2007-09-01

Prototype Speech Corpus

Dermot Campbell
Technological University Dublin, dermot.campbell@tudublin.ie

Yi Wang
Technological University Dublin, yi.wang@tudublin.ie

Ciaran McDonnell
Technological University Dublin, ciaran.mcdonnell@tudublin.ie

Recommended Citation

Campbell, D., Wang, Y. & McDonnell, C. (2007) Prototype Speech Corpus. *EuroCALL 2007.* Coleraine, UK. 5-8 September.

Funder: EU FP6

Prototype Speech Corpus

EuroCALL
2007

DIT's prototype speech corpus allows language learners and researchers access to real, informal dialogues–not just the transcripts of dialogues. Because the dialogues were created at a very high acoustic level they are capable of being slowed down using a time-scaling tool for more detailed study of speech production. The recording methodology used produces a natural dialogue exhibiting all the features of native-to-native interchanges.

The speech corpus is capable of serving the needs of students and researchers of spoken language. LinguaTag is the first prototype tool within the language work package in the FP6 EU project SALERO to automatically tag speech for integration with lip synchronisation in the field of animation. It can also be used as a tagging tool for other, linguistic features such as speed of delivery and formulaicity.

Erman and Warren (2000) have calculated that formulaic sequences constitute 58.6% of the spoken English discourse they analysed. Initial studies within SALERO suggest that sections spoken at speeds considerably above the average for the individual speaker can constitute sequences of formulaic language and are characterised by a tonal contour which is flatter than normal. It is anticipated that research into this phenomenon will yield useful insights into native speaker prosody.

The presentation will demonstrate the speech corpus prototype and discuss aspects of the tagging system of interest to linguists. It will be of interest to language teachers and students of English at all levels and researchers in the field of spoken language.

Keywords: natural speech, corpus, tagging, slow-down, speed, prosody, SALERO