Can voice recognition technology identify a masked jihadi?

January 7 2016, by Ian Mcloughlin, University Of Kent

A masked face but experts still have his voice to go on. Credit: Video screengrab

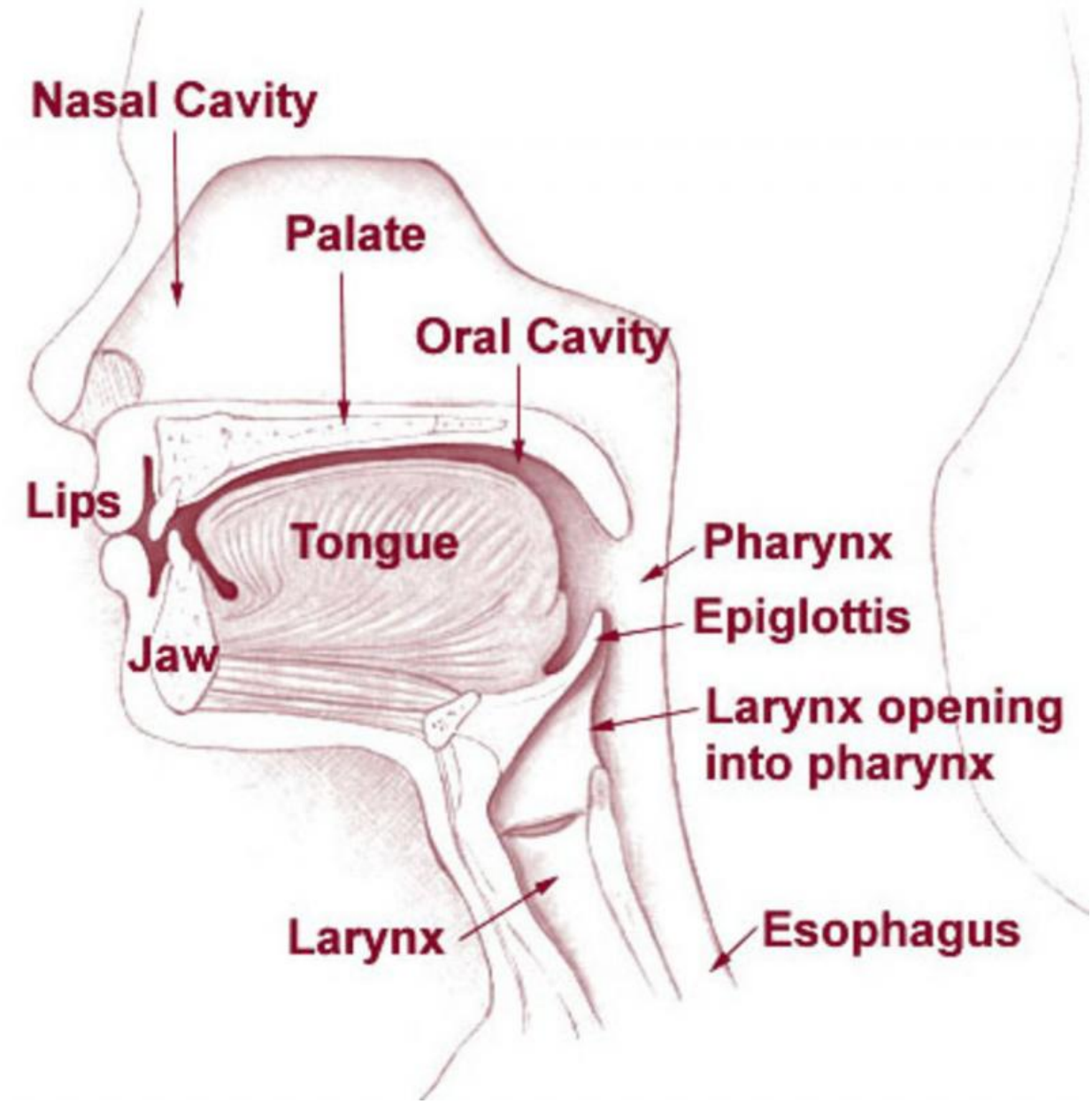
The latest video of a masked Islamic State jihadist apparently speaking with a British accent led to him being tentatively identified as Muslim convert Siddhartha Dhar from East London. Voice recognition experts

were reportedly working with UK intelligence services using voice analysis. But how does this technology work and what is it capable of?

