Speechmatics

The Speechmatics Approach to Global English.

Accent-Independent Speech Recognition.

September 2020

WHITEPAPER

Global English

One Model to Rule Them All

Historically, to get the most accurate results from speech recognition technology, specializing was key. When confronted with accents, dialects and other regional variations in speech, specialist language packs were developed to ensure reliable results.

Speech recognition is evolving and improving. More people, businesses and products are looking to use voice technology to provide better services to their customers, creating efficiencies in their organizations or augmenting their lives through the use of voice-enabled technologies. The demand for voice technology has increased and is required to serve more markets, geographies and people. While demand for language coverage is constant, it's the support of multiple accents and dialects that make up the most widely spoken languages that presents the biggest challenge.

Since their launch, virtual personal assistants (VPAs) such as Siri and Alexa have faced well-documented issues with certain accents for English language recognition, particularly Scottish and Irish. This has led to many users being forced to modify their speech patterns to be understood, adapting their voices to the technology. At Speechmatics, our any-context speech recognition engine adapts to its users no matter their accent or dialect.

By harnessing advancements in neural network architectures, and applying proprietary language training techniques, Speechmatics delivers a single English language pack supporting all major accents and dialects. Removing the need to use multiple languages packs for English dialects optimizes operational efficiencies as well as reducing the overall deployment footprint. This reduces the overhead costs for customers regardless of application or use case.

Global English not only delivers simplified deployment capabilities, but also leads the market in accuracy against English models designed for specific accents and dialects.

Global English can recognize and transcribe any audio – especially long-form audio – no matter the English accent (how words are pronounced) or dialect (a regional variant) of the language spoken.

Variations in Speech

How Speech Recognition Deals With Variations in Speech

Speech in a single language can vary according to location, group or individual idiosyncrasies, including accents, use of grammar and vocabulary.

In the extreme, these variations may prevent speakers of the same language from understanding one another and presents a significant challenge for speech recognition.

Speechmatics is the first and only company to pioneer a new approach when dealing with accent and dialect variations.

The Traditional Approach

Traditionally, speech recognition has dealt with significant variations of accents and dialects by producing different, customized language packs to ensure accuracy. Time-consuming and laborious, this process involved a whole new set of models trained on data from each particular subset of speakers.

This creates complexities for ASR vendors managing an extensive and growing number of variants for each language they support, slowing down innovation and time to market of the latest versions of their language packs.

The Speechmatics Approach

Speechmatics is the first and only company to pioneer a new approach when dealing with accent and dialect variations. We know that the way people speak and use language is very different depending on a broad range of factors including their country, specific region within a country, industry or use case.

Speechmatics is unique in our approach to the creation and operation of languages. Each language includes many attributes. The language attributes include, but are not limited to, support for accents, dialects, number of speakers and types of speech. This approach to building language packs makes it as simple as possible for our customers and partners to deploy and operate the Speechmatics solution and deliver the best possible.

Speechmatics' Global English Rationale

Rather than creating numerous English language packs for each specialist variant, we created a single, comprehensive language pack, accurately encompassing as many variations of English as possible. For most real-world applications, this gives the most reliable, accurate and efficient performance for our customers and partners. By implementing a new accent-independent approach – improving and harnessing recent advances in technology and data gathering – we are able to simplify the traditional approach, dramatically improving the accuracy and ROI, while reducing complexity and time to market. accuracy across all applications and use-cases.

Global English Competitor Comparison

Proving the Theory

To test our approach, we created the Global English language pack. We then compared its performance using test sets comprising of a number of accent and dialects against those of our competitors (see 'Figure 1: One model to rule them all').

Speechmatics' Global English language pack was the best option in every instance.

Next Level

Next-Generation Global English

Taking accuracy to the next level

Since the release of Global English in 2018, Speechmatics has continued to improve its market-leading accuracy.

Figure 1: One Model to Rule Them All

In this table we compare our Global English model with those of other providers of speech recognition for the most common English accents. Numbers represent accuracy, defined as the percentage of words correctly transcribed by the speech recognition engine.

In every case it was better to use the Speechmatics Global English (EN) language pack for transcription rather than our competitor's variant specific language packs.

