A ticklish problem

Studying accent variation from speech recordings

The book *Mr. Tickle* is a childhood favourite for many. But this fantastical story of a man with "extraordinary long arms" is helping statisticians derive a more realistic understanding of the differences in regional accents.

By **Marius A. Tirlea**, **Shahin Tavakoli**, **Davide Pigoli** and **John A. D. Aston**

Mr. Tickle has been a well-loved children's story for almost half a century and has been read countless times since 1971, the year it was first published. In that time the story has remained unchanged, but the telling of it will have sounded different depending on where it was read. From London to Liverpool to Newcastle, the words are the same – but the accent of the reader might vary considerably.

The study of dialect variation has a long and interesting history; understanding differences in vocabulary and pronunciation has possible uses in many social sciences, since the way a language is spoken changes according to social status; it is also used to shape group identification.

These differences have typically been studied by comparing phonetic transcriptions of the spoken word, rather than the spoken word itself.[1] However, with the help of *Mr. Tickle* and a database of recorded speech, statisticians are developing new methods for analysing accents, allowing for a more realistic and interactive interpretation of the way speech varies across geographic regions.

Mr. Tickle's extraordinary long reach

To demonstrate this approach, we use a data set of recorded speech from the British Library, which compiled a corpus of recordings of people from different parts of the UK reading an adaptation of *Mr. Tickle* (bit.ly/2kLnSLe). There were 154 recordings made between 2007 and 2009. The readers were all children and young adults, between the ages 12 and 23.

With these recordings, we develop methods to draw information directly from speech in an objective manner. To do this, we introduce the notion of "statistics on sounds". Instead of doing statistics on observations that are numbers or vectors, here each of our observations is a sound, represented by its so-called mel frequency cepstral coefficients (MFCCs; see "Sound as data", page 33). The goal is to create a modern version of the isogloss map, which is the traditional way in which accent changes across a geographical area are represented.

Figure 1 provides an example of an isogloss map for the word "fast", showing different pronunciations of the "a" sound in different regions of England.[2] The map is based on accent transcriptions compiled by fieldworkers in the 1950s who surveyed the speech of local residents (those studied were usually older people, native to the recording location, and with limited formal education). This extensive study, called the Survey of English Dialects, was directed by Prof. Harold Orton at the University of Leeds, and has become the authoritative

work in the area.[3] Using the transcription obtained, it has been possible to draw many maps like this one, whose boundaries define the regions where speakers pronounce a specific word or sound in a certain way.

However, there are several limitations with these kinds of maps. Primarily, the process of transcription of how the word or vowel is pronounced is subjective: it is entirely dependent on how the fieldworker writes down the accents they hear, and there is no guarantee that two fieldworkers hearing the same person would transcribe the speech in the same way, despite the training they have received. In addition, the reported variation is necessarily too little within each region of an isogloss map, and too sharp across boundaries. This happens because the number of distinct regions must be low to ease interpretation. But it gives the misleading impression that speakers from a relatively large geographical area are identical in their speech, and that individuals on opposite sides of artificial boundaries differ drastically in their pronunciation.

"Statistics on sound" offers an improvement by making it possible to consider concepts such as sound variations over a geographical continuum, as well as sound interpolation – that is, determining what an accent might sound like in a place without an associated recording. When data at a geographic location is available, it is also possible to use it to draw continuous maps that reflect accent variations.

The result should be a more accurate estimate of the way in which pronunciation varies in a country such as England. Moreover, by working directly on the sound process, we are able to produce sounds as outputs of our analysis, which can then be interpreted by linguists.

The *Mr. Tickle* recordings provide the data we need to achieve all this, but it is not a perfect data set. For each recording, the age of the speaker and the place of recording are provided. But in the absence of further information about the speakers' background, we must assume that the accent of each individual is representative of the geographical location where the recording was taken. In addition, while the reach of *Mr. Tickle* recordings covers much of England, with a few locations in the rest of the UK, the number of speakers at each location varies considerably, ranging from locations with only two speakers, to one location with 18 recordings. Furthermore, there is a noticeable concentration of locations near London.

Finally, as these recordings were not made under laboratory conditions, they present some relevant background noise, and each person has their own speaking rhythm, and the fluency of speech varies between recordings.

For this analysis, a model was constructed for the word "fast", since this word contains only one vowel sound, "a", the pronunciation of which varies significantly across England. In the data set considered, there are 139 recordings of the word "fast" available for use in construction of accent variation maps, and Figure 2 shows how the recordings were distributed between locations.

