INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Vowel characteristics in the assessment of L2 English pronunciation

Calbert Graham, Paula Buttery, Francis Nolan

ALTA Institute
Department of Theoretical and Applied Linguistics, University of Cambridge, UK
crg29@cam.ac.uk, pjb48@cam.ac.uk, fjn1@cam.ac.uk

Abstract

There is considerable need to utilise linguistically meaningful measures of second language (L2) proficiency that are based on perceptual cues used by humans to assess pronunciation. Previous research on non-native acquisition of vowel systems suggests a strong link between vowel production accuracy and speech intelligibility. It is well known that the acoustic and perceptual identification of vowels rely on formant frequencies. However, formant analysis may not be viable in large-scale corpus research, given the need for manual correction of tracking errors. Spectral analysis techniques have been shown to be a robust alternative to formant tracking. This paper explores the use of one such technique – the discrete cosine transform (DCT) – for modelling English vowel spectra in the productions of non-native English speakers. Mel-scaled DCT coefficients were calculated over a frequency band of 200-4000 Hz. Results show a statistically significant correlation between coefficients and the proficiency level of speakers, and suggest that this technique holds some promise in automated L2 pronunciation teaching and assessment.

Index Terms: DCT coefficients, automated assessment, vowel, acquisition

1. Introduction

There is considerable interest in developing automated assessment systems as a more resource-efficient alternative to the traditional approach of using human assessors to evaluate second language (L2) proficiency. Computer-assisted language learning (CALL) systems provide the user with the freedom to self-regulate their learning in a way that is not possible in a traditional classroom setting. In CALL systems, the computer acts interactively as an aid in the presentation and assessment of the material to be learned. However, despite advances in the development of CALL systems, pronunciation assessment and teaching continue to lag behind other components of language competence (e.g. writing and listening). One potential reason for this is the limitation on such systems to process and evaluate oral responses within a computer interface [4].

The vast majority of studies on automated assessment of speaking proficiency examine the use of ASR technology (e.g. [4], [5], [6], [7], among others). However, given the multiplicity of factors contributing to the quality of oral proficiency, it is a great challenge deriving useful metrics to quantify assessment. The traditional ASR approach focuses on speaking rate, likelihood-based pronunciation features, and so on, arguably without sufficient attention to features derived from acoustic phonetic measurements [8]). Acoustic phonetic features may contain important perceptual cues that are used by humans to judge pronunciation that may be useful in L2 speech assessment. It is evident from a review of the existing literature that there is much need to develop empirically grounded criteria that link assessment metrics to concrete linguistic features that can be made explicit to the L2 learner. Such an approach would also enable CALL system developers to provide relevant feedback to learners who may want to improve their proficiency (e.g. to progress from the basic A level to a more advanced C level on the Common European Framework of References for languages (CEFR) assessment scale). One such concrete linguistic feature is vowel quality, which will be the focus of this paper.

Research suggests a relationship between vowel production accuracy and L2 speech intelligibility, with various effects of the L1 vowel system (e.g. [1], [8]). Thus, assessing vowel production of a speaker can provide a perceptually valid index of speech intelligibility [22], and when applied to non-native speech acquisition a metric based on vowel characteristics could be a relatively easy way to provide feedback within a CALL context.

It is well known that formant frequencies contain the primary information for the perceptual distinction of vowels [3]. Recent studies that have examined the use of formant characteristics in the automated assessment of speech intelligibility (e.g. [9]; [10]) have established a link between proficiency and vowel formant characteristics. It is regrettable, however, that no serious attempt has been made in the vast majority of these studies to normalise speaker differences in vocal tract sizes. This makes it extremely difficult to interpret or generalise from the reported findings. What is needed, therefore, is a robust measure that normalises speaker variation due to physiological or anatomical differences while leaving phonetic variation intact.

[11] applied a normalisation procedure in modelling formant frequencies as a metric in automated pronunciation assessment. However, these authors conceded that although formant analysis has proven to be a useful metric when applied to good quality data, it may not be a viable measure in automated assessment given the need to manually correct tracking errors. Critical band analysis has been proposed as a vowel identification technique when formant tracking is problematic [2] due to formant tracking error or poor data quality.

