Phonetic and prosodic features in automated spoken language assessment

♩♪ Calbert Graham, ♪ Andrew Caines, ♪ Paula Buttery

♩Phonetics Laboratory, ♪ALTA Institute: DTAL, University of Cambridge

1. INTRODUCTION

The traditional approach of using trained human assessors to evaluate spoken language proficiency has proven to be both expensive and time-consuming. This has led to renewed effort to develop tools for the automated assessment of non-native speech. Previous work on automated assessment has generally not focused on acoustic features, although it has long been established that listeners use such features in judging the naturalness of L2 pronunciation. Including such features in assessment will go a long way in making the implicit knowledge examiners use in assessing pronunciation more explicit, and can be very useful, for instance, in a computer-assisted language learning (CALL) pronunciation training programme. In this paper, we describe one strand of research in our ALTA[1] project and discuss the role of phonetic features in automated assessment of non-native speech. We will briefly discuss some of the challenges we face in automatically measuring phonetic/prosodic features, and how we go about striking a compromise between what may be a linguistically meaningful feature and what is actually measurable given the constraints of our system.

Keywords: automated assessment, CALL, prosody, non-native, speech

2. CORPUS DESCRIPTION

The dataset discussed in this paper comes 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). Currently, our pilot dataset consists of over 1000 candidates and 20,000 recordings. Candidates are speakers of various source languages: Gujarati, Hindi, Urdu, Thai, Spanish, Portuguese, and others. As preparation for our phonetic analyses, the data was orthographically transcribed using multiple crowd-sourcers and a speech recogniser according to the procedure described in [1]. The transcribed data was then automatically segmented and aligned using a Hidden Markov Model Toolkit (HTK) MLP-based algorithm to determine word and phone boundaries.

[1] A partnership between Cambridge Assessment and the University of Cambridge.

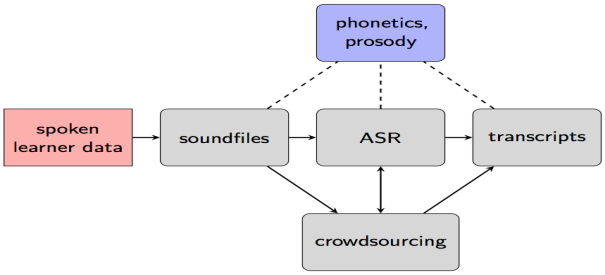
2.1. Features extracted from audio

We have extracted the following features (1) Audio-based: f0, Energy (mean, minimum, maximum, mean absolute deviation); (2) Fluency-based: long silence, silence duration, disfluencies, words, phones (mean and mean-weighted features similar to audio-based features). See [2] for a complete list. From our current feature set the system performs remarkably well with a Pearson's correlation of 0.8 with the human graders.

2.2. Phonetic features

One of the central goals of our project is to identify linguistically meaningful, criterial features in spoken language assessment, and to build a CALL system that allows candidates to improve their pronunciation based on these criteria. To that end, we seek to examine phonetic and prosodic features that can also be integrated into a pronunciation training system, as shown in Figure 1 below.

Figure 1: Phonetic and prosodic features in automated assessment systems



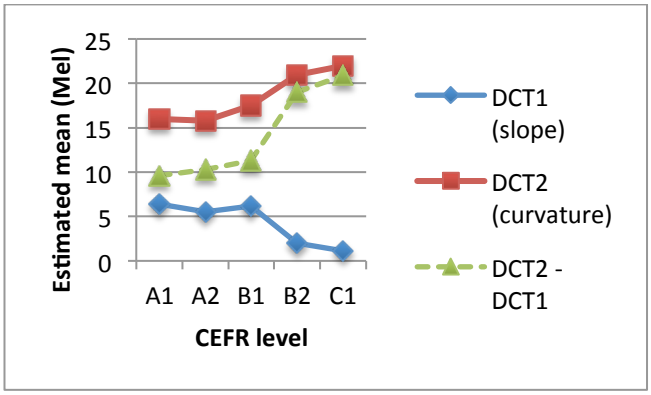
We aim to capture both the segmental and the suprasegmental (hereafter prosodic) level in speech production. Examples of features we are currently investigating include: vowel metrics, phonetic vowel reduction and sentence boundary detection. In preparation for phonetic analysis the Master Label File (MLF) output of the HTK of a pilot dataset was converted to Praat Textgrids using a script. The candidates were mostly L1 Gujarati speakers, who were multiply graded and placed into different CEFR levels. In the next section we summarise some of our on-going research.

