When dragons listen

Viktor Tron

Technology has come a long way in recent years and voice-related software has not been an exception. Speech recognition is the name of the technology that underlies practical applications like voice command control or text input by voice dictation. The utility of speech recognition is clear in situations where a hand-free scenario is desirable such as mobile voice control or in-car navigation. But such is the case of people with special needs. Many people with Parkinson's find relief using voice to text solutions to circumvent prooobleems wheeen typing with a shaaaaaky hand. Voice input that allows voice dictation is by now ubiquitous and available out of the box on any reasonably smart phone.

Voice recognition is the term used for technology that recognises a voice in the sense that it identifies an individual by their voice. Just like fingerprints or iris patterns it can be used for authentication but it also found a genius application as a potential early diagnostic tool for Parkinson's based on the insight that PwP have subtly different speech characteristics. Voice tools such as these rely on techniques that turn the continuous audio input into digital representation. This domain of acoustic engineering is called digital signal processing. It is the front-end to any speech recognition solution this is the component that is responsible for filtering out background noise, enhancing speech-specific acoustic features of audio input not unlike the human auditory system.

Speech recognition goes further in that involves transforming a highly ambiguous low level sensory input to high level structured information. When it works, gadgets using them will often give the impression that they understand language. So how does a speech recognition system actually work? Although there has been a lot of incremental steps improving recognition accuracy and performance, the underlying modelling technique is virtually the same as decades ago. Yet only in the last few years did the exponential growth of computating capacity, storage and network speeds allow the technology to reach the end user with a quality that we find useful.

The components of asr map neatly to the two aspects of human speech processing. The bottom up system maps digital representation of the acoustic signal to sounds (abstract units roughly corresponding to phones, think letters) and is called the acoustic model. The top down system maps sounds to sequences of words or utterances and is called the language model. These systems are implemented as sophisticated probabilistic models parameters of which are learned from annotated data: lots and lots of gigabytes of text and hours of pre-recorded speech. When the acoustic and language models are combined the result is a ranked list of guesses. The acoustic model is meant to capture our knowledge of how certain sounds are realized or in other words what is the range of audio characteristics that could plausibly be regarded as an instance of say the 'e' sound. The speech signals that could be categorised as 'e' show a massive variation and modulated by expectations of background noise, speaker's gender and whole array of contextual factors. This warrants the use of statistical models and huge amounts of data to estimate ranges of variation.

The language model is meant to capture our knowledge that sounds are sequenced to form words and the order that words follow each other is to a certain extent predictable. The language model is most commonly modelled as a stochastic sequence, i.e the probability of a word occurring is conditioned on (a few) preceding words. Somewhat surprisingly this simple model is able to integrate sources of contextual knowledge including grammatical structure of a language, topic or common usage. Once again though estimating these probabilities requires training based on huge amounts of text. Having vast amounts of good quality annotated data to train these statistical models hold the key to speech technology and has proved the main barrier of entry in this market.

Written language has still a fair amount of ambiguity or degrees of freedom for interpretation but compared to the messy medium of sound, text is a haven of clarity. This is exactly why the opposite problem of turning text to speech is relatively trivial compared to speech to text. This simpler scenario of text to speech (reading aloud) can however illustrate the main issue in language technology: ambiguity and context. The word 'read' is ambiguous between two pronunciations, one that rhymes with 'red', the other with 'reed'. How do we know which one to say when we encounter a piece of text containing this word. The answer lies in contextual knowledge. We can use the local context of other words to help make a choice: in the context of 'have you read the news?', the local linguistic context of other unambiguous words is able to change the odds, the variant that rhymes with 'red' should be pronounced.

In another context like 'do you read the daily mail?' the variant that rhymes with 'reed' is chosen. There are situations where we go beyond the local linguistic context to do the disambiguation. For instance the word 'reading' is pronounced different if we know that it is in the context of travel destinations or hobbies. This illustrates that information used in narrowing down possibilities can come from a wide variety of sources but also points to the fuzzy nature, we can never be entirely sure and computers as yet are still more often wrong than humans. This is exactly true of speech recognition except that instead of a clearly represented written input and a choice of two pronunciations we have the noisy medium of speech and usually a much wider range of choices. When humans understand language they are able to integrate all aspects of the context to help interpret speech. We do it so effortlessly and automatically that we often hear what we wanna hear. While it can then lead to misunderstanding if our anticipations are strong yet this top-down component is crucial in driving efficient speech processing in that it filters out less plausible options.

A similar process underlies other strategies such as speaker adaptation. Some people have a low voice (low average pitch) while others have a high voice: the same absolute pitch is not the same across speakers. In other words if we talk to someone with a lower voice, our anticipation of pitch range is adjusted. In computer models of speaker adaptation this can be modelled by using relative pitch instead of absolute pitch ie a measure of pitch change relative to some speaker dependent average. This simple technique is called normalisation and is used not only for pitch but a wide variety of speech features. Speaker adaptation can be viewed as just another example of contextual disambiguation. It is known that people understand a familiar voice much better.

By the same token you can familiarize your computer to your speech characteristics by reading in text already known to the system. This enrolment phase as they call it in dictation systems enables the recogniser to adapt to your voice before you start a new utterance. Automated call centres or public voice control interfaces (say, in a lift) can also use speech recognition, often much to our frustration. Since anyone can call in, pretraining is not an available option. The task is still manageable if the context restrict the possible choices, for instance the system only needs to recognise a number to identify which floor you are asking the elevator to take you to or pick which station you want as your destination in a train booking system.

In general then the task of speech recognition is easier if we have more context more knowledge (less ambiguity) in either the domain (what is said) or the signal (how it is spoken). On one end of the spectrum we got dictation: ideally, the user trained the system to their voice and speaks in a clear tone with little background noise however the domain is relatively open. On the other end is a call centre system or voice command control interface that needs to be prepared for noisy channel and anyone calling in but the domain is restricted: the system only needs to distinguish between a handful of options. A voicemail to text system where anyone can call from anywhere and leave a message about pretty much anything on the other hand will always preform worse than the examples above. Simply because both signal and domain are relatively unconstrained.

People often think a recogniser won't understand their accent. This is a myth. Individual speech characteristics, noisy channels and context means that speech displays a massive variation. Within this variation, accent differences are deceptively minute. The statistical models that are trained to generalize across this wide variety of realisations of sound are robust enough to cope with the broadest of accents. Where they are more likely to fail is open class items like surnames, use of rare but short words.

By Grantscharoff (Own work) [CC-BY-SA-3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons