Presentation Abstract

L1 Identification from L2 Speech Using Neural Spectrogram Analysis

Calbert Graham, Phonetics Laboratory, University of Cambridge

English has become the most widely spoken language globally with the vast majority of its speakers using it as a second language (L2). It is well-known that the characteristic features of these different varieties of English are highly influenced by the speakers' native languages (L1s). Understanding the speech features that contribute to the foreign-accentedness of a speaker's L2 English may be useful in foreign language learning (e.g. in pronunciation remediation systems) and in forensic speaker profiling (e.g. by helping an investigator to narrow down the scope of an investigation).

The main objective of this project is to model L1-L2 interaction and uncover discriminative speech features that can identify the L1 background of a speaker from their non-native English speech. In modelling L1-L2 interaction, traditional phonetic analyses tend to measure the similarity of an L2 speaker's production (of specific phonemes or prosodic units) as compared to that of a native speaker, based on a pre-selected set of acoustic features. However, apart from being time and expertise consuming, the set of extracted features may not be sufficient to capture all the traces of the L1 in the L2 speech that are needed to make an accurate classification. Deep learning has the potential to address this issue by exploring the space of features automatically.

In this talk I will report a series of classification experiments involving a deep convolutional neural network (CNN) based on spectrogram pictures. The classification problem consists of determining whether English speech samples from a large spontaneous speech corpus are spoken by a native speaker of SSBE, Japanese, Dutch, French or Polish.

The input to the CNN are spectrogram images extracted from 30-second speech samples. In order to make the features more transparent and therefore interpretable by phoneticians, the experiment also compares accuracy rates in training the classifiers on (1) spectrogram pictures of phonetically segmented vocalic, consonantal and inter-segmental intervals vs. on (2) spectrogram pictures without any explicit phonetic segmentation (i.e. extracted at fixed time intervals).

Overall, results showed that the system can identify the 5 English varieties with a high level of accuracy based on spectrogram pictures. Findings also suggest that although spectrogram images without phonetic segmentation have the highest level of accuracy in the experiments, training the classifiers on certain combinations of phonetically modelled spectrogram images can produce results with comparable accuracy rates.

Our preliminary conclusions are that:

1. Spectrogram images contain a wide of range of information to successfully trace the L1 background of speakers when they speak in their L2, which makes this approach superior to traditional feature-extraction methods.
2. Unlike traditional phonetic approaches, deep learning based on spectrogram images lacks transparency and is therefore difficult to interpret.
3. However, an integrative approach that combines deep learning with phonetic modelling (to make the source of the discriminating features more transparent) can potentially be very useful in phonetic research.