*INTERSPEECH 2016*
September 8–12, 2016, San Francisco, USA

# CloudCAST - Remote Speech Technology for Speech Professionals

*Phil Green*[1], *Ricard Marxer*[1], *Stuart Cunningham*[1], *Heidi Christensen*[1],
*Frank Rudzicz*[2,3], *Maria Yancheva*[3], *André Coy*[4],
*Massimiliano Malavasi*[5], *Lorenzo Desideri*[5] *and Fabio Tamburini*[6]

[1]University of Sheffield, United Kingdom,
[2]Toronto Rehabilitation Institute, Canada,
[3]University of Toronto, Canada,
[4]University of West Indies, Jamaica,
[5]AIAS Onlus Bologna, Italy
[6]University of Bologna, Italy

## Abstract

Recent advances in speech technology are potentially of great benefit to the professionals who help people with speech problems: therapists, pathologists, educators and clinicians. There are 3 obstacles to progress which we seek to address in the CloudCAST project:

- the design of applications deploying the technology should be user-driven,
- the computing resource should be available remotely
- the software should be capable of personalisation: clinical applications demand individual solutions.

CloudCAST aims to provide such a resource, and in addition to gather the data produced as the applications are used, to underpin the machine learning required for further progress.
**Index Terms**: assistive technology, clinical applications of speech technology

## 1. Introduction to CloudCAST

In this position paper we introduce CloudCAST, a Leverhulme Trust International Network funded from January 2015 for 3 years. The network partners are The University of Sheffield (United Kingdom), AIAS Onlus Bologna (Italy), The University of the West Indies (Jamaica), and the University of Toronto (Canada).

In recent years, there has been significant progress in Clinical Applications of Speech Technology (CAST), notably in diagnosis of speech disorders [1], tools to correct pronunciation and improve reading skills [2], recognition of disordered speech [3] and voice reconstruction by synthesis [4]. The aim of CloudCAST is to facilitate progress in this domain by providing a freely-available platform for worldwide collaboration.

We aim to place CAST tools in the hands of professionals who help clients with speech and language difficulties, including therapists, pathologists, teachers, and assistive technology experts. We intend to do this by means of a remotely-located, internet-based resource 'in the cloud' which will provide a set of software tools (free-of-charge, if possible) for personalised speech recognition, speech diagnosis and interactive spoken language learning. Following a user-centred design methodology, we will provide interfaces which will make these tools easy to use for professionals and their clients, who are not necessarily speech technology experts.

There are various models for user-centred design [5], among which the ISO standard 9241-210 [6] is prominent. This standard for human-centred design processes includes six guiding principles (P):

- P1. understand the user, the task and environmental requirements;
- P2. encourage early and active involvement of users;
- P3. be driven and refined by user-centered evaluation;
- P4. include iteration of design solutions;
- P5. address the whole user experience;
- P6. encourage multi-disciplinary design.

This paper reports our progress and describes how we have engaged with user groups (P1, P2 and P6 above).

The CloudCAST resources will also facilitate the speech data collection necessary to inform the machine learning techniques which underpin this technology: we will be able to automatically collect data from systems which are already in use, as well as provide a database scheme for collecting and hosting databases related to this domain.

Our 3-year aim is to create a self-sustaining CloudCAST community to manage future development beyond our current funding period.

While CloudCAST will build on previous work by its partners and others, we believe that it offers several 'unique selling points', including:

- The resource will be available worldwide, and free of charge.
- We will provide interfaces, resources and tools targeted at several kinds of users, including:
  - Developers, who want to embed CloudCAST technology into their own applications, for instance voice control of domestic robots,
  - Speech professionals, who want to use CloudCAST technology to work with their clients, for instance, to devise personalised therapy exercise programmes,
  - End users, for whom applications are developed, e.g., children learning to read,
  - Speech technologists, who are improving or adding to the CloudCAST technology itself.
- The technology will be based on open source toolkits such as Kaldi for automatic speech recognition and OpenHab for smart homes [7, 8].



http://dx.doi.org/10.21437/Interspeech.2016-148

- Subject to ethical constraints, we will collect speech data and metadata from every CloudCAST interaction. All this material will therefore be available for re-training the technology, and for analysis. In this way,
  - we will be able to personalise the technology for each End User,
  - by pooling the data, we will address the problem that for abnormal speech the large datasets needed for speech technology development are not available,
  - we will be able to underpin and evaluate improvements in analysis and classification of speech disorders.

