2004-01-01

DIT-CALL: naturally speaking but slow

Dermot Campbell
Dublin Institute of Technology

Marty Meinardi
Dublin Institute of Technology

Bunny Richardson
Dublin Institute of Technology

Eugene Coyle
Dublin Institute of Technology

Olivia Donnellan
Dublin Institute of Technology

See next page for additional authors

Authors
Dermot Campbell, Marty Meinardi, Bunny Richardson, Eugene Coyle, Olivia Donnellan, Pak Kui Leung, Ciaran MacDonaill, Elmar Jung, and Charles Pritchard

DIT-CALL – naturally speaking, but slow

Dermot Campbell, Marty Meinardi, Bunny Richardson, Eugene Coyle, Olivia Donnellan, Pak Kui Leung, Ciaran MacDonaill, Elmar Jung, Charles Pritchard

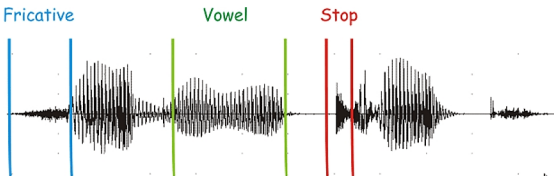
This paper introduces a new development in speech technology and describes its planned application in an EFL context. Special attention is paid to the pedagogical potential of the resultant language tool.

Introduction

'DIT' serves two purposes here, standing both for the Dublin Institute of Technology, where the research group is located, and also for 'Digital Interactive Tools'. The project itself is funded by Enterprise Ireland, which recognises the potential for the study to produce a marketable product. In fact the project has already led to a world-wide patent being taken out.

The three strands involved in the project are language/linguistics, digital signal processing (DSP) and the multimedia interface. Each strand has a postgraduate student and a principal investigator and is managed by a project manager. Dr Eugene Coyle has overall responsibility for academic standards. The principal technological innovations lie in the DSP strand (Reduced Speed Playback algorithms). The Language Teaching & Learning/Linguistics strand provides the creative motivation and methodological innovation. The interface strand is the locus of the digital media software innovation of the project. The combination provides an innovatory enabling technology.

At the heart of the project is the slow-down algorithm Adaptive Overlap Add (AOLA), which can perform time-scale modification (TSM) of speech signals without changing the pitch of the speech sample. The current ANSI-C implementation operates at approx 30 times real-time on recorded sounds and streamed sound TSM is currently under investigation. Sound quality improvement with artefact reduction is also under development. In part this will be addressed by tailoring the application of the algorithm to the speech signal depending on whether a segment represents a stop, a fricative or a vowel. Silence and pauses will also influence the quality of the final, slowed signal. Different speech segments (voiced, unvoiced, vowel etc.) may benefit from being scaled at different rates, i.e. implementation of a non-linear scaling algorithm. In real speech, some segments are more influenced by speaking rate than others. To maintain intelligibility and naturalness, different time scaling factors need to be applied to different segments of speech.



The language/linguistic strand provides the learning environment in which the slow-down algorithm is applied and will provide input to the further refinement of the AOLA algorithm. Its initial target is Mandarin speaking students of English, and in particular Chinese English language teachers.

The multimedia strand is creating a functionally elegant and user-friendly interface for content managers and the student end-user. One of its tasks is to capture the recorded speech and ensure that the linguistic functionality established by the language strand is met.

NS Speech Flow

The ability to slow down recorded, natural – not prepared – speech is particularly relevant to English, given the importance of stress patterning in this language. Cauldwell (2001) convincingly refutes the idea that English is a stress-timed language, but nevertheless stress plays a vital role in native-speaker (NS) delivery and comprehension. DIT-CALL concentrates its attention on the non-stressed elements of the speech stream which cause English learners considerable problems. The greater the number of syllables which occur between two stressed elements, the greater the tendency for speech flow features such as elision, assimilation and co-articulation to occur.

Chinese students of English often have competence in syntactical and lexical aspects of the language but experience considerable difficulties with the spoken language, in both production and comprehension. Their secondary schooling does not prepare them for NS speech flow and when they first encounter it, they have trouble in scanning NS delivery in real time. DIT-CALL aims to assist learners in honing their listening skills so that they can more easily link the reduced forms of unstressed NS–NS speech to their 'passive' linguistic competence. It is not designed for beginners or students with significant language deficits, but for the competent user of the language who has not been exposed to native speech or native speakers.

