Automatic detection of sociolinguistic variation in forced-alignment

Abstract

The emergence of forced alignment and automatic vowel extraction is arguably one of the most important methodological advances in modern-day sociolinguistics, particularly with the current trend of employing 'big data' on an unprecedented scale (Fruehwald 2015). Forced alignment software time-aligns orthographic and phone-level transcriptions with a corresponding audio file, which facilitates a more efficient and more reliable analysis of linguistic variation. This study investigates the possibility of using one such tool, FAVE-align (Rosenfelder et al. 2011) to fully automate the coding of three variable rules, namely: (th)-fronting, (td)-deletion, and (h)-dropping.

The methodology employed here mirrors that of Yuan & Liberman (2011) and Milne (2014), in that it involves expansion of a standard pronunciation dictionary to include multiple phonemic transcriptions of single lexical entries; each extra entry reflects the surface output of a variable phonological process. When encountering words with multiple dictionary entries, the aligner selects the most appropriate transcription based on how closely the competing acoustic models fit the observed speech signal.

The accuracy of FAVE's variant discrimination is evaluated by comparing its results to manually-coded human judgements, as well as inter-transcriber agreement rates, on an hour-long sociolinguistic interview conducted with a native British English speaker from Manchester in the north of England. The results provide encouraging evidence for this innovative method of token-coding; tokens of (h) are coded with 85.54% accuracy, comparable with the accuracy rates for the voiced segments of (th) (82.22%) and (td) (86.49%). Interestingly, the discriminative performance of FAVE struggles most with the voiceless segments involved in these latter two processes (/θ/ and /t/, respectively); these errors are largely false positives, where FAVE has incorrectly coded for application of the phonological rule. It is unsurprising that FAVE would mistake the voiceless, lenited quality of word-final /t/ in consonant clusters as nothing more than silence or faint background noise. A comparably low accuracy is found for the presence of /θ/, which FAVE often mistakes for /f/, though the fact that a similar rate of discrepancy is shown between human transcribers suggests that these errors shouldn't necessarily be attributed to FAVE's performance, but rather to the subtle nature of this alternation for this particular speaker.

There is also evidence that this method of automation suffers in faster speech rates, which is an important consideration given the messy nature of sociolinguistic interview data and how speech rate can vary dramatically within a single conversation; crucially, this is not found to have any inhibiting effect on the performance of human coders and the agreement rates between them.

The accuracy rates reported here are promising, and have methodological implications for coding sociolinguistic interviews on a large scale. Future developments could further reduce the degree of error by self-training new, speaker-specific acoustic models. More advanced systems could also seek to implement some computational pseudo-phonology, that is, to try modelling a phonological system, including all stochastic processes, so that the aligner tests all possible surface realisations before settling on the closest match to the observed speech signal.

References

Fruehwald, J. 2015. Big data and sociolinguistics. Presented at *Penn Linguistics Conference 39,* 19 March 2015. Available at: <https://jofrhwld.github.io/italk/big_data.html>.

Milne, P. 2014. *The variable pronunciations of word-final consonant clusters in a force aligned corpus of spoken French*. University of Ottawa dissertation.

Rosenfelder, I., J. Fruehwald, K. Evanini & Y. J. Yuan. 2011. *FAVE (Forced Alignment and Vowel Extraction) Program Suite*. Available at: <http://fave.ling.upenn.edu>.

Yuan, J. & M. Liberman. 2011. Automatic detection of "g-dropping" in American English using forced alignment. In *Proceedings of 2011 IEEE Automatic Speech Recognition and Understanding Workshop,* 490-493.