## 2. Challenges for CloudCAST

CloudCAST's success requires meeting a number of technical, scientific and more general challenges:

- The technology will run remotely, but in many applications it must deliver results rapidly, within a few seconds.
- The technology should improve its performance as it is used, by adaptation to the data it is collecting.
- It will not be possible to control the conditions under which the tools are used to the extent that one might like. For example, diverse recording devices and recording conditions may make normalisation challenging.
- There must be shared functionality of tools over applications. For instance, pronunciation tutors and reading tutors have much in common.
- There must be interfaces, and guides to these interfaces, which are suitable for each user-group listed above.
- There must be a scheme which protects the security and privacy of CloudCAST users and their data.
- There is understandable resistance to technology from some speech professionals, based on bad experiences.
- For this reason, and others, the technology must adapt to its user, rather than the other way round.
- There must be a strategy for developing a self-sustaining CloudCAST community.

Our intention is to commence with three exemplar applications: small vocabulary command-and-control with disordered speech, a literacy tutor and a computer aid for therapists. These are described after the next section, in which we introduce the common speech technology resource that will support them.

## 3. Speech technology resource

CloudCAST must provide an interactive speech recognition service which allows developers to retrain or adapt the acoustic and language models, keep up to date with technical advances and control the level of detail in the recogniser output. The client should have instant feedback about the recognition process, such as partial decodings, and have access to fully detailed results such as phone-level alignments and posterior probabilities. Crucially, interactions of clients with CloudCAST should supply additional data to improve the recognition process and the training of future models. To provide this flexibility we have adopted the Kaldi toolkit [7], a well-known free software library widely used in the research community partly due to its modular and flexible architecture as our recognition platform. The arguments leading to this choice are outlined in [9]

The architecture of CloudCAST (Figure 1) can be split into the exemplars, the frontend, and the backend. The exemplars are services using CloudCAST, for instance, webapps that perform literacy tutoring or command-and-control (see next section). The frontend is the visible CloudCAST website, from which users can manage their recordings, developers can obtain API keys, professionals can create models, and so on. Finally, the backend is the server which consumes audio from the exemplars and provides speech recognition results. The backend is also in charge of applying the parameter changes that the exemplars may request to the recognition process.

Both the frontend and the backend have access to a common storage space and database for models, recordings, and authentication details. The frontend and backend are both backed by worker processes, whose roles are to perform computationally intensive tasks, such as the training of the models and actual speech recognition, which may be run in separate devices. This split ensures the scalability of the system.

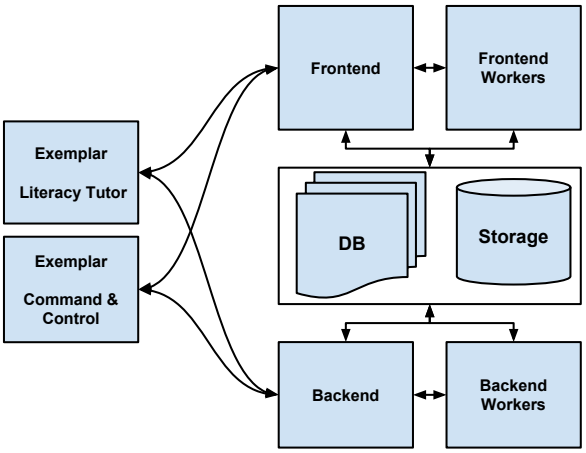


Figure 1: Architecture of the CloudCAST resource.

To implement the speech recognition service (the backend) we have decided to build on kaldi-gstreamer-server developed by Tanel Alume [10]. Kaldi-gstreamer-server is a distributed online speech-to-text system featuring real-time speech recognition and a full-duplex user experience where the partially transcribed utterance is provided to the client. The system includes a simple client-server communication protocol and scalability to many concurrent sessions. The system is open source and based on free software, allowing us to deploy recognisers developed at Sheffield within the CloudCAST framework [11].

To facilitate the creation of services in CloudCAST, we are developing a speech recognition client in JavaScript based on the existing library dictate.js. The proposed client extends dictate.js with multiple types of interactions with the server, such as swapping grammars, models and other parameters, as well as interpreting the different results provided by the server.

## 4. Exemplars and User engagement

### 4.1. Literacy tutor

An automated literacy tutor is being developed as one of the exemplars that will showcase the flexibility and overall utility of the tools provided. Speech-enabled literacy tutors have been developed and deployed with some success (see [9] for a brief overview). These tools, however have limitations relating to

accessibility and flexibility, that the proposed exemplar aims to address.

Initially the tool will be developed for use by Jamaican children learning to read English as a first language and by bilingual students in Italy, to practice speaking in English. In an effort to employ a user-centred approach to designing this exemplar, we have engaged in a number of interactive sessions with intended user groups. Consultations with Italian speech and language therapists and teachers of English have taken place. The recommendations coming out of those meetings are that the tutor design should account for the fact that the users are non-native speakers of English and are learning English as a second language.

Most recently, members of the network met with speech and language therapists, teachers, representatives from the Ministry of Science and Technology and the literacy unit at the Ministry of Education from Jamaica. Insight was gained into some of the issues that could assist in the successful implementation of the literacy tutor in the local context. The need for integration with the existing curriculum and the use of culturally relevant reading material was highlighted. It was suggested that buy-in and continued use of the tutor, and possibly other tools provided under the initiative, would be enhanced if the exemplar was integrated into the training regime for trainee teachers. Issues of accessibility were also highlighted. In a low-resource country access to computers, even within schools, is not a given. However, mobile devices with Internet access are more readily available and are, in some cases, provided by the Ministry of Science and Technology. Thus, the exemplar should be developed for easy implementation on these devices. Once basic functionality is achieved, end-users will be recruited to interact with the tutor and provide feedback for improvement of the design.

### 4.2. Environmental control

The command and control exemplar will provide a service that will allow, for instance, manipulation of multiple devices in a smart home either directly with speech commands or through voice communication with assistive robots. Home automation systems and the increasingly popular Internet of Things (IoT) can provide great support to people with disabilities but require acceptable user interfaces.

There are several ways in which CloudCAST will improve on existing speech-based interfaces. Current systems devised for assistive technology or for the mainstream market are unsuitable, in terms of performance, for many potential users. Common limitations are the inability to be completely hands-free and poor recognition performance.

Command and control systems are particularly useful for subjects with mobility issues. In many cases these people also experience speech disorders for which available speech recognition systems perform poorly. The possibility of using personalised speech models could greatly enhance the recognition accuracy and therefore the reliability of the system. Furthermore the speech material produced by such users will be of great value to improve future speech models for other users with similar issues.

A first consultation with a group of potential users and assistive technology professionals took place at the Ausilioteca of the Associazione Italiana Assistenza Spastici (AIAS) in Bologna. Several issues with traditional voice controlled systems were raised: the complexity and non-accessibility of installation, the steep learning curve for users, and most importantly the low recognition accuracy for disordered speech.

Several solutions are proposed to overcome these challenges. The usage of context-dependent restrictive language will render the system significantly more robust to speech disfluencies, environment noise and recognition ambiguity. The potential of the exemplars can be extended through the use of specific open source servers dedicated to the integration of home automation technologies and IoT solutions, such as Openhab [8]. Employing such solutions permits easy installation and discovery of new devices. Furthermore, through reflection methods available in these systems the voice command layer can be adapted and constructed automatically.