Learning Environment

The DIT-CALL toolkit is designed to work on a standard desktop PC and will provide the student with a series of 10 lessons which contextualise the linguistic features to be practised. The scenarios cover first arrival in Ireland, through registration and living in Dublin to attending lectures at DIT. While the exercises will be graded to suit learner background and ability, no concessions are made on the model recordings. Only native–native interchanges can provide the linguistic features which require slowing down to aid comprehension, since it is the physical speed of delivery which cause these features in the first place.

A series of exercises will help develop in the learner listener the ability to segment the blur of the speech flow and re-produce it in a more appropriate mode, whether that be slow colloquial as advocated by Gillian Brown (1990) or Jennifer Jenkins' EIL (2000).

Slow-down Function

The application of AOLA will be controllable by the student so that any slow-down factor between 100% and approximately 50% can be selected in order to enhance intelligibility. There is no absolute limit to the slow-down facility, but extremely slow speeds do not add any extra clarity. Should the learner wish to adopt the NS speech as a role model, then s/he can speed up the play-back function gradually until native or near-native speeds have been reached. Slowing down the speech flow, of course, does not produce orthographic forms, but all Chinese learners will be familiar with the citation forms of the individual 'words' spoken in continuous NS speech. It is the 'messiness' of the speech stream (Cauldwell 2002) which causes the difficulty, and therefore which must be tackled in a systematic fashion.

The algorithm could also be used to speed up recorded speech, and there is a possible application in the area of voice mails, but this is not for discussion in the current paper.

Learner Model

While NS–NS recordings are used in order to display and cope with the characteristic features of the speech blur, secondary recordings will be included in order to provide a production model appropriate for students who do not necessarily wish to sound like native speakers of English. This latter model is inappropriate for many learners, who may not even have the opportunity to talk to native speakers, but choose instead to use EFL, ESL or even EIL as their preferred model.

The current version of DIT-CALL will not stress student production or emphasise the comparison between student model and master model, but rather aim at establishing the link between what is heard and what has already been internalised in citation/orthographic form. In order to do this, a series of exercises has been elaborated which is currently being implemented by the multimedia interface strand.

Exercise Types

A number of exercises have been drawn up to capitalise on the ability to 'capture' the normally unheard–at least by the learner. These inculde:

Comprehension
A range of comprehension exercises, at various grades of difficulty, test global comprehension and allow the learner to home in on the contextualising video clips and sound recordings. Other types of comprehension exercise are multiple choice questions, true/false quizzes and cloze tests.

Word Order
This exercise type utilises the multimedia capabilities of the programme platform to promote listening skills in the student, who must scan short-term memory in order to reconstruct the original recording. Since the semantic content of the utterances is already given, listening and re-listening focusses on the manner of speech production. As with all activities, the AOLA slow-down algorithm is available to the user at any time, should this prove beneficial.

Distractors
Listening environments vary in quality, and correspondingly not all recordings will be of studio quality. Some recordings – such as public announcements – will imitate PA systems, with their frequently degraded acoustics. Another exercise type will add noise randomly, such as banging doors, or ringing phones. Students can thus train themselves to exploit the redundancy built into the speech stream.

Production Exercises
Students will be invited to imitate a model utterance, which raises the thorny question of which model to select. This choice will be facilitated by parameters input by the student at the start of the exercise. If s/he logs a wish to sound like a native speaker, then exercises can be offered by the programme which facilitate this model.

For less advanced students a 'second pass' model can be made available. This is akin to the 'slow colloquial' model advocated by Gillian Brown, but less sustained. The term is meant to convey a slightly less speaker-oriented version of an originally not fully comprehended utterance. It is still colloquial, not yet slow and intended by native speakers to clear up any acoustic problems with the original utterance.

The present authors agree with Gillian Brown in advocating 'slow colloquial' as a model of merit, in that it is easy to obtain naturalistic recordings using NS speakers accommodating NNS listeners. It is a sustainable model – unlike 'second pass' and comes closer to EIL than any other NS model. A parallel study by Bunny Richardson will look at the possibility of establishing a lingua franca core for capturing L1 influenced English production using a modified speech recognition programme with a view to providing enhanced corrective feedback to the learner. Her findings will ultimately feed into the further development of DIT-CALL.

Other recordings will accommodate students who – for whatever reason – wish to retain an L1 influence in their EFL production. Master recording will be made making use of EFL speakers who display a nativised but sufficiently clear production as to ensure intelligibility.

In the current version of DIT-CALL students will be invited to record their own attempts at imitating the speech model selected. They will be able to compare master and student versions and slow down the master recording where appropriate. A later development of the programme will concentrate on enhancing feedback to the student, but the current version will concentrate on listening skills rather than speech production.