The sounds of the words were transformed to MFCCs – a time-indexed vector representation of the sound that is useful for statistical analysis – which were then smoothed and

Marius Tirlea is a third-year mathematics undergraduate student at Trinity College, Cambridge

Shahin Tavakoli is Research Fellow at the Statistical Laboratory, University of Cambridge and Darwin College, Cambridge

Davide Pigoli is Research Associate at the Statistical Laboratory, University of Cambridge and Wolfson College, Cambridge

John Aston is Professor of Statistics at the Statistics Laboratory, University of Cambridge and a Trustee of the Alan Turing Institute

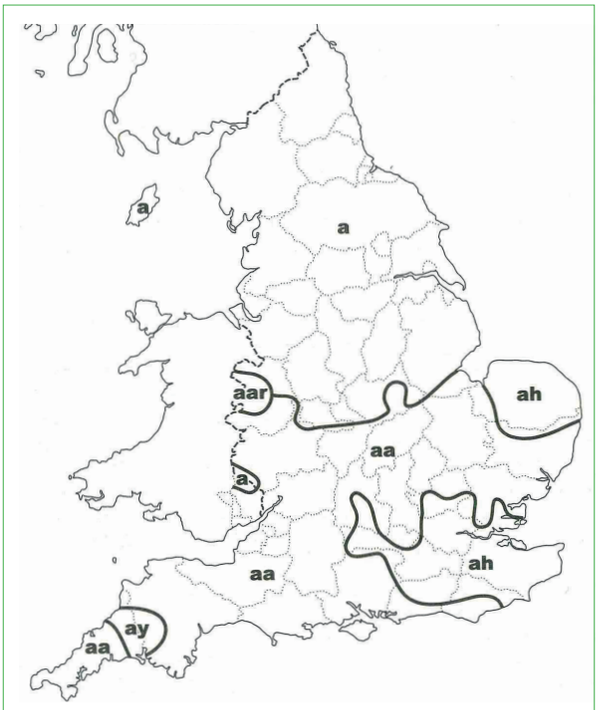

FIGURE 1 Example of isoglosses for the word "fast" in England, from *An Atlas of English Dialects*.[2] Republished by permission of Oxford University Press

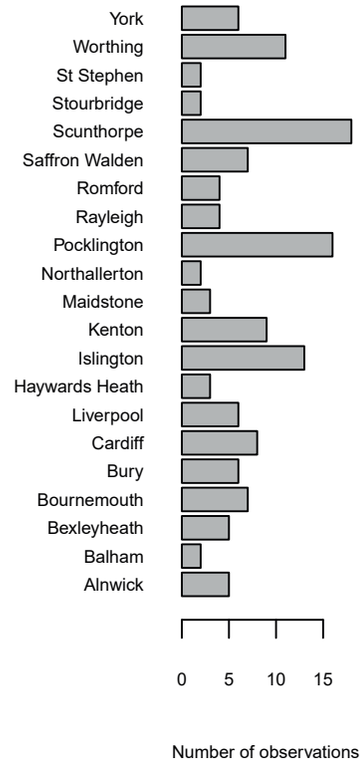

FIGURE 2 Number of recordings of the word "fast" in each location in England, drawn from the British Library's archive of *Mr. Tickle* recordings