Most of us can, when we hear a voice we know well, recognise who is speaking after just a few words, while less familiar voices might take a little longer. If the context and content of the words spoken are familiar, that makes it easier still. Generally, machines face the same constraints when trying to compare recordings and find a match.

Computational systems that aim to establish who people are from their voices – speaker identification – differ in whether they aim to detect: the presence of a single known speaker; to match speech to one of several known speakers; detect what's recognisable from an unknown recording; or verify that a recording of speech was indeed from the expected speaker.

Modern systems tend to take a big data approach, where machine learning algorithms are trained with large sets of known recordings so they can recognise individual speakers' vocal features. The idea is that the important features that discriminate between different speakers are learned automatically. In contrast, older methods specified which type of linguistic and phonetic features of speech were thought important in order to compare them between speakers.



The surfaces involved in producing speech. Credit: National Cancer Institute

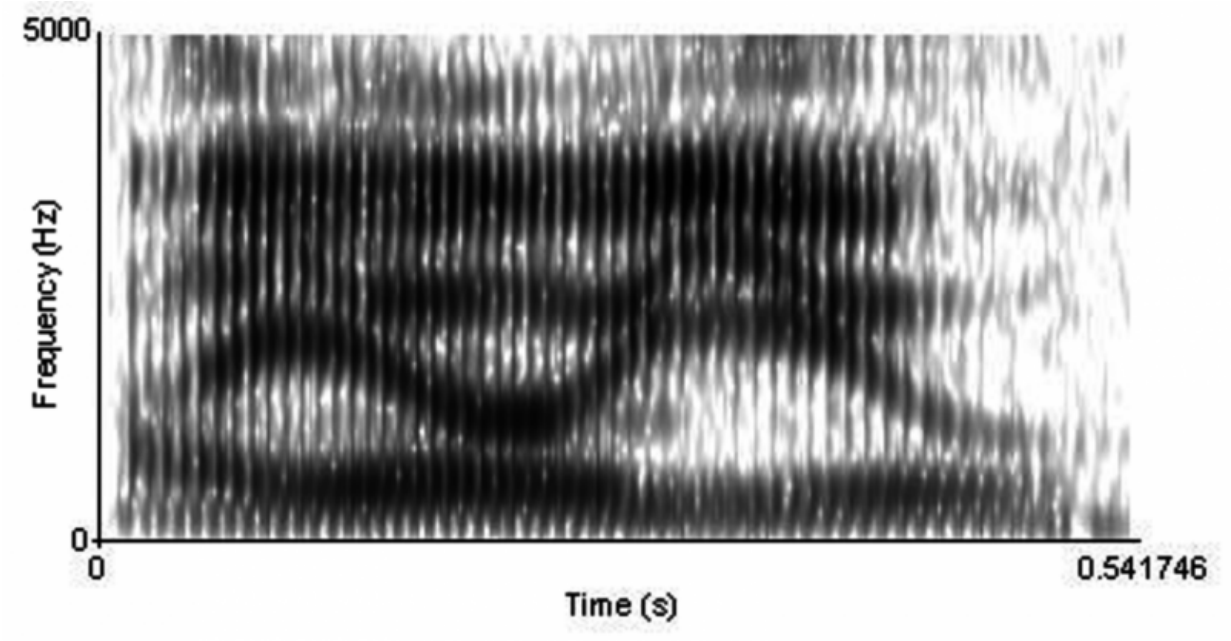
While we don't really know what combination of features is best for voice recognition, we can classify them as either acoustic or linguistic.