IPA Recognition

Cauldwell (2002) points out the limits of comprehension exercises based on listening texts, claiming that this type of activity sharpens semantic awareness rather than promoting accurate listening to the manner of articulation in the original recording. In this type of exercise students are presented with speech samples which offer several characteristic features of streamed speech, with its reduced forms, elisions etc. They will then be able to choose the best fit from several IPA versions in an associated table. Experience has shown that Chinese students are quite proficient in the use of IPA and this exercise will reinforce and build on this capability. In addition it will encourage the student to scan and re-scan the recorded passage so as to listen to the manner of delivery, rather than the content, thus addressing Richard Cauldwell's concern over comprehension exercises. This activity also goes some way towards addressing John Fields comment: '…the learner may be unable to segment the word out of a piece of connected speech' Field (2003)

He also agrees with Brown that stressed syllables are the most important part of speech for the learner to focus on and he states that: '…in connected speech, such syllables are 'islands of reliability': louder and longer than unstressed ones.' (p. 329). He notes that: …'unlike readers, listeners do not have regular indications of where words begin and end.' (p.327). Field also reiterates that the difficulties for foreign learners of English especially occur in connected speech, which again reinforces the reasons for the DIT-CALL programme to present the audiovisual material as native natural speech as these speakers are most likely to give examples of connected speech. He explains that segments of speech and even entire phrases in connected speech are reduced. Cauldwell (2001) , has also found that: 'The stream-like characteristics of everyday spontaneous speech change familiar words to such an extent that they become unrecognisable'. He warns against not identifying this phenomenon in lesson materials and so misguiding students into thinking that they will hear the citation form of speech in reality. He advises that the goal for listening work should be: '…to make students familiar and comfortable with the real-time acoustic blur of the stream of speech, and the way in which this stream is shaped by speakers to communicate meanings in all contexts.' (p.8).

It is felt that it is most beneficial for the target group of the DIT-CALL programme, predominantly Chinese speakers of English, that the audio- and video material should use Natural English, as spoken in a non-citation form by native speakers (NS). Brown (1990) suggests that 'students need help in learning to interpret the spoken form of the language and, in particular, the form of the phonetic signal', and a task-based approach to listening skills should be adopted (p.146). As they will be presented with authentic NS speech, spoken at natural speed, students may or may not have understood the entire passage, but by presenting them with original un-modified material, they will have to try and work out what the speakers must have meant and this way they have to use the same everyday listening skills that a native speaker uses, and which they themselves would use in their own native language. For this purpose, the audio-visual material shall not be graded to different learner levels, although the actual exercises will be. It is important that learners are exposed to language where there is an occurrence of 'diminution of phonetic information at the segmental level' (Brown 1990 p. 59). This means presenting them with authentic and preferably unscripted audio-visual material of NS communication which is not slowed down or simplified but captured in as natural a form as possible. Wardhaugh (1993) noted that: 'It is now generally agreed, for example, that reading a passage aloud produces a very careful, formal type of speech because it puts people on their 'best' linguistic behaviour', thus not representing natural native

speech, or what Labov (1972a) has termed the 'vernacular style' meaning the way people talk casually, without feeling that they are observed or listened to.

The skills that a NS uses in order to decode incoming signals are not necessarily available to NNSs. It has frequently been pointed out by various researchers such as Brown and Jenkins amongst others, that NSs process language 'top-down' but that there is evidence to suggest that NNSs process language 'bottom-up', which for example means that NNSs do not rely on context to review misunderstood chunks in the L2. Field (2003) states that: 'Once learners have constructed a set of expectations for a text, they are notoriously reluctant to revise them, even if evidence comes in that contradicts them'. (p. 325). He feels that the commonest perceptual cause of breakdown of understanding is the fact that: '…the learner may be unable to segment the word out of a piece of connected speech'. (p.327). The difficulty for the leaner lies in the problem with being able to know where a word begins and ends in English connected speech. This means that NSs and NNSs must be viewed separately when discussing what constitutes intelligibility. Meierkord (1998) even sees NNS as creators of a separate language and states that NNSs: '… establish a special variety of English, which is effective in informal conversations… Due to their cooperative behaviour, speakers manage to communicate successfully despite their restricted linguistic means.' (p.13)

As in other exercises, the AOLA slow-down function can be employed to help determine which IPA string represents the best fit.

