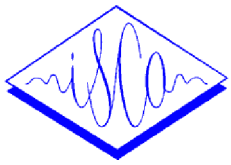

Speech Prosody 2010
Chicago, IL, USA
May 10-14, 2010

Automatic Differentiation Between Accents of Native and Non-Native English, and the Significance of Prosody

Ladan Baghai-Ravary

Phonetics Laboratory, University of Oxford
Ladan.Baghai-Ravary@phon.ox.ac.uk

Abstract

This paper analyses 25 different accents of English, determining differences using both prosodic and non-prosodic features. A Hidden Markov Model aligner phonetically labelled the data. Prosodic features of the phonemes, in the same linguistic and phonetic contexts, were calculated. For non-prosodic comparisons, Dynamic Time Warping (DTW) was used to measure segmental acoustic differences between the phonemes of a given accent with those of other speakers. The discriminative ability of the prosodic features was compared with that of the segmental acoustic score, quantifying the relative utility of prosody and segmental acoustic information in identifying accent.

Index Terms: accent differences, HMM alignment, phoneme labelling, prosodic features.

1. Introduction

To perform automatic classification of accents, it is possible to draw on a number of features of speech, but most previous work in this area has concentrated on acoustic-phonetic features, or phonemic realisation (substitution, insertion or deletion of phonemes, e.g. Schaden [1]).

Fung and Liu [2], went somewhat further than most: they used energy, pitch, formant parameters, and lexical manipulation to differentiate between native and Cantonese-accented English. Although they used two features (pitch and energy) which are affected by prosody, they did not calculate them in a way which suppressed variations due to non-prosodic factors. Even Teixeira *et al.* [3], who used a standard HMM-based recogniser to classify six different accents of speech, made no explicit use of prosodic features.

In this paper we examine the relative performance of a method based on acoustic-phonetic features, with one based solely on features reflecting the prosody of an utterance. This is based on phonemic transcriptions of citation-form pronunciations derived from a Standard Southern British English lexicon. These transcriptions were aligned with each utterance, of whatever accent, to produce labelled segments of speech, which were then analysed for segmental (acoustic) and prosodic content. The non-native English speakers often substituted, deleted or inserted different phonemes, but we have not treated these any differently from sub-phonemic differences in pronunciation.

For the 'prosodic' analysis, four features have been selected which are strongly influenced by prosody, although they do not fully characterise it. These features are also relatively immune to phonemic changes. The speech is compared in terms of the phonemes' durations, their pitches, a measure of voicing, and their intensity, all normalised relative to the neighbouring phonemes.

For the segmental acoustic analysis, dynamic time warping (DTW) and a form of Itakura-Saito distance measure [4] was used to quantify the similarity of each phoneme. The Itakura-Saito distance was made symmetrical, so that the order of accent comparison would not affect the results (described in Section 3.2.2).

2. The Data

For this study, we have used data from 'The Speech Accent Archive' of George Mason University [5]. This consists of many recordings of a single English passage read by speakers of many different languages. The passage was designed to include a wide range of phonemes and to show effects which were thought to be useful in differentiating accents:

"Please call Stella. Ask her to bring these things with her from the store: six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother, Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station."

We selected a subset of this data, choosing only those languages which had 9 or more speakers. This gave 24 