Automated assessment of non-native speech using vowel formant features
Calbert Graham, Francis Nolan, Andrew Caines, Paula Buttery

ALTA Institute, Phonetics Lab
DTAL - University of Cambridge

The traditional approach of using trained human assessors to evaluate spoken language proficiency has proven to be both expensive and time-consuming. This has resulted in considerable interest in finding ways to automate the assessment of non-native speech. Toward this goal, a large number of studies have investigated the use of ASR technology to assess L2 speaking proficiency automatically (e.g. Witt, 1999; Mostow et al., 1994; Neumeyer et al., 2000; Hönig et al., 2010, among many others). However, the vast majority of these studies focus on speaking rate, likelihood-based pronunciation features, and so on, without sufficient attention to features derived from acoustic phonetic measurements (Chen et al., 2010). In the case of vowels, duration has been used as a measure of fluency, but this has proven to be an inadequate measure in the automatic assessment of non-native speech mainly because it fails to utilise important perceptual cues used by humans to judge pronunciation. Research suggests that assessing the vowel space of a speaker can provide a perceptually valid index of speech intelligibility (e.g. Ferguson and Kewley-Port, 2002). When applied to non-native speech acquisition, a metric based on vowel characteristics could be a useful, relatively easy way to provide feedback within, for instance, a computer-aided language learning (CALL) context.

In this paper we explore the possibility of using vowel formant features for automated assessment of second language pronunciation. The pilot study is being carried out as part of a wider research initiative involving researchers from several areas of expertise (engineering, computing, linguistics) who are working toward developing techniques to automatically evaluate non-native oral and written English proficiency skills. Vowel space features from read and spontaneous speech in a test of business English (Gujarati L1 speakers) were rank correlated with pronunciation scores provided by expert judges. The data was transcribed orthographically via crowdsourcing, then automatically segmented and aligned using an HTK-based algorithm to determine word and phone boundaries. Praat (Boersma & Weenink, 2014) was used to extract F1 and F2 measures automatically at the midpoint of the following target SSBE vowels which appeared as full vowels in stressed syllables of content words: /ɪ:/, /ɨ/, /ɜ:/ /ʊ/, /ɻ/, /ɛ/, /ɛ/, /ʌ/. Unlike in SSBE, Gujarati does not distinguish between /ɪ:/ & /ɨ/ or between /ɜ:/ & /ʊ/; similarly, in Gujarati /ɛ/ & /ɛ/ are realised as the phoneme /ɛ/, and /ɜ:/ & /ʊ/ as /ʊ/. F1 and F2 formant measures were expressed on the ERB-rate scale and normalised using the Lobanov procedure (1971). We then calculated the Euclidean distance ratio for each speaker (i.e. the distance of a vowel to other 'anchor' vowels produced by a speaker).

The difference between target vowel pairs (e.g. between /ɪ:/ and /ɨ/) was shown to correlate with pronunciation scores such that lower proficiency level speakers did not appear to make a distinction between the counterparts in a vowel pair (Fig. 1) - a possible transfer effect of their L1, whilst the more proficient speakers were making a clear distinction (Fig 2). The finding appears to support the use of vowel space characteristics as a metric in the automated assessment of non-native speech and underscores the importance of taking L1-specific characteristics into account. The work is being extended to include L1 Gujarati data as well as more speakers of different L1 backgrounds.
Fig. 1 Mean values of F1 and F2 for L1 Gujarati (Basic English) speakers for target vowels at vowel midpoint (the 50% temporal position).

Fig. 2 Mean values of F1 and F2 for L1 Gujarati (advanced English) speakers for target vowels at vowel midpoint (the 50% temporal position).

References