2003). “Voice quality and expressive speech”. *Proceedings of the 1st JST/CREST International Workshop on Expressive Speech Processing,* Kobe, Japan, 19-28.

[29] Ní Chasaide, A. and Gobl, C. (2004). “Decomposing linguistic and affective components of phonatory quality”. *Proceedings of the 8th International Conference on Spoken Language Processing, INTERSPEECH 2004,* Jeju Island, Korea, Vol. 2, pp. 901-904.

[30] Ní Chasaide, A. and Gobl, C. (2004). “Voice quality and f0 in prosody: towards a holistic account”. *Proceedings of the 2nd International Conference on Speech Prosody,* Nara, Japan, 189-196.

[31] Gobl, C. and Ní Chasaide, A. (2003). Amplitude-based source parameters for measuring voice quality. *Proceedings of the ISCA Tutorial and Research Workshop VOQUAL03 on Voice Quality: Functions, Analysis and Synthesis,* Geneva, 151-156.

[32] Campbell, N & Mokhtari, P., “DAT vs. Minidisc — Is MD recording quality good enough for prosodic analysis?”, Proc ASJ Spring Meeting 2002, 1-P-27.

[33] Campbell, W. N., Marumoto, T., “Automatic labelling of voice-quality in speech databases for synthesis”, in Proceedings of 6th ICSLP 2000, pp. 468-471, 2000.

[34] Gobl, C., Bennett, E. and Ní Chasaide, A. (2002). Expressive synthesis: how crucial is voice quality? *Proceedings of the IEEE Workshop on Speech Synthesis,* Santa Monica, California, paper 52, 1-4.

[35] Campbell, W. N. “Processing a Speech Corpus for CHATR Synthesis”. Proceedings of The International Conference on Speech Processing pp.183-186, 1997.

[36] Yanushevskaya, I., N Chasaide, A. and Gobl, C. (2008). Cross-language study of vocal correlates of affective states. Proceedings of the 9th International Conference on Spoken Language Processing, INTERSPEECH 2008, Brisbane, Australia, 330-333.

[37] Campbell, W. N. and Black, A. W. “CHATR a multi-lingual speech re-sequencing synthesis system”. Technical Report of IEICE SP96-7, 45-52, 1996.

[38] Campbell, W. N., “Multi-Lingual Concatenative Speech Synthesis”, pp.2835-2838 in Proc ICSLP’98 (5th International Conference on Spoken Language Processing), Sydney Australia 1998.

[39] Abair – Romhchainteoir don Ghaeilge. A text-to-speech synthesiser for the Irish language. – http://www.abair.tcd.ie/

[40] N Chasaide, A., Wogan, J., Raghallaigh, B., N Bhriain, ., Zoerner, E., Berthelsen, H. and Gobl, C. (2006). Speech Technology for Minority Languages: the Case of Irish (Gaelic). Proceedings of the 9th International Conference on Spoken Language Processing, INTERSPEECH 2006, Pittsburgh, 181-184.

[41] Dalton, M. and N Chasaide, A. (2005). Tonal alignment in Irish Dialects. Language and Speech, 48 (4), 441-464.

[42] Dalton, M. and N Chasaide, A. (2007). Melodic alignment and micro-dialect variation in Connaught Irish. In C. Gussenhoven & T. Riad (Eds.), Tones and Tunes: Studies in Word and Sentence Prosody, Vol. 2, Mouton de Gruyter, Berlin, 293-315.

[43] Gobl, C. and N Chasaide, A. (2002). Dynamics of the Glottal Source Signal: Implications for Naturalness in Speech Synthesis. In E. Keller, G.Bailly, A. Monaghan, J. Terken and M. Huckvale (Eds.) Improvements in Speech Synthesis, Wiley and Sons, 273-283.

[44] Mahshie, J. and Gobl, C. (2003). Estimating glottal parameters in nasalized speech: an analysis by synthesis. Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, 2193-2196.

[45] N Chasaide, A., Dalton, M., Ito, M. and Gobl, C. (2004). Analysing Irish Prosody: a dual linguistic/quantitative approach. Proceedings of the SALTMIL Workshop at LREC2004: First Steps in Language documentation for Minority Languages, Lisbon, 60-63.

[46] OReilly, M., N Chasaide, A. and Gobl, C. (2008). Cross-Dialect Irish Prosody: Linguistic Constraints on Fujisaki Modelling. Proceedings of the 9th International Conference on Spoken Language Processing, INTERSPEECH 2008, Brisbane, Australia, 886-889.

[47] Campbell, N., "Robust real time face tracking for the analysis of human behavior". Nick Campbell, Damien Douxchamps, in Machine Learning & Multimodal Interaction, Springer's LNCS series, 4892, pp.1-15, Dec. 2007.

[48] Viola, P., and Jones, M., "Robust Real-Time Face Detection", International Journal of Computer Vision, vol. 57, no. 2, pp. 137-154, May 2004.

[49] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 27, no. 3, pp. 305-317, Mar. 2005.

[50] Campbell, W., N., "A multimedia database of meetings and informal interactions for tracking participant involvement and discourse flow", in Proc LREC 2006, Lisbon.

[51] see for example the COST 2102 International School held recently at Trinity on Development of Multimodal Interfaces: Active Listening and Synchrony: https://www.cs.tcd.ie/research_groups/clg/COST2102.IS2009/

[52] see also Special Session on Synchrony & Active Listening at the forthcoming Interspeech conference: http://www.interspeech2009.org/conference/specialsessions.php

[53] To access pdf files of the above refs, please visit: http://feast.atr.jp/˜nick/cv/cv.html