Test Set Accent:	AU	CA	GB	IE	IN	NZ	US	Average
Speechmatics Global English	**97**	**98**	**92**	**92**	**91**	**94**	**97**	**94**
Microsoft	91	94	89	87	90	91	93	91
Leading Cloud Provider	91	94	86	85	86	86	93	89
Google Cloud Speech-to-text	86	94	74	75	52	81	91	79

AU – Australian, **CA** – Canadian, **GB** – British, **IE** – Irish, **IN** – Indian, **NZ** – New Zealand, **US** – American

Test sets comprised of approximately 4 hours of diverse audio and transcribed text. Accented test files included variations in gender, age and region. We know accuracy results are always dependant on the test set used. If you would like to know further details about our test set, please get in touch.

Solving the problem of audio featuring multiple speakers each with a different accent.

How to Calculate Accuracy

Accuracy is calculated by the percentage of incorrectly transcribed words within a media file (known as word error rate (WER)) subtracted from a potential 100%. WER is used as a common metric of comparing ASR providers but only considers a single and small element in the overall 'accuracy' of the transcription output. WER does not reflect the nuances of transcription such as punctuation, readability and formatting to make the output easier to consume for a human or automated process. Understanding the goals and requirements of speech processing are crucial to deliver value – not only to our customers and partners – but to the customers of their services that utilize Speechmatics' voice technology.

Real-World Benefits

For businesses with staff and customers across the world, it is not always possible or effective to select a single accent-specific language pack. Customers contacting national contact centers have a broad range of accents; call monitoring of multinational workforces must decipher numerous different forms of accented English, and live TV interviews feature guests from across the world.

ORIGINAL LINKS BROKEN

Ease of Use

This single, multi-use solution means users do not need to identify which English variant is being spoken. Solving the problem of audio featuring multiple speakers each with a different accent, or where speaker accents are not known in advance, one comprehensive language pack provides reliable results over a broader range of speakers. Global English solves the challenge of having to run audio files through the any-context speech recognition engine multiple times to capture all the different accents using accent-specific language packs.

Getting the Right Output for You

Delivering a unified language pack means that users from all over the world can use Global English to deliver the best possible accuracy with minimal operational complexity. However, it also needs to cater for users in those different regions. Global English provides the ability for users to specify rules to control the output selecting either American or British spellings.

