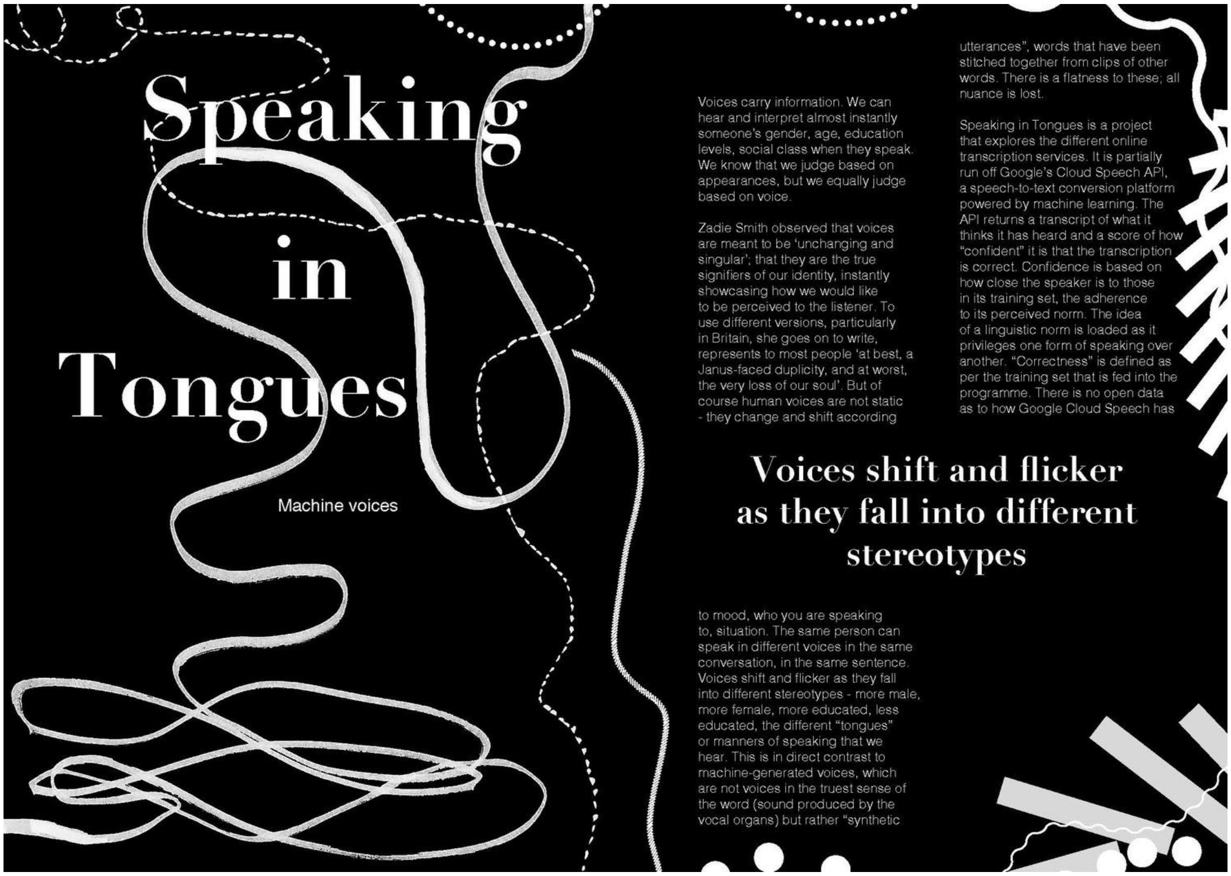
Speaking in Tongues

Syrup Magazine, October 2017



Speaking in Tongues

Machine voices

Voices carry information. We can hear and interpret almost instantly someone's gender, age, education levels, social class when they speak. We know that we judge based on appearances, but we equally judge based on voice.

Zadie Smith observed that voices are meant to be 'unchanging and singular'; that they are the true signifiers of our identity, instantly showcasing how we would like to be perceived to the listener. To use different versions, particularly in Britain, she goes on to write, represents to most people 'at best, a Janus-faced duplicity, and at worst, the very loss of our soul'. But of course human voices are not static - they change and shift according to mood, who you are speaking to, situation. The same person can speak in different voices in the same conversation, in the same sentence. Voices shift and flicker as they fall into different stereotypes - more male, more female, more educated, less educated, the different "tongues" or manners of speaking that we hear. This is in direct contrast to machine-generated voices, which are not voices in the truest sense of the word (sound produced by the vocal organs) but rather "synthetic utterances", words that have been stitched together from clips of other words. There is a flatness to these; all nuance is lost.

