### Description of the Proposed Research

## Focus on Actions in Social Talk; Network-Enabling Technology (FASTNET)

Current speech technology is founded upon text. People don't speak text, so there is often a mismatch between the expectations of the system and the performance of its users. Talk in social interaction of course involves the exchange of propositional content (which *can* be expressed through text) but it also involves social networking and the expression of interpersonal relationships, as well as displays of emotion, affect, interest, etc. A computer-based system that processes human speech, whether as part of a robot, an information-providing service, a translation device, or an entertainment system, must not only be able to process the text of that speech, but also be able to interpret the intentions, or *acts,* of the speaker. It is not enough for a machine just to know what a person is saying; it must also know what that person is *doing* with each utterance as part of a discourse.

Previous work carried out by the Lead Applicant in Japan has shown that more than half of interactive speech in everyday conversations takes the form of nonverbal utterances which cannot adequately be transcribed into text. These affective speech sounds carry important interpersonal information related to the states, intentions, and beliefs of the discourse participants, and to the progress of the social interaction, and they form a small set of highly variable sounds in which most of the information is carried by prosody and tone-of-voice. It is this component of speech that makes it such a rich and expressive medium for human interaction, and it is a component that is not yet well modelled by machine processing.

For the development of future speech technology, it is therefore essential to collect a representative corpus of spoken interaction wherein participants display the full range of their daily speech strategies and to use that material to train new modules for interactive speech processing (whether for synthesis or recognition) that can make use of this higher-level information.

### What is the question that this proposal addresses?

This research proposal specifically addresses the question of how speech technology should be produced which is capable of processing not only the lexical content of an utterance, but also its underlying intentions. A human interlocutor intuitively interprets the nonverbal information in speech and tone-of-voice to aid in the interpretation of each utterance in context. It has been shown that a machine can be programmed to perform similar interpretation of speech utterances in Japanese, and the present research intends to generalise and further develop these findings using speech data from Irish and Irish English. Its academic goal is to show that the use of nonverbal utterances in conversation is a characteristic of human speech in general, not limited to only one particular culture or language. The technical goal of the research is to produce devices specifically adapted to interactive or conversational speech that will enable a friendlier and more efficient speech interface for public services and entertainment.

### Why is this question significant?

The research is significant in a multiplicity of ways. While principally being concerned with the advancement of speech synthesis to enable that technology to produce more natural-sounding utterances and to convey affective information as well as linguistic content, the research also enables parallel developments for speech recognition, so that the subtle nuances of human interactive speech can be more efficiently recognised and processed. It also offers, through the combination of these two technologies, the advancement of interactive systems for modular inclusion into e.g., robots, games, information-devices, and multimedia content provision.

Networking is an essential component of human interaction, and a large part of the content of a spoken conversation has as much to do with networking as it does with with the transfer of propositional content. In many cases, it is not '*what you say*', but '*the way that you say it'* that counts in a conversation, yet this component is missing from all current speech technology. It will require a significant paradigm shift in the way we process speech to incorporate this non-propositional information alongside the content of the message to provide a fuller expression (or understanding) of an utterance. The term 'network-enabling' in the project title reflects this component of communication. It recognises that social actions are the essential component of intercourse, and that actions, rather than words, are the prime units to be processed in a discourse.

There is growing international interest in such multimodal interaction processing (see e.g., UC (Universal Communication) in Japan, AMI (Augmented Multimodal Interaction) in Europe, and CHIL (Computers in the Human Interaction Loop) in the US) and the collection of multimodal conversational speech data was identified as a principal task by the LREC (Language Resources and Evaluation Conference) last year.

**How will the question be addressed?**

The initial stages of the research will involve the collection and transcription of a representative corpus of Irish speech, covering the 3 main dialects, 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 integrate software and tools developed separately at ATR in Japan and at Trinity College in Ireland for the signal processing and 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 through artificial means.

The second stage of the research involves testing the signal processing results in the context of speech synthesis, 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 use in Irish-language classrooms, but we envisage its incorporation into more sophisticated commercial applications, such as machine interpretation, robotics, and customer-services, in the third stage of the research. 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 ro relax informally and maximise the range of speaking styles and formats..

We anticipate unfunded official collaboration in this research from the teams of Prof R. Cowie (Queens, Belfast), Prof T. Sadanobu (Kobe University) and Prof H. Kashioka (NAIST, Japan) initially, and will cooperate closely with members of the EU COST Action 2102: Cross-Modal Analysis of Verbal and Non-verbal Communication, of which the Lead Applicant recently was elected to the Management Committee. Since this research issue is of interest to all members of the EU community, we anticipate working more closely with other groups as the project develops.

**What is the value of this research to Ireland? (max. 250 words)**

The Lead Applicant came to Trinity as a Stokes Research Professor to join with the two Co-Applicants to improve Signal Processing of Voice Quality. He aims to prove that previous findings based on Japanese speech are also applicable to other languages and cultures, and considers Irish and Irish-English to be languages that are worthy of support in the development of their speech technology. In addition to the commercial applications mentioned above, there is also a small but necessary market for teaching Irish as a world language, and the proposed research will immediately be tested in that context, providing advanced resources, both materials and technologies, for the teaching of Irish. Having a state-of-the-art conversational-speech corpus of Irish, and speech technologies tuned especially for Irish will provide the people of Ireland with necessary resources for inclusion as an equal member in the EU community. By bringing in top-quality researchers from abroad, and training Irish graduates, this research project will boost the country’s competitiveness with respect to speech technology and processing.

*An Charraig Aonair* is a symbol of Ireland, standing isolated off its coast in much in the same way that Ireland itself stands isolated from mainland Europe, but serving as its westernmost component. This project, codenamed FASTNET, will burn a lilght from these shores that will be seen across the world. It will take the lead in introducing a paradigm shift in speech processing that will help ordinary people to make use of advanced technology in a simple and natural way.