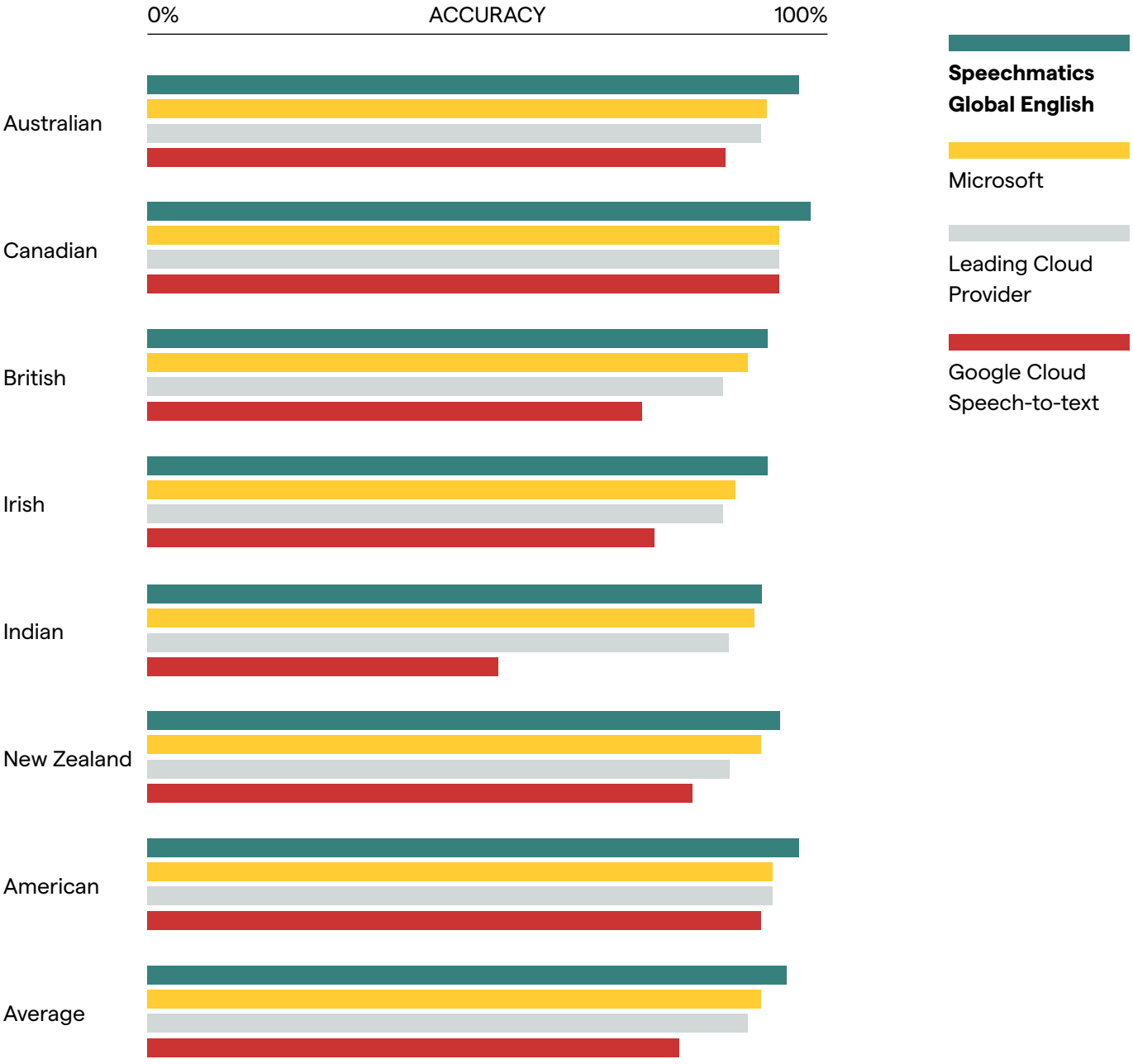
Fewer Models to Maintain and Update

By focusing resources on maintaining and updating fewer models, Speechmatics can increase quality, improve accuracy and ensure reliability of the smaller number of models we maintain. Global English always uses the same models, giving our customers and partners a consistent result.

Real World

Figure 2: Accuracy % Comparison

Figures taken from August 2020. Test sets comprised of almost 6 hours of diverse audio and transcribed text covering multiple use cases. Accented test files included variations in gender, age, region and ethnicity.



The testing in figure 1 and 2 was performed by Speechmatics using real-world audio files. This provides representative data on how Speechmatics' any-context speech recognition engine compares in terms of accuracy to other ASR vendors. Accuracy is calculated by 100 - Word Error Rate. No consideration is made to punctuation or any other factor that might contribute to whether one transcription provider is more accurate or better than another.

Speechmatics is already delivering leading levels of accuracy and investing into new ways of solving problems.

How Did We Do It?

As an industry pioneer, Speechmatics has taken advantage of recent advances in machine learning and applied proprietary language training techniques allowing this more universal, accent-independent and any-context approach to succeed where it never would have before.

Improved Algorithms

Speech recognition has advanced hugely in recent years, giving step-change improvements in a field used to marginal gains. In particular, modern neural network architectures are capable of generalizing across variations in speech by using representation learning. Deep neural networks feature multiple layers between input and output, allowing Speechmatics to filter everything but the phonetics. This effectively gives us the performance of a variety of specialized models, all in one comprehensive language pack.

Greater Computing Power

Single modern servers are more powerful than old room-filling supercomputers. This astonishing rise in computer power, coupled with re-purposing of GPUs – from playthings of gamers into serious computing machines – gives masses of computing power. This allows us to train models, based on more data, capable of supporting more variations.

More Data Available

By investing more time gathering data from a wide range of sources, Speechmatics has created a huge and diverse training corpus, allowing us to train models with a much wider range of applications than ever before.

Traditionally, speech recognition – like other machine learning and big data organizations and products – relies on huge amounts of labeled data to achieve real-world results. While this approach is good in the short to medium term, as we continue towards the future where results are getting better, it will become harder to continue to drive improvements from a mechanism of training that relies on big data alone. Small data becomes increasingly important to support precise and focused use cases as data also needs to be more representative to deliver exceptional levels of accuracy.

This brings the data problem further into scope. To reach the next stage of improvements, there will be a requirement to invest significant amounts of time to collect and label data for machine learning. While Speechmatics is already delivering leading levels of accuracy, we are investing into new ways of solving problems to enable high levels of accuracy without a growing commitment to labeling data that is not sustainable.

Future-Proofing

Speechmatics is committed to undertaking regular comparisons against other providers with frequent testing and benchmarking to ensure we provide the best automatic speech recognition on the market. By moving from multiple specialist language packs to a more comprehensive, single language pack, we have streamlined our portfolio and maximized the resources available for Global English.

Fast, accurate, reliable and now more flexible, convenient and inclusive, Global English offers users speech recognition for the future.


Speechmatics