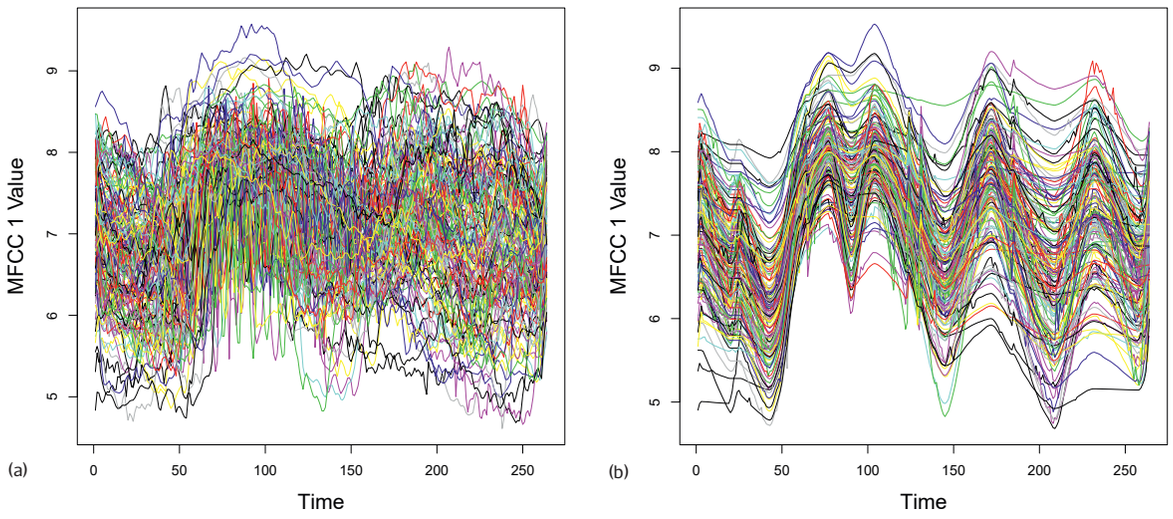

FIGURE 3 First mel frequency cepstral coefficient for the vowel in the word "fast", before and after preprocessing: (a) raw first MFCC; (b) smoothed and aligned first MFCC

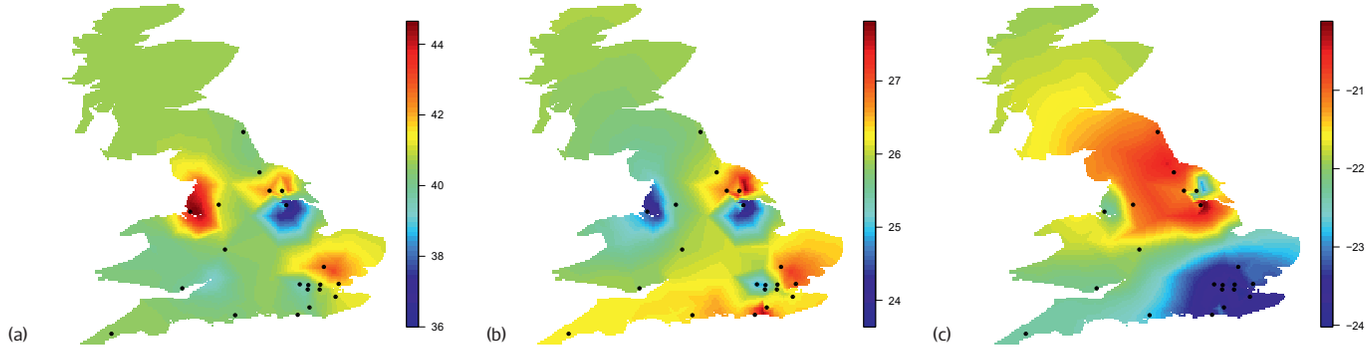

FIGURE 4 Plots of the triangulated grid over England used for the analysis (see "Statistical model") and of the projection of estimated mean sound fields into the first three principal components. Dots represent the locations of the available *Mr. Tickle* recordings. (a) First principal component score field; (b) Second principal component score field; (c) Third principal component score field

time-aligned to account for differences between speakers' pronunciation speeds (see Figure 3). Indeed, you can imagine that two different speakers may be reading faster or slower from the *Mr. Tickle* adaptation which makes point-to-point comparison in real time meaningless. We need to "play" the recordings at different (and not constant) speeds so that the different parts of the word (and, in particular, the vowel we are interested in) are matched across the speakers. In this way, we can find out the portion of the MFCCs corresponding to the vowel sound. These segments of the MFCCs provide the data to explore how the pronunciation of the vowel in "fast" varies across England.

"Statistics on sound" makes it possible to consider sound variations over a geographical continuum

Maps of accents

We chose to model the geographical variation of the vowel sound using a *mean field plus error* model (see "Statistical model"). This is essentially a non-parametric regression model, which assumes that at each location in England there is an average pronunciation (a stereotypical accent), which changes smoothly as we move across the country. This *mean sound field* can therefore be used to produce accent maps of England.

However, producing such maps is not straightforward, as sounds are complex high-dimensional objects. There is, in particular, no single accent map that one can create to convey the complex reality of accent change. But it is possible to create a series of maps reflecting different modes of accentual change, for instance by using dimension reduction techniques such as principal component analysis (see Figure 4). The maps that are produced depend on the choice of how smooth the transition from one accent to another is. We can investigate this by visual inspection of the output maps, and by listening to the quality of the reconstructed sounds over a range of values of the (spatial) smoothness parameters.

The third principal component direction, shown in Figure 4(c), is particularly easy to recognise as it is very similar to the accent regions shown in Figure 1. Indeed, it is known that there is a contrast between the pronunciation of the "a" sound in northern and southern accents: in the North, "fast" has a short, open front vowel as in "pat", whereas in the South and South-East it has a long back vowel ("aah"), similar to the vowel in "part". It is possible to recognise this behaviour in the principal component maps in Figure 4, while at the same time we see a continuous transition between the two regions, which makes it easier to appreciate accent variability as it really occurs. It is also possible to listen to the reconstructed sounds on the map (bit.ly/2mS8iOh) and explore directly the transition in pronunciation.

