SPEECHMATICS

THE SPEECHMATICS APPROACH TO GLOBAL ENGLISH

Accent independent speech recognition

DECEMBER 2018

Introducing Global English

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

Times have changed, and speech recognition is evolving and improving.

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 in order to make themselves understood, adapting their voices to the technology. At Speechmatics, the technology is adapting to users.

By harnessing recent advances in algorithms and neural network architectures, Speechmatics can now deliver one English language pack supporting all major accents and dialects. Removing the need to use multiple languages packs for English dialects means customers will benefit from simplified deployments as well as a reduction in the overall footprint. In turn this reduces the overhead costs for customers regardless of application or use case.

Global English can recognise and transcribe any audio – especially long-form audio – no matter the English accent or dialect spoken.

VARIATIONS IN SPEECH

How speech recognition deals with variations in speech

Speech in a single language can vary according to location, group or even 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 present a significant challenge for speech recognition.

Already world experts in the traditional approach, Speechmatics has become the first company to pioneer a new approach when dealing with accent and dialect variations specific to English.

The traditional approach

Traditionally, speech recognition has dealt with significant variations of accents and dialects by producing different, customised 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.

Historically, Speechmatics has produced North American, British and Australian versions of our English speech recognition language packs.

The traditional British language pack does indeed perform better on British-accented speech than a traditional North American language pack. However, working out how much granularity to choose became difficult. Even within a nation there are often distinct accents and different use cases with distinct vocabularies, and strong cases could be made for modelling them all.

The modern Speechmatics approach

Already world experts in the traditional approach, Speechmatics has become the first company to pioneer a new approach when dealing with accents and dialect variations specific to English. Rather than dealing with a confusing gaggle of specialist variants, for English we have now 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.

By improving and harnessing recent advances in technology and data gathering, we are able to simplify the traditional approach, dramatically improving the results and ROI.

Global English competitor comparison:

Proving the theory

To test our approach we created a 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 always the better option.

Next Generation Global English:

Taking accuracy to the next level

To prove that having one single language pack to cover all major accents and dialects for English is the best option for our customers – we tested our model against our previous specific dialect models using a test set comprising of different accents, noise levels and speaker characteristics. (see 'Figure 2: Next Generation Global English').

Based on our test results, Next Generation Global English always resulted in the best 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 – 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	ZA
Speechmatics Global English	**96%**	**98%**	**92%**	**90%**	**89%**	**94%**	**92%**	**97%**
Google Cloud Speech-to-Text	95%	94%	90%	89%	89%	93%	88%	95%
IBM Watson	87%	95%	56%	87%	78%	88%	83%	95%
Microsoft Video Indexer	88%	91%	83%	87%	76%	85%	82%	90%

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

Figures taken December 2018. Test sets comprised of approximately 4 hours of diverse audio and transcribed text. Accented test files included variations in gender, age and region.

Figure 2:

Next Generation Global English

In this table we compare our Next Generation Global English model with our previous models for the most common English accents using test sets from real world applications.

Numbers represent accuracy – the percentage of words correctly transcribed by the speech recognition engine.

Test Set Accent:	Accuracy
Speechmatics Next Generation Global English	**75%**
Speechmatics AU specific	69%
Speechmatics GB specific	62%
Speechmatics US specific	69%

AU – Australian, **GB** – British, **US** – American

Figures taken December 2018. Test sets comprise of content from a variety of domains such as news and entertainment, business meetings, financial teleconferencing and webcasts, podcasts, and journalist interviews.

As an industry pioneer, Speechmatics were able to take advantage of recent advances in the field that have allowed this more universal, generic approach to succeed where it never would have before.

Real-world benefits

For businesses with staff and customers across the country, it is not always possible or effective to select a single accent-specific language pack. Customers contacting national call centres 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.

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.

Fewer models to maintain and update
By focusing resources on maintaining and updating fewer models, we can increase quality, improve accuracy and ensure reliability of the smaller number of models we maintain.

Consistency
Global English always uses the same models, giving the customer a consistent result.

How did we do it?

As an industry pioneer, Speechmatics were able to take advantage of recent advances in the field that have allowed this more universal, generic 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 generalising across variations in speech by using representation learning. Deep neural networks feature multiple layers between input and output, allowing us to filter everything but the phonetics. This effectively gives us the performance of a variety of specialised 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 the recent repurposing of GPUs, from playthings of gamers into serious computing machines, gives masses of computing power.

This allows us to train bigger 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, we have created a huge and diverse training corpus, allowing us to train models with a much wider range of applications than ever before.

Future-proofing

Speechmatics are committed to undertaking regular comparisons against other providers with frequent testing and benchmarking, to ensure we provide the best automatic speech recognition possible. By moving from multiple specialist language packs to a more comprehensive, single language pack, we can streamline our portfolio and maximise 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.

Brookmount Court
Kirkwood Road, Cambridge
CB4 2QH United Kingdom

www.speechmatics.com