The present study seeks to explore the use of one such technique – the discrete coefficient transformation (DCT; [3]) in the automated assessment of L2 pronunciation. Furthermore, we examine the effects of two normalisation procedures: Mel-scale warping and the Lobanov normalisation. Finally, we also explore the potential effect of native language setting in the phonetic realisation of English

http://dx.doi.org/10.21437/Interspeech.2016-1630

vowels by non-native speakers. Specifically, this study addresses three primary research questions:

1. Using DCT coefficients as a measure: how do the spectral characteristics of speakers vary as a function of proficiency?
2. What is the effect of applying an extrinsic normalisation technique (i.e. the Labanov normalisaton)?
3. Gujarati speakers do not make a distinction between tense and lax vowels (/i:/ vs. /ɪ/ and /u:/ vs. /ʊ/, whilst Thai speakers do, but in a way that gives primacy to duration over vowel quality cues, unlike English. We ask: is there an influence of the L1 vowel settings of non-native speakers on their L2 pronunciation?

2. Method

2.1. Speakers

80 speakers (age range 20-35 years) whose native language was either Gujarati (30 females 18 males) or Standard Thai (18 females, 14 males) were randomly selected from a larger pilot dataset. Based on the judgement of several expert graders, the speakers were placed into 5 proficiency levels according to the CEFR: A1 (15 speakers); A2 (15 speakers); B1 (21 speakers); B2 (15); C1 (14 speakers). There were no speakers at the C2 level in English.

2.2. Dataset

The dataset was from a Cambridge English BULATS test of business English comprising elicited spontaneous speech (in the form of a short bio and a monologue testing the business knowledge of the candidate). The data was recorded in BULATS testing centres in Gujarat, India and in Bangkok, Thailand (at 44.1 KHz sampling rate and a 16-bit resolution). In total, there were 1300 recordings in the dataset used in this study.

2.3. Analysis

2.3.1. Data processing

The data were orthographically transcribed using multiple crowd-sourcers and a speech recogniser according to the procedure described in [12]. The transcribed data were then automatically segmented and aligned using an HTK-based algorithm to determine word and phone boundaries. All data analysis was carried out in EMU-R [3], which has a two-way interactive interface to Praat ([14]) for converting annotations and R for signal processing (see [3, pg. 39] for more on this.) To avoid the effects of vowel centralisation due to lack of stress and the influence of /r/ consonants, which would affect feature values, only vowel tokens that appeared as full vowels in stressed syllables of content words and which did not precede an /r/ were included in the analysis reported in this study. As we aimed to explore the possibility of using vowel characteristics in automated assessment, no manual corrections were made to the phone-aligned data and no further contexts were specified.

2.3.2. Discrete Cosine Transform

The discrete transform (DCT) is a transformation that decomposes a signal into a set of coefficients at half cycles (k=0, 0.5, 1…1/2(N-1)) and provides numerical correlates for trajectory shape. When applied to the spectrum, it has been shown that the DCT coefficients are equivalent to cepstral coefficients [23].

The DCT decorrelates the spectral coefficients and allows them to be modelled with diagonal Gaussian distributions. The number of parameters needed to represent a frame of speech is significantly reduced, which in turn reduces memory and computation requirements. In spectral analysis, it has been shown that the first three coefficients (C0, C1, C2) are proportional to the mean, linear slope and curvature of the signal respectively [13]. In the present study, we converted the speech data to EMU format, then warped the frequency axis to the Mel scale and calculated DCT coefficients for vowel targets over a frequency band of 200Hz up to and including half the Nyquist rate of 8000Hz. This was done in Emu-R (according to the procedure described in [3]). The formula for the DCT used in the study was according to [15]: $C_m \cos(\theta)$ is the cosine function to model the trajectories; for an N-point Mel spectrum, x(n), extending in frequency from n=0 to N−1 points, the mth DCT-coefficient C_m (m=0,1,2) was calculated with DCT form:

$$C_m = \frac{2k_m}{N} \sum_{n=0}^{N-1} x(n) \cos\left(\frac{(2N+1)\, m\pi}{2N}\right)$$

Where k_m is $\frac{1}{\sqrt{2}}$ when m=1, and 1 when m ≠1.