Acoustic and linguistic features

Acoustic features are characteristics of how humans produce speech. When we speak, air is expelled from our lungs, travels up the trachea, through the larynx and out of our mouth and nose. As it passes it vibrates against vocal cords which when relaxed or contracted change the frequency of vibration, and so the pitch of our voices.

Several parts inside the vocal and nasal cavities such as the tongue, teeth and lips – known as articulators – modify sounds to create different resonances – called formants – to produce other variable characteristics of speech. What we hear as speech is a combination of all these interactions of air passing through these cavities and over and between these body parts.

Each of us has unique speaking characteristics: the way our lungs exhale, vocal cords resonate, articulators act all produce unique sounds. One person's "a" can be very different from another's – and that's just one of the 44 phonemes (the smallest unit of sound that make up words) in the English language. The way our speech blends the phonemes together and moves from one to another is also different, as is the speed at which this happens. Consider the difference between the steady tempo, rounded vowels of a English country accent with the faster, staccato speech common in bigger cities.



A time/frequency spectrogram of the phrase ‘I owe you’. Credit: Jonas.kluk

Linguistic features relate to which phonemes we choose to use and in what sequence, rather than how they're produced. If I say "tomahto" and you say "tomayto" then we have spoken the same word, with a different choice of phonemes. There are a vast number of alternative pronunciations, based on familiarity and often on regional and generational differences. The choice of word, different words, grammatical patterns, characteristic pauses, stresses, sentence structures or phrases also presents a way to identify different speakers.

At a higher level still is the meaning of the words themselves. We tend to make different choices in what we say and how we choose to say it – how direct, or confrontational, or evasive, or intellectual our way of speaking is. If you've ever met someone and thought they speak like a lawyer, teacher, or artist, then the patterns you recognise can be recognised by computers too.

Making sense of it all

In computational terms, first the linguistic and acoustic features are isolated, condensing the large amount of data into manageable sets of features that succinctly captures their important nuances. Then pattern matching is used to compare these to those from another recording.

Features of speech that can be automatically extracted include pitch, formant frequency, vocal tract length, and the rate at which syllables are spoken. Some modern methods operate better with lower-level features that require less processing and offer less intrinsic meaning to human ears. These are typically two-dimensional maps of time and frequency, such as spectrograms.

Once complex speech has been reduced to a set of more simplified representative features, then a process of generalised pattern matching is applied, establishing how best to make a comparison, and how closely patterns match. Given enough good quality speech to analyse, we can convincingly match the speaker to one person from among a small group of suspects. The more speech we have from both sets to compare, the better the match. In this case, experts had several recordings of Dhar giving interviews when still in the UK.

With no suspects to go on the task would be near-impossible, like searching for a needle in a haystack. But what we can learn and infer about a speaker from a recording can itself reduce the haystack to a more manageable size. For example, expert listeners can narrow down the home region, age, gender, emotion and maybe infer something about a speaker's education. In some cases speech experts can determine where a speaker was born, whether their parents spoke another language, and whether they have lived elsewhere more recently. Perhaps even when they left the UK.

Science fiction or reality?

While much is shrouded in secrecy, speaker identification technology is thought to be used by national security agencies such as GCHQ in the UK, the NSA in the US, and Public Security Bureau in China and so on. It's also widely believed that voice prints are captured at airport immigration counters in some countries, which perhaps explains why you may be asked a meaningless question or two during processing – after all facial recognition is already widespread at airports, why not for voices too?

Commercial voice matching technology from the likes of GoVivace, iFlytek, IBM and Nuance is probably at least a generation behind that used by governments. How useful the technology is at present is debatable, but it is used daily by financial institutions as a means of speaker validation – offering proof that they are who they claim to be.

Voice print analysis has been used in criminal cases since the 1970s, with mixed success, and usually for the less demanding task of proving that speech in a given recording belongs to a particular speaker. Trying to match one speaker from a huge set of possibilities that may not even include the correct match is far more difficult. But it's not impossible, and systems are improving.

This article was originally published on The Conversation. *Read the* original article.

Source: The Conversation