Segmentation into IPA chunks

A somewhat higher order exercise, this task requires students to segment an IPA stream (corresponding to a WAV file which can be re/played as often as necessary) into its word-like components. The reasoning behind this exercise is to highlight to the student the phonemic elements missing in the speech stream, compared to the internalised and expected citation forms. The necessity to 'provide adequate training in strategies which compensate for gaps in word recognition' has been recognised by John Field (Field 2000); also 'Awareness of this kind of feature can aid learners in producing these clusters, as well as recognizing what has been omitted.' ibid. The need for such an awareness is mirrored in Gillian Brown's (1990) comment: 'The problem for foreign learners is that so much disappears in the stream of normal speech that it is not clear to them (learners) how many words there are supposed to be in an utterance and where their boundaries might lie.' (p. 150). She feels that the most salient part of the word is the stressed syllable, 'since this is the best and most stable feature of the word's profile, and to those words in the stream of speech which are stressed, since these mark the richest information-bearing units.' (p. 151). Chela-Flores (2001) believes that priority should be given to: '…factors that contribute more significantly to an intelligible communication…and suprasegmental aspects such as rhythm and certain features of intonation, which have been considered significant phonological factors in helping to organize speech into units of information.' (p.88). Like Brown, she advocates that pronunciation teaching material should ideally focus on meaningful chunks rather than separate segments as she finds that this would pose fewer problems for the learner later on, as it '…reduces the problem of transfer from the segment or word to larger units…'. She adds that when learners concentrate too much on segments of pronunciation, the flow of speech will be interrupted (p.88-89).

The exercise by its nature encourages scanning and re-scanning the recorded NS speech until the link between 'speech blur' and orthographic form is established.

Note taking
Without wishing at this stage to enter the controversy over whether listeners process top-down or bottom-up, this exercise attempts to promote listening skills in a two-stage approach. Initially the student is required to listen to a longer recording of a lecture and note down key words only. These will be compared by the programme with key words pre-selected by the author and a reconciliation effected. Once the main ideas have been captured the student is then invited to expand on the keywords – from memory, as much as possible – to produce the equivalent of lecture notes. 'Cheating' will be allowed in the form of re-playing the lecture file and applying the slow-down facility, but this could be penalised in test mode.

This approach is an attempt to counter Chinese students' tendency to copy verbatim what is being said and to train them in the art of listening globally for key information and re-forming their notes around these central concepts.

Conclusion

The DIT-CALL project is not only a fruitful interaction between the disciplines of language, engineering and computer science, but also enlightening as regards interdisciplinary research. Effective communication is at its core, and as linguists (speaking only for Strand-1 participants) we tend to be fixated on form and function rather than the wider, richer view of communication which *includes* language. The project exemplifies the compromise between what is desirable and what is possible and feasible. The real novelty of the undertaking is perhaps not just the AOLA algorithm itself, but rather the focus of all three strands on maximising its yield for linguistic purposes.

There is also an issue in balancing the deliverables required by an external funder and the requirements of a third-level educational institution. But the positive side of this creative tension is that it helps keep all strands focussed on what is very much an applied research project.

Should any reader wish to have a sample of the slow-down algorithm, please send an e-mail to: dermot.campbell@dit.ie or dcampbell@eircom.net.

Acknowledgements

This work was funded by the Enterprise Ireland administered Advanced Technologies Research Programme (ATRP) 2001, - Project ATRP/01/203, "DITCALL - Digital Interactive Tools for Computer Assisted Language Learning."

References

Brown, G. 1990. 'Listening to Spoken English', second edition. London: Longman.
Cauldwell, R. 2001. Phonology for Listening: Relishing the Messy. Published on the internet at www.speechinaction.com
Chela-Flores, B. 2001. 'Pronunication and language learning: An integrative approach'. IRAL 39 (2001), 85-101. Walter de Gruyter
Field, J. 2000. 'Not waving but drowning: a reply to Tony Ridgeway'. ELT Journal 54/2: 186-95.
Field, J. 2003. 'Promoting perception: lexical segmentation in L2 listening'. ELT Journal Volume 57/4 October 2003, Oxford University Press. pp. 325-333

Jenkins, J. 2000. The Phonology of English as an International Language. Oxford: Oxford University Press.
Labov, W. 1972a. Language in the Inner City: Studies in the Black Vernacular. Phildelphia: University of Pennsylvania Press
Meierkord, C. 1998. 'Lingua Franca English: characteristics of successful non-native-/ non-native-speaker discourse.' Unpublished article previously presented as a poster at the 11th AILA World Congress of Applied Linguistics, 1996, Jyvaskylä, Finland.
http://webdoc.gwdg.de/edoc/ia/eese/artic98 /meierk/7_98.html
Wardhaugh, R. 1993.'Investigating Language. Central Problems in Linguistics.' Oxford: Blackwell Publishers.