Taking this forward

We have outlined a way in which isogloss maps can be enhanced, using statistical methods to update the traditional issues associated with the representation of accent variation across a geographical area. The use of the British Library's *Mr. Tickle* recordings shows how this can be achieved even with noisy, crowdsourced data.

For the analysis to be conclusive on dialect features, however, the data must be representative of the linguistic characteristics of the population; this is not necessarily the case for the *Mr. Tickle* data, which includes only children and young adult speakers. Therefore, we cannot claim that the maps produced describe well the speech pattern of the general population. *Mr. Tickle*, rather, gives an exemplar of the type of analysis that it is possible to perform.

The same approach can be applied to more comprehensive linguistics corpora, such as the British National Corpus, with a preliminary investigation of this data in Tavakoli *et al.*[4] Moreover, speakers' accents do not solely depend on their geographical position, but also on socio-economic factors, and this needs to be taken into account when moving to richer data sets. For example, different maps can be plotted for different socio-economic classes, or for males and females.

Applications of the method outlined here are not restricted to dialect studies; for example, one can apply the same ideas to quantify variation between similar phonetic structures in different languages with the same common root (see bit.ly/2kLOfAD for more details), giving us the chance to really understand our linguistic history and diversity simply from the sounds we all use every day. ■

Sound as data

In order to perform statistics on sounds, a proper encoding of the sound is necessary. We choose here to represent sounds in a time-frequency domain and align them in time to account for individual variation in speaking rate. The time-frequency representation we choose is the mel frequency cepstral coefficient (MFCC) representation, because it provides a principled lower-dimensional representation of sounds, and it works well for speech resynthesis.

A typical MFCC transformation consists of the following:

1. Perform a discrete Fourier transform on the (sampled) waveform of the audio recording.
2. Convert the resulting frequency spectrum to the mel scale, where the mel value, m, is calculated from the frequency, f, by

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$

   This is done using overlapping, non-linearly spaced, triangular windows.
3. Take the logarithms of the resulting mel values, and then take the discrete cosine transform of these to obtain the MFCCs.

There exist many modifications and variations of this procedure for MFCC synthesis in the literature, as authors seek incremental improvements in the performance of implemented speech recognition or parametric speech synthesis systems. We use the MFCC proposed in Erro *et al.* as it yields high-quality, natural sounding resynthesised speech.[5] However, the underlying principles are the same.

MFCCs also have the property that their construction mimics the fashion in which humans perceive sound.[6] This further emphasises the suitability of using MFCCs to construct the model, since one of our goals was to be able to synthesise and play back words using our model.

Statistical model

We model the MFCCs time-dependent vector $Y_{ij} \in \mathbb{R}^p$, corresponding to the jth speaker at position X_j, by the following mean plus error model:

$$Y_{ij}(t) = \mu(X_j, t) + \varepsilon_{ij}(t)$$

In other words, we have some mean field μ which is an unknown function of space and time, along with a spatially independent random error ε, which affects individual speakers' properties of the MFCCs. In this case, the mean μ is the "average" form of the relevant MFCCs at a given location. Our expectation is that the estimates for $\mu(x, .)$ will contain the relevant information about the dialect feature at location x in the domain of interest.

We define first a triangular mesh over England, as can be seen above. We then estimate the value of μ at each node of the mesh by means of a Nadaraya–Watson estimator, which is a locally weighted average, where the weights are given by an isotropic Gaussian kernel, and the distance between nodes is taken to be the graph distance.

References

1. Kretszchmar, W. A. (1996) Quantitative areal analysis of dialect features. *Language Variation and Change*, **8**(1), 13–39.

2. Upton, C. and Widdowson, J. D. A. (1996) *An Atlas of English Dialects*. Oxford: Oxford University Press.

3. Orton, H. (1962) *Survey of English Dialects: An Introduction*. Leeds: E. J. Arnold for the University of Leeds.

4. Tavakoli, S., Pigoli, D., Aston, J. A. and Coleman, J. (2016) Spatial modeling of object data: Analysing dialect sound variations across the UK. arXiv:1610.10040 [stat.ME].

5. Erro, D., Sainz, I., Navas, E. and Hernaez, I. (2014) Harmonics plus noise model based vocoder for statistical parametric speech synthesis. *IEEE Journal on Selected Topics in Signal Processing*, **8**(2), 184–194.

6. Huang, X., Acero, A. and Hon, H.-W. (2001) *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Upper Saddle River, NJ: Prentice Hall PTR.