The following vowels were chosen to give a good overall coverage of the vowel space: ae [æ], eh [ɛ], ih [ɪ], iy [i:], oh [ɒ], uh [ʊ], uw [u:]. Coefficients C1 and C2 equate to F1-F2 for vowels; thus we were able to derive a vowel discriminant of averaged positions similarly to the F1xF2 vowel plane. It is worthwhile noting that there are well-established clustering methods that can be applied using a probabilistic framework. However, these are most relevant in contexts where there is no regard for the phonetic interpretation of the data. The quality of tense-lax vowels is highly dependent on F1-F2 settings and the DCT method, as a F1-F2 parameterisation, takes dynamic changes in the target vowels into account. Similar features have been used in previous studies (e.g. [23]), although not for non-native vowel production data.

2.3.3. Log Euclidean distance ratio

One way of quantifying vowel-space differences between speakers is to measure the Euclidean distance between a vowel and the centre of the vowel space. Whereas the Lobanov normalisation is applied to remove speaker variation (due to sex and age, etc.), the log Euclidean distance ratio, which is calculated per speaker as the distance of a vowel to other vowels they produce, is generally applied in sociolinguistics research e.g. [18] to address a specific hypotheses about, for example, coarticulatory effects and vowel shifts. In the study reported in this paper, it is applied specifically to determine whether Gujarati and Thai speakers at different CEFR levels in English were making a distinction between the tense and lax vowel pairs. We hypothesised that the Gujarati speakers will have difficulty differentiating these vowel pairs, given that they do not make such a distinction in their native language. For the Thai speakers, we hypothesised that they would likely not have such a difficulty as tense-lax distinction exists in their L1, notwithstanding the fact that they tend to rely more on duration cues than on vowel quality [24], whereas English speakers primarily use vowel quality differences. The

following form where x, y and a, b represent vowel spectra vectors (DCT coefficients), gives the Euclidean distance: $distance\ ((x,y),(a,b)) = (x-a)^2 + (y-b)^2$.

3. Results

Out of the 80 speakers, the datasets for three of the Gujarati females were rejected as there were no vowel tokens in their speech (there were only silent pauses and other non-linguistic noises.) The 77 speakers produced vowel tokens as follows: ae (272), eh (308), ih (487) iy (460), oh (409), uh (44), uw (296). /uh/ was produced by only some speakers (mostly at the C1 level) and therefore excluded from further analysis.

3.1. Mel-scaled DCT coefficients

In order to address RQ1 whether the CEFR level of the speakers correlates with their vowel production (measured by MEL-scaled DCT coefficients) we calculated the Spearman rank order correlation coefficient between CEFR levels and the DCT-coefficients.

Overall, the results indicated a statistically significant correlation between CEFR level and the two DCT measures for six of the vowels, as shown in the vowel space lattice in Figure 1.

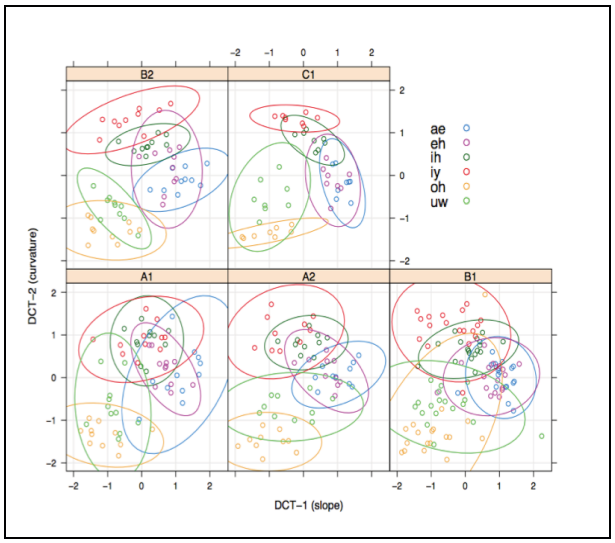

Figure 1: *Vowel space by CEFR level.*

For the vowel /ae/, for example, the Spearman's rho revealed a statistically significant relationship between CEFR level and DCT measures: DCT-1 (r_s=.50, p<.01), DCT-2 (r_s=.72, p<.01). The results were generally the same for the Mel-scaled transformed data and the Lobanov transformed data, although the latter was less correlated with gender (r_s=.18, n.s.) than the former (r_s=.25, n.s.).

