Visualising vocalic variability in space and time – an automatic exploration of "found data"

Malisz Zofia, Jenny Öqvist**, Per Fallgren*, Jens Edlund* , and David House* (*Department of Speech, Music and Hearing, KTH in Stockholm, **Institutet för språk och folkminnen in Stockholm)*

State-funded cultural heritage archives harbour vast amounts of speech data. This type of data has typically been collected as historical records, ethnographic research, dialect studies and documentation of popular culture. In speech technology, this is sometimes referred to as found data, in order to contrast it with data that has been recorded expressly for purposes of speech analysis.

The broader aim of the present work is to adapt and develop existing speech technology methods to facilitate the accessibility of found data to social sciences and humanities research (SSH). More specifically, our SSH partners are currently interested in describing the diachronic and dialectal changes of specific vowels in the Stockholm area using found data. Towards this aim, we process resources housed by the Swedish Institute of Language and Folklore – an archive of ca. 13 thousand hours of digitised speech recorded over close to a century.

We tackle two challenges related to this task: how to enable efficient browsing through several hundreds of thousands of vowel tokens and how to overcome the problem that formant tracking is notoriously unreliable.

We use either the t-SNE (van der Maaten and Hinton, 2008) or the SOM (Kohonen, 1998) method in a similar manner as the t-SNE has recently been used to visualise and sonify bird calls (Google AIExperiment by Mann et al.). Unlike formant analysis, we do not extract any acoustic parameters from the vocalic segments but use greyscale spectrogram images of the vowels and submit those to t-SNE for dimensionality reduction and clustering.

The method is likely to cluster those speech sounds separately that have mistakenly been identified as vowels and to put vowels of similar quality close to each other. With this, we hope to achieve a more robust method of organising very large quantities of data of less-than-perfect acoustic quality.

Manual exploration of the structured representation of unruly data subsets, will allow us to take the next step and meaningfully cast vocalic variability onto geographical maps. However, the proposed techniques have been shown to be less useful for diachronic and geographical visual representations (Juuso and Kretzschmar 2016). For this step, we instead aim to implement a cellular automaton (ibid).

References

L.J.P. van der Maaten and G.E. Hinton (2008). Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research* 9(Nov):2579-2605.

Kohonen, T. (1998). The self-organizing map. *Neurocomputing*, *21*(1), 1-6.

McDonald, K., Tan, M., Mann, Y. https://aiexperiments.withgoogle.com/bird-sounds

Juuso, I., & Kretzschmar Jr, W. A. (2016). Creation of Regions for Dialect Features Using a Cellular Automaton. *Journal of English Linguistics*, *44*(1), 4-33.