eNTERFACE Workshop Application

Name: Stephen Wilson
Email: stephen.m.wilson@ucd.ie
Telephone: +35317162404, +353863964354

Education: October 2001 – present
University College Dublin, Ireland
PhD candidate
MUSTER project (www.muster.ie)

August 2000 – June 2001
NUI Maynooth
Higher Diploma in Information Technology (2.1)

October 1994 – June 1999
Trinity College, Dublin
BA (Hons) German and English (2.1)

Project Description: Recent research within the area of automatic speech recognition (ASR) has moved away from traditional phoneme (triphone) models towards nonlinear phonological models where speech is viewed as multiple tiers of phonological or articulatory features. This trend has also recently manifested itself in the Audio-Visual Speech Recognition (AVSR) domain, where visual speech is viewed in terms of multiple streams of visual features rather than a single string of visemes. My research seeks to define a useful set of visual features that can be employed in the investigation of further research questions within the field. An existing audio-visual database (http://www.ece.clemson.edu/speech/cuave.htm) consisting of multiple speakers is hand-labelled with respect to certain visual speech gestures. This takes the form of a multi-tiered annotation structure, with separate tiers being defined for each visual gesture. The tiers include, but are not restricted to: *lip-spreading ,lip-rounding, tongue-protrusion, dental-visibility.* A parallel phonemic transcription is carried out and syllable boundaries are inserted. The outcome is a fully labelled bimodal corpus with annotations showing multi-tiered representations of visual speech gestures with complete phonemic and syllabic labelling. The annotated data is then used to automatically learn associations between phonemes and visual gestures. The associations are learned with respect to a language's phonotactics, (the legally permissable combinations of sounds), as defined by the syllables in the training corpus. The learned associations can be used to make predictions about expected visual features given phonemic input, which can then be confirmed or modified by an expert user as necessary. The modified data feeds back into the learning cycle, leading to a more refined set of associations.
The phoneme-to-visual feature mappings can then be used to augment the Time Map Model, a specific model of computational phonology, with additional tiers of information that pertain to visual speech data. The model is tested in various modes, effectively turning tiers "on" and "off" and examining the relative usefulness of the visual feature set in the task of speech recognition. The desired output is a ranked catalogue of visual features and their phonemic cognates.