We further conducted a mixed effects model with DCT coefficient (two levels: DCT-1 and DCT-2) as the dependent variable and CEFR level (5 levels) and native language (two levels: Thai vs. Gujarati) as the random factors. The results showed a main effect of CEFR level on both coefficient measures (DCT-1: F=(20, 230) = 3.97, p<=.001; DCT-2: F(20, 230) = 5.94, p<.001). There was no main effect of native language. Bonferroni posthoc tests revealed that on both dimensions the three lowest CEFR level speakers (A1, A2, B1) were different from the B2 and C1 speakers (all at p<.001). B2 and C1 speakers realised a significantly larger distance between the two measures (suggesting a wider vowel space). This finding (depicted in Figure 2 below) is also evident in the previous analysis (Figure 1) which shows the separation between vowels becoming much more distinct as the B2 level and C1 levels.

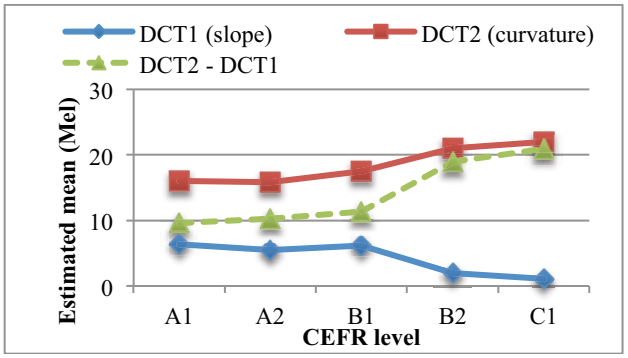

Figure 2: *Estimated DCT coefficient mean by CEFR level.*

3.2. Log distance ratio

For space reasons, only the results for /i:/ vs. /ɪ/ will be presented here (however, note that similar results were obtained for /u:/ vs. /ʊ/.) The result of a mixed model with Euclidean distance as the dependent variable and proficiency and native language as the random factors showed a significant difference in the distance from /i:/ to the vowel centre according to proficiency (F(1,28)=4.028, p<.01), but not according to native language. Bonferroni post hoc tests revealed that, with the exception of the C1 level, all groups failed to realise a statistically significant difference between the tense-lax vowel pair. The overall patterns are given in Figure 3.

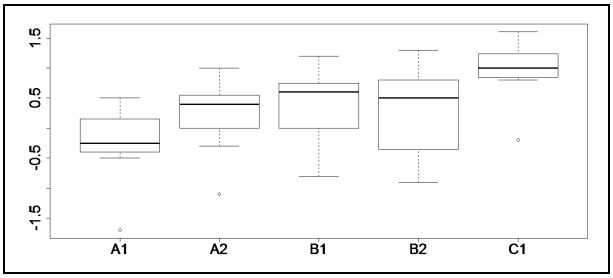

Figure 3: *Log Euclidean distance ratio (y-axis) of /i:/ to the centroids of /ɪ/ by CEFR level (x-axis). Midpoint on y-axis is intermediate between the two vowels.*

4. Summary and discussion

4.1. Summary

We compared the phonetic realisation of vowel characteristics in the English of native Gujarati and Thai speakers in order to determine (a) whether non-native vowel productions are correlated with proficiency level and (b) the utility of vowel characteristics as a metric in CALL-based assessment of pronunciation. While formant frequencies have been found to correlate with CEFR level, their application to automatic

assessment has been problematic, given the need to apply complicated formant tracking algorithms [3]. We therefore applied an equivalent measure that encodes the overall shape of the spectrum whilst being robust when applied to low quality data. The study revealed that:
(1) normalisation does not yield better results (only minor increase for gender)
(2) DCT-1 and DCT-2 of vowels /ae/ [æ], /eh/ [ɛ], /ih/ [ɪ], /iy/ [i:], /oh/ [ɒ], /uw/ [u:] correlate with CEFR level
(3) better rated speakers show a relatively larger vowel space
(4) high proficiency speakers show a further distance of their tense vowel /i:/ to their lax vowel /ɪ/
(5) there was no significant difference between speaker groups who have a tense-lax distinction in their native language and speakers who do not.

