Cambridge English Centenary
Symposium on Speaking Assessment

Computer recognition of learner speech

Helmer Strik, Radboud University

Standard ‘Automatic Speech Recognition’ (ASR) systems are generally employed to recognize words. For instance, speech-driven dictation systems convert speech spoken into a microphone to text, strings of words appearing on a screen. The ASR system itself consists of a decoder (the search algorithm) and three 'knowledge sources': the language model, the lexicon, and the acoustic models. The language model (LM) contains probabilities of words and sequences of words. Acoustic models are models of how the sounds of a language are pronounced. The lexicon is the connection between the language model and the acoustic models. It contains information on how the words are pronounced, in terms of sequences of speech sounds. Therefore, the lexicon contains two representations for every entry: an orthographic transcription representing how a word is written, and a phonological transcription representing how a word is pronounced. Since words can be pronounced in different ways, lexicons often contain more than one entry per word, i.e. the pronunciation variants, which indicate possible pronunciations of one and the same word.

ASR of native speech is already complex because of many well-known problems such as background (speech) sounds, (low) signal-to-noise ratio (SNR), end-point detection, pronunciation variation, and dysfluencies. However, ASR of learner speech is even more complex, since the grammar, the words used, and the pronunciation can deviate considerably, thus affecting all three 'knowledge sources' of the ASR system (language model, lexicon, and acoustic models, respectively). In the ASR community, it has long been known that the differences between native and non-native speech are so extensive as to degrade ASR performance considerably. Furthermore, native and non-native speech can differ in many (sometimes unexpected) ways, e.g. non-native speech often contains more broken words for (cold) reading, and many more filled pauses in spontaneous speech.

As the quality of speech technology improved, more and more researchers tried to apply it to language learning, sometimes with disappointing results. Some researchers were skeptical about the usefulness and effectiveness of ASR-based Computer Assisted Language Learning (CALL) programs: evidence gathered in different lines of research seemed to confirm that either speech technology was not mature enough, or ASR-based CALL programs were not effective in improving second language (L2) skills. For the sake of our own research, we studied this literature thoroughly and gradually acquired the impression that, while it is undeniable that speech technology still presents a number of limitations, especially when applied to non-native speech, part of this pessimism is in fact due to misconceptions about this technology and CALL in general.

For instance, in some studies unsatisfactory results were obtained when standard dictation systems were used for CALL. Such dictation systems are not suitable for L2 training or for recognition of L2 speech, as CALL requires dedicated speech technology. Apart from the fact that the majority of dictation packages are developed for native speakers,

the major problem here is that CALL and this technology have differing goals and thus require different ASR approaches. The aim of a dictation package is to convert an acoustic signal into a string of words and not to identify L2 errors, which requires a different, more complex procedure. Consequently, the negative conclusions related to the use of dictation packages should be related to those specific cases and not to ASR technology in general.

The following fragments were obtained from the website http://www.ict4lt.org/en/en_mod4-1.htm on 1 June 2013 (the website states the document was last updated on 19 April 2012):

> 1.3 Which skills can be assessed?
>
> Speaking: Very limited as yet. Automatic Speech Recognition (ASR) software is developing rapidly but it is still too unreliable to be used in accurate testing.
>
> To assess speaking skills solely by a computer, using Automatic Speech Recognition (ASR), is a very complex task and research in this area is developing rapidly. ASR can be motivating for students working independently, but computers are still not completely reliable as assessors.

So it seems that the authors acknowledge that ASR can potentially be useful, but are still skeptical about the quality.

Many of the first studies that considered employing ASR technology in the context of L2 learning focused primarily on the automatic assessment of different aspects of L2 oral proficiency, in particular L2 pronunciation. The results showed that automatic testing of certain aspects of oral proficiency was feasible: the scores obtained by means of ASR technology were strongly correlated with human judgments of oral proficiency.

In general, such automatic scores were calculated at a rather global level, for instance for several utterances by the same speaker, because in this way more reliable measures could be obtained. Such measures might be suitable, and in certain cases even preferable, for testing purposes, for assessing the problems of individual speakers, for providing overviews of words or phonemes that appear to be difficult and suggesting remedial exercises for the problematic cases. However, such overall measures are generally not specific enough for practice and feedback purposes.

For training, error detection is required, a procedure by which a score at a local (e.g. phoneme) level is calculated. In general, the relation between human and automatic grading improves if longer stretches of speech are used, i.e. complete utterances or a couple of utterances. Such cumulative measures can also be adopted for error detection, for instance by combining the scores of several utterances. This can be useful to assess the problems of a specific speaker, to obtain an overview and suggest remedial exercises for the problematic cases. However, for remedial exercises immediate feedback based on local calculations is to be preferred.

To sum up, ASR of non-native speech is indeed complex. Still, if one carefully takes account of its limitations it can already be applied usefully. In several projects we have developed speech technology for language learners and have studied different aspects related to the use of such technology. For instance, we investigated the performance of the speech technology used, the way in which it could best be implemented in applications, how such applications should be designed and how feedback can best be provided to the learners, how language learners experience the use of speech technology, including its effect on their motivation, etc. In our presentation we will provide an overview.