Voices shift and flicker as they fall into different stereotypes

Speaking in Tongues is a project that explores the different online transcription services. It is partially run off Google's Cloud Speech API, a speech-to-text conversion platform powered by machine learning. The API returns a transcript of what it thinks it has heard and a score of how "confident" it is that the transcription is correct. Confidence is based on how close the speaker is to those in its training set, the adherence to its perceived norm. The idea of a linguistic norm is loaded as it privileges one form of speaking over another. "Correctness" is defined as per the training set that is fed into the programme. There is no open data as to how Google Cloud Speech has

Voices carry information. We can hear and interpret almost instantly someone's gender, age, education levels, social class when they speak. We know that we judge based on appearances, but we equally judge based on voice.

Zadie Smith has observed that voices are meant to be 'unchanging and singular'; that they are the true signifiers of our identity, instantly showcasing how we would like to be perceived to the listener. To use different versions, particularly in Britain, she goes on to write, represents to most people 'at best, a Janus-faced duplicity, and at worst, the very loss of our soul'. But of course human voices are not static - they change and shift according to mood, who you are speaking to, situation. The same person can speak in different voices in the same conversation, in the same sentence. Voices shift and flicker as they fall into different stereotypes - more male, more female, more educated, less educated, the different “tongues” or manners of speaking that we hear. This is in direct contrast to machine-generated voices, which are not voices in the truest sense of the word (sound produced by the vocal organs) but rather “synthetic utterances”, words that have been stitched together from clips of other words. There is a flatness to these; all nuance is lost.

Google’s Cloud Speech API, a speech-to-text conversion platform powered by machine learning, returns a transcript of what it thinks it has heard and a score of how “confident” it is that the transcription is correct. Confidence is based on how close the speaker is to those in its training set, the adherence to its perceived norm. The idea of a linguistic norm is loaded as it privileges one form of speaking over another. “Correctness” is defined as per the training set that is fed into the programme. There is no open data as to how Google Cloud Speech has been trained, but by looking at the transcriptions that are generated, one can guess that it was trained on closed-caption television (there is a proliferation of anime characters, for example, appearing in the most innocuous of phrases) but it is not clear which voices inside those programmes it has been told are “correct”. Indistinct voices, quiet voices, accented voices, voices that do not match a television norm do not currently work as well with this API. It does not have enough input data to properly map what is being said by these types of voice into anything meaningful. As the user interacts with it, they are questioned by a version of what the programme thinks it has just heard, leading to surreal, uncanny and absurd conversations as the user either tries to correct or go with the train of thought.

A streaming version of Google Cloud Speech was only recently released (November 2016) and it is still in beta. It is possible, in the terminal, to see it “think” - how it finds and then discards words as it hears more and more of a sound, creating OuLiPo-esque poetry when printed on the screen. Speaking in Tongues II is a piece that will exist only for a finite amount of time. The accuracy will improve dramatically as it gets more and more feedback from the continued use of Google products. According to Google scientists, in six months time, the imperfections will be ironed out and it will be possible to have a conversation that is perfectly repeated and mirrored back, rather than being fragmented and Beckett-like. But as the programme becomes more smooth, it also becomes easier to ignore, and as it becomes easier to ignore, it also becomes easier not to question the assumptions around correctness in accent and dialect that sit behind it.

The other side of speech-to-text: text-to-speech. In their project ‘WaveNet’, Google DeepMind are attempting to create a realistic human voice that can be generated when text is fed into it. Instead of using a database of short speech fragments that are reassembled and re-pieced together to give the mechanical “non-voice” mentioned above, WaveNet will directly model the raw waveform of the audio signal (at a rate of around 16,000 samples per second). Even at an early stage, it is impressively accurate at sounding like a human. And because these can be put together from any combination of the dataset there is the real possibility that it will create new accents and dialects that have been entirely made by technology.

Furthermore, because of the way it has been designed, WaveNet can still generate speech without text input. It makes up what to say, mimicking what it thinks people sound like to create strange babbling sounds. Speaking in tongues, in linguistic terms, is the fluid vocalising of syllables that do not have any readily understandable meaning, usually linked to a religious practice where it is believed to be an indication of the divine trying to communicate through a language that cannot hold all meaning. WaveNet, to us, at the moment, is perhaps speaking in tongues.