4.2. Discussion

The results revealed a statistically significant correlation between the Mel-scaled DCT coefficients and CEFR level for most of the target vowels. This finding is noteworthy since the Mel-scale is an auditory perceptual scale with direct correspondence with the way humans perceive speech. However, the link between perception and vowel quality articulation is a complicated one, and automated systems can never make the kinds of perceptual distinctions humans make when judging speech. One main difficulty posed for automated non-native assessment is how to normalise variation in gender. We investigated the use of an additional data reduction technique – the Lobanov normalisation. Previous research suggests that this procedure outperforms others in normalising speaker differences [16]. However, when applied to our data, the results revealed no overall significant change in correlation between Mel-scaled DCT coefficients and proficiency level, on the one hand, and Lobanov-normalised DCT coefficients and proficiency level, on the other. However, the fact that the Lobanov-normalised data were less correlated with gender than the MEL-scaled coefficients suggests that the procedure may possibly be the better option when there is need to jointly model male and female speech, as is the case with automated assessment of the phonetic realisation of vowel quality.

Overall, the findings showing that there is general correlation between DCT-based vowel features and pronunciation proficiency, as judged by human assessors, is an interesting one. This suggests that vowel characteristics should indeed be taken into account in the teaching and assessment of pronunciation. DCT coefficients are closely related to articulation and the axes of a vowel quadrilateral, given that they encode the overall shape of the spectrum which for vowels is determined by F1 (proportional to vowel height) and F2 (proportional to vowel backness). This makes it useful for CALL or other applications where clustering methods may not be ideal, such as when the teaching and assessment of pronunciation are linked to the phonetic interpretation of the speech data. The results further suggest that some characteristics for accurate discriminacy of vowels (and thus to speech intelligibility) may only emerge at the B2 level (at least for the non-native speakers in this study.) It might be argued that the assessment of non-native proficiency need not be modelled on native speech, as a lower proficiency level learner may indeed aim to reach a higher level of proficiency on the CEFR, rather than to attain native-level proficiency at the outset. However, this finding suggests that it may nonetheless be worthwhile to compare non-native speakers with native English speakers in order to adequately discriminate between the more advanced levels (e.g. C1 and C2 CEFR levels.)

Whereas native speakers of Standard Southern British English (SSBE) make a phonemic distinction between tense and lax vowel pairs, Gujarati speakers do not. We hypothesised that Gujarati speakers may therefore have difficulty in producing this distinction, as a transfer mechanism from their L1. We predicted the opposite result for the Thai native speakers, given that their native language differentiates between tense and lax vowels – albeit in a way that contrasts with the target language by giving primacy to duration cues over vowel quality cues.

The fact that learners of different L1 backgrounds were indistinguishable by the vowel quality metrics was somewhat surprising and suggests some degree of generalisability of the findings. It further suggests that learners who are acquiring a second language with a vowel system generally similar to their own may not have an advantage over learners whose L1 and L2 vowel systems are less similar. Nonetheless, it might also be argued that the strategy used by the Thai learners may actually be different from the Gujarati speakers. For instance, it is possible that the Thai speakers were indeed making a distinction, but by using their native language devices. This would lend support to the idea that it may be necessary to take the L1 vowel system of speakers into account in the development of CALL-based pronunciation teaching and assessment. Findings in second language acquisition research suggest a strong theoretical basis for this ([e.g. [1], [20], [21].

5. Conclusions

We have presented the results of a study examining the utility of spectral features of vowels in the automated assessment of non-native English speech. Previous studies ([9], [10]) have used vowel space size (or vowel dispersion) as a measure of accuracy in L2 pronunciation. However, the theoretical basis for this assumption remains unclear. There is also difficulty in normalising speaker variation, and in the use of formant analysis in automated assessment. This study examined the use of vowel spectral features as a robust alternative to formant analysis.

The results of the study suggest a link between DCT coefficients (DCT-1 and DCT2 being analogous to F1 and F2 formant measures, respectively) and the pronunciation proficiency of speakers. An approach such as this that links features to concrete linguistic phenomena would make assessment more explicit and provide a direct path to second language teaching within, for example, a CALL medium. We further show that assessment of L2 pronunciation may be enriched by an approach that takes the native language phonology of speakers into account. This would inevitably mean that future studies such as this one would need to compare speakers of various other native language backgrounds.

6. Acknowledgements

This paper reports on research supported by Cambridge Assessment, University of Cambridge. Many thanks to Jonathan Harrington and Raphael Winkelman for their support with Emu-R.

7. References

[1] Bohn, O and Flege, J. (1990). Interlingual identification and the role of foreign language experience in L2 vowel perception. *Applied Psycholinguistics*, 11, 303-328.

[2] Palethorpe, S., Wales, r., Clark, J. E., & Senserrick, t. (1996). Vowel classification in children. *Journal of the Acoustical Society of America*, 100, 3843–51.

[3] Harrington, J. (2010). Phonetic Analysis of Speech Corpora. Wiley-Blackwell, Oxford.

[4] Witt, S. (1999). Use of Speech Recognition in Computer-assisted Language Learning, Ph.D. thesis, University of Cambridge, 1999.

[5] Mostow, J., Roth. F., Hauptmann, A. & Kane, M. (1994). A prototype reading coach that listens," in Proc. AAAI, 1994, pp. 785–792.

[6] Neumeyer, L., Franco, H., Digalakis,V. & Weintraub, M. (2000). Automatic Scoring of Pronunciation Quality, Speech Communication, vol. 30, pp. 83–93, 2000.

[7] Hönig, F., Batliner, A., Weilhammer, K., Nöth, E. (2010). Automatic Assessment of non-Native Prosody for English as L2. Speech Prosody.

[8] Bohn, O and Flege, J. (1992). The production of new and similar vowels by adult German learners of English. *Studies in Second Language Acquisition*, 14, 131-158.

[9] Chen, L., Evanani, K. & Sun, X. (2010). Assessment of non-native speech using vowel space characteristics. Speech Prosody 2010.

[10] Patil, A., & Gupta, C. (2010). Evaluating vowel pronunciation quality: Formant space matching versus ASR confidence scoring.

[11] Graham, C., Nolan, F., Caines, A., Buttery, P. (2015). Using vowel formant characteristics in automated assessment. PaPE, University of Cambridge (accepted).

[12] Van Dalen, R., Knill, K., Psiakoulis, P., & Gales, M. (2015). Improving Multiple-Crowd-Sourced Transcriptions Using a Speech Recogniser. ICASSP.

[13] Watson, C. & Harrington, J. (1999) Acoustic evidence for dynamic formant trajectories in Australian English vowels. *Journal of the Acoustical Society of America*, 106, 458–68.

[14] Boersma, P. & Weenick, D. (2014). Praat: Doing phonetics by computer, version 5.4.02, http://www.praat.org.

[15] Watson, I. & Harrington, J. (1999). Acoustic evidence for dynamic formant trajectories in Australian English vowels. *Journal of the Acoustical Society of America,* 106, (1), 458-468.

[16] Adank, P., Smits, R., van Hout, R. (2004). A comparison of vowel normalization procedures for language variation research. *Journal of the Acoustical Society of America,* 116(5). 3099-3107.

[17] Lobanov, B. (1971). Classification of Russian vowels spoken by different speakers. *Journal of the Acoustical Society of America*, 49, 606-8.

[18] Buckmaeir, V., Harrington, J., Reubold, U., & Kleber, F. (2014). Synchronic variation in the articulation and the acoustics of the Polish three-way place distinction in sibilants and its implications for diachronic change. INTERSPEECH, Singapore.

[19] Glasberg, B., & Moore, B. (1990). Derivation of auditory filter shapes from notched-noise data. Hearing Research, 47, 103-38.

[20] Graham, C., & Post, B. (2012). L2 English stress production by Japanese native speakers. Proceedings of BAAP, Leeds, 26-28.

[21] Graham, C. & Post, B. (2013). Realisation of tonal alignment in the English of Japanese-English late bilinguals. INTERSPEECH, Lyon, Causal Productions.