2.2.1. Vowel quality metric

Previous research has established the relationship between vowel production accuracy and intelligibility in the L2 ([3], [4]). If confirmed as a reliable metric in

automated assessment, vowel articulation could be an excellent candidate on which to provide feedback in a CALL context. Research has shown that formant frequencies contain the primary information for the distinction of vowels [5]. Our previous pilot work [6] suggests a correlation between the vowel space of speakers (measured from formant tracking data) and their oral proficiency in the L2. However, we also observed that formant tracking may not be entirely practical with the mixed quality of the data in our corpus, which is generally unavoidable given the varying conditions in which tests are recorded in testing centres around the world. As an alternative to formant tracking we explored the Mel-scaled Discrete Cosine Transformation (DCT) method (see [5] for a detailed description). We then calculated the log Euclidean distance ratios between target vowels (e.g. the distance of [ɪ] and [i:] to a fixed anchor point (e.g. [æ]) to test the specific hypothesis of whether Gujarati speakers were making a distinction between tense and lax vowels in their L2 English, given that their L1 does not have this distinction. The log Euclidean distance ratios for the Guajarati speakers suggest a strong correlation between their vowel space and their Common European Framework of Reference (CEFR) level in English. More specifically, there is a correlation between the extent to which they realised a distinction between tense and lax vowels and their CERF levels (as shown in Figure 2).

Figure 2: Estimated DCT coefficient mean by CEFR level

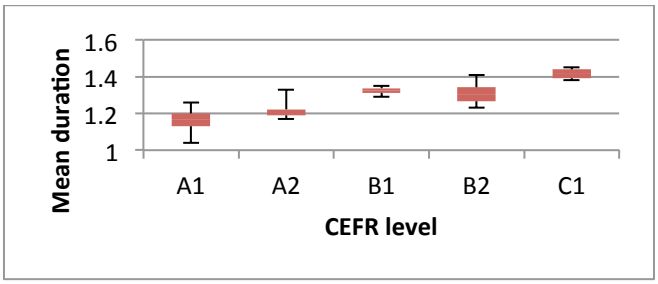


There is inherent difficulty in automatically creating accurate phone alignments. However, we find that this becomes much less problematic once we define our parameters consistently across conditions and extend our training data to include learners of different proficiencies, L1s, speaking conditions, and so on.

2.2.2. Stress and vowel reduction

In English there is a well-documented tendency for vowels in lexically unstressed position to be reduced (see e.g. [7]) when they do not carry lexical stress. This is characterised by the reducing of the acoustic correlates (f0, energy, duration) relative to what would be typical when the vowel is stressed. According to Bolinger [7], stress reduction can be seen as a possible measure of the rhythm of a language. Previous research suggests a strong link between the level of vowel reduction and the proficiency of the speakers (e.g. [8]). In our study, we calculated the ratio of all stressed to unstressed vowels produced by each speaker (the use of ratios means no further normalisation was necessary), and examined how effective this measure is if applied automatically (no manual corrections applied). The results suggest a very high correlation with proficiency.

Figure 3: Ratio of stress to unstressed vowel duration by CEFR level



2.2.3. Sentence boundary detection (SBD)

We also use prosodic information in the task of SBD in continuous speech, taking pitch and duration measurements to model spoken behaviour. This model is combined with a language model and sentence-length model in a log-linear fashion to decide on an optimal set of sentence breaks for a given recording, à la [9]. Where our work differs from [9] is that they worked with native speaker data whereas we apply the same methods to non-native learner data [10].

3. DISCUSSION AND CONCLUSION

Our starting point in this project is that automated assessment systems need to reflect the intuitions and implicit criteria used by human assessors to judge pronunciation. Given the reported link between phonetic and prosodic competence and the intelligibility of L2 speech, we explored ways to implement these features in the assessment of non-native English. Vowel quality and stress reduction are both easy to implement in automatic system as well as to teach in a CALL medium. SBD may also be useful in assessing the fluency of non-native speakers. Overall, the results that begin to emerge appear to support this view. In the future, we plan on extending our work by looking at higher-level intonation and other prosodic features, which we hope will ultimately lead to the building of a phonetically grounded CALL pronunciation training system. We also intend that the tools and methodologies we develop can be used in similar research on other source languages apart from English, and welcome the opportunity to further collaborate.

4. REFERENCES

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

[2] Van Dalen, R., Knill, K., Psiakoulis, P., & Gales, M. (2015). Automatically Grading Learners' English Using a Gaussian Process. SLaTE Workshop.

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

[4] 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.

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

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

[7] Bolinger, D. (1986). Two kinds of vowels, two kinds of rhythm. Ms., distributed by Indiana University Linguistics Club, Bloomington, IN.

[8] Graham, C. (2011). A case study of Japanese acquisition of American English stress. Journal of the Phonetic Society of Japan, 15, (3), 87.

[9] Lee, A. & Glass, J. (2012). Sentence detection using multiple annotations. Proceedings of INTERSPEECH 2012.

[10] Caines, A., Moore, R., Graham, C., Buttery P. Sentence boundary detection (in prep).