### 4.3. Speech therapy

Speech therapy helps improve communication ability and produces benefits in terms of quality of life and participation in society. It is however time-consuming, and patients rarely receive sufficient therapy to maximise their communication potential [12, 13].

In articulation therapy speech therapists work with patients on the production of specific speech sounds and provide feedback on the quality of these speech sounds. Our previous research shows that computer programs using speech recognition can improve outcomes of speech therapy for adults with speech difficulties [14, 15] In CloudCAST we will develop a web-based application enabling therapists and clients to work together to specify and perform speech exercises. We have run a workshop with therapists in Jamica, where some 7 professionals serve a population of 2.5 million. In such a situation our approach has considerable potential:

- Clients can practice in their own time, on their own equipment, thus freeing up scarce professional resources
- Therapists can monitor and review the progress of their clients
- Subject to ethical approval, speech material collected in this way by many therapists can be used to increase our knowledge and understanding of speech disorders.

A crucial aspect of therapy is the quality of feedback given to the client. We have previously developed techniques for using ASR to provide objective feedback to patients practising their speech [14, 16]. This feedback can be given to patients when they are practising either with a therapist or on their own between therapy sessions [15].

## 5. Data collection and repository

CloudCAST will also serve as a data repository for the distribution of existing databases and for the acquisition of new databases, along with provided tools for that collection. Below we discuss the first database that will become freely available in CloudCAST, TORGO, and the database scheme we will use to represent future data collection

### 5.1. Establishing a new Italian dysarthric speech database

To support the work towards having an Italian environmental control exemplar, we have begun collecting a suitable database of Italian speakers with varying levels of dysarthric speech.

The prompts are digits, command words as well as polysyllabic content words. These have first been extracted from the lexicon provided with the APASCI speech corpus [17] and then we have selected the 100-most frequent words by measuring their frequency using the CORIS/CODIS Italian reference corpus [18].

### 5.2. TORGO

TORGO, the first corpus to be available through CloudCAST, consists of aligned acoustic and articulatory data from individuals with and without cerebral palsy, which is a speech disorder caused by disruptions in the neuro-motor interface [19] that affect the intelligibility of speech but not the comprehension of language [20].

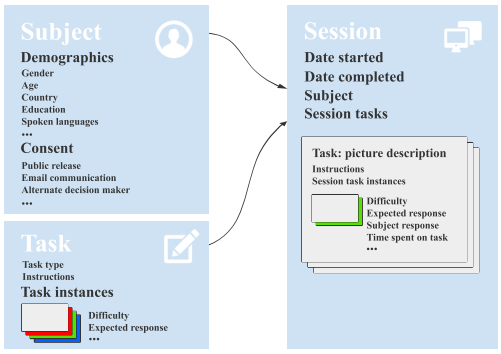


Figure 2: *Simplified database schema, arranged into three core sections: Subject, Session, and Task.*

### 5.3. Future database schema

New users of CloudCAST can immediately use our database framework for representing the data. To a large extent, this framework is designed to be generic to all speech recording tasks, and not all components need to be utilized. The database schema is broken down into three core sections: the subject, the task, and the session. A high-level overview of the data representation is shown in Figure 2.

The subject section generally involves aspects related to the speaker, including demographics, levels of permission to use the data, and factors affecting the subject's language quality, such as country of origin, country of residence, spoken languages, history of smoking, and education level. The task section specifies the language task (e.g., picture description, conversation, reading of text, repetition of audio) along with a bank of available task instances (e.g., pictures to be used in the picture description task). The system supports a variety of question and answer types, including text, speech, multiple-choice, and fill-in-the-blank, with the ability for easy extension to new types. Each task instance is optionally rated with a level of difficulty, measured across arbitrary dimensions (e.g., phonological complexity, syntactic complexity). Information related to automatic scoring of tasks is stored along with each task instance, where appropriate (e.g., the correct answer to a multiple-choice question). Each subject can be associated with a number of recording sessions, and each session can be associated with a number of task instances. The session section stores the subject responses to specific task instances every time they interact with the system. This includes their language data, as well as metadata such as total amount of time spent on each task, and date of completion.

This database is designed to be extensible to future needs, and will be especially useful to streamline data organization to projects that otherwise have a more clinical focus. It enables (i) longitudinal subject assessments, due to the ability to accommodate multiple language task instances in order to avoid 'the learning effect' over time, (ii) dynamic variation of task instance difficulty and type based on subject performance, and (iii) automated scoring of subject performance where appropriate.

### 5.4. Ethics

As part of the CloudCAST initiative, we will be seeking to collect speech data from individual participants. To do so, we must ensure that we fully respect their personal data. As part of this process, professionals who initiate a service through CloudCAST will need to first confirm that they are abiding by the local ethics and governance rules, and we will be promoting international standards for the handling and sharing of such data [21, 22]. In addition to patient safety during data collection, these ensure that data access are controlled, and additional security steps are essential to secure patient's information [23].

CloudCAST is not merely a selection of tools, but it is also meant to be a repository and aggregator of data. Local ethics boards may, by default, not allow data to be transmitted to servers overseas, which is why we propose that protocols for enrolled programs allow participants to opt-in at different levels of engagement. At the most basic level, users can make use of CloudCAST services without uploading their data. The second level uploads data but does not share it with others, and the third retains and distributes data to other speech researchers.

## 6. Related activities

**Helping people who breath using a ventilator**

CloudCAST has an opportunity to work with the Princess Royal Spinal Cord Injuries Unit at Sheffield Teaching Hospitals which treats people with a high-level spinal cord injury (SCI) who use a ventilator to breathe. People with SCIs cannot use traditional interfaces such as keyboards, touchscreens and buttons and, unsurprisingly, there are no commercial speech recognisers which can cope with their altered speech patterns. This CloudCAST exemplar will apply our cloud-based speech technology in a hospital setting for the first time.

**JSALT Workshop** CloudCAST will provide a data hub and cloud-based resource for one of the third Frederick Jelinek Memorial Workshops in 2016 focusing on "Remote monitoring of neurodegeneration through speech".

## 7. Conclusions

CloudCAST aims to create a self-sustaining community of academic and speech professionals which will continue to grow after its 3 year funding period. It is our belief that only by collaborating in this way can we make the benefits of speech technology available to those who need it most and at the same time create the knowledge bases for further technical improvement. To attain critical mass we need to widen the participants beyond the initial partners. If you are interested, please contact us by registering on our website: http://cloudcast.rcweb.dcs.shef.ac.uk/

## 8. Acknowledgements

CloudCAST is an International Network (IN-2014-003) funded by the Leverhulme Trust.

# 9. References


[1] "PEAKS a system for the automatic evaluation of voice and speech disorders," *Speech Communication*, vol. 51, no. 5, pp. 425–437, 2009.

[2] O. Saz, S.-C. Yin, E. Lleida, R. Rose, C. Vaquero, and W. R. Rodrguez, "Tools and technologies for computer-aided speech and language therapy," *Speech Communication*, vol. 51, no. 10, pp. 948–967, 2009.

[3] H. Christensen, S. Cunningham, C. Fox, P. Green, and T. Hain, "A comparative study of adaptive, automatic recognition of disordered speech," in *Proc Interspeech 2012*, Portland, Oregon, US, Sep 2012.

[4] C. Veaux, J. Yamagishi, and S. King, "Towards personalized synthesized voices for individuals with vocal disabilities: Voice banking and reconstruction," in *Proceedings of 4th Workshop on Speech and Language Processing for Assistive Technologies, SLPAT2013*, 2013, pp. 107–111.

[5] S. Blackburn and P. Cudd, "An overview of user requirements specification in ICT product design," in *Proceedings of the AAATE workshop: The social model for AT Technology Transfer*, Sheffield, UK, 2010.

[6] "Iso 9241-210: 2009. ergonomics of human system interaction - part 210: Human-centrered design for interactive systems," Switzerland, 2009.

[7] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, Dec. 2011.

[8] "openHAB." [Online]. Available: http://www.openhab.org

[9] P. Green, R. Marxer, S. Cunningham, H. Christensen, F. Rudzicz, M. Yancheva, A. Coy, M. Malavasi, and L. Desideri, "Remote speech technology for speech professionals - the CloudCAST initiative," in *6th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, 2015.

[10] T. Alumäe, "Full-duplex speech-to-text system for Estonian," Kaunas, Lihtuania, 2014.

[11] H. Christensen, I. Casanueva, S. Cunningham, P. Green, and T. Hain, "Automatic selection of speakers for improved acoustic modelling : Recognition of disordered speech with sparse data," in *Spoken Language Technology Workshop, SLT'14*, Lake Tahoe, Dec 2014.

[12] J. Law, Z. Garrett, and C. Nye, "Speech and language therapy interventions for children with primary speech and language delay or disorder," *Cochrane Database of Systematic Reviews*, no. 3, 2003.

[13] P. Enderby and L. Emerson, *Does Speech and Language Therapy Work?* London: Singular, 1995.

[14] R. Palmer, P. Enderby, and S. P. Cunningham, "Effect of three practice conditions on the consistency of chronic dysarthric speech," *Journal of Medical Speech-Language Pathology*, vol. 12, no. 4, pp. 183–188, 2004.

[15] R. Palmer, P. Enderby, and M. Hawley, "Addressing the needs of speakers with longstanding dysarthria: computerized and traditional therapy compared." *International journal of language & communication disorders / Royal College of Speech & Language Therapists*, vol. 42 Suppl 1, pp. 61–79, Mar. 2007.

[16] M. Parker, S. P. Cunningham, P. Enderby, M. S. Hawley, and P. D. Green, "Automatic speech recognition and training for severely dysarthric users of assistive technology: the STARDUST project," *Clinical Linguistics & Phonetics*, vol. 20, no. 2-3, pp. 149–156, 2006.

[17] B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter, and M. Omologo, "Speaker independent continuous speech recognition using an acoustic-phonetic italian corpus," in *ICSLP94*, 1994, pp. 1391–1394.

[18] R. R. Favretti, F. Tamburini, and C. De Santis, "Coris/codis: A corpus of written italian based on a defined and a dynamic model," *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World. Munich: Lincom-Europa*, 2002.

[19] R. D. Kent and K. Rosen, "Motor control perspectives on motor speech disorders," in *Speech Motor Control in Normal and Disordered Speech*, B. Maassen, R. Kent, H. Peters, P. V. Lieshout, and W. Hulstijn, Eds. Oxford: Oxford University Press, 2004, ch. 12, pp. 285–311.

[20] R. D. Kent, "Research on speech motor control and its disorders: a review and prospective," *Journal of Communication Disorders*, vol. 33, no. 5, pp. 391–428, 2000.

[21] Council for International Organizations of Medical Sciences (CIOMS), "International ethical guidelines for biomedical research involving human subjects," http://www.cioms.ch/publications/layout_guide2002.pdf, 2002, accessed 6 March 2016.

[22] World Health Organization (WHO), "Standards and operational guidance for ethics review of health-related research with human participants," accessed 6 March 2016, ISBN 978 92 4 150294 8.

[23] F. Ozair, N. Jamshed, A. Sharma, and P. Aggarwal, "Ethical issues in electronic health records: A general overview," *Perspectives in Clinical Research*, vol. 6, no. 2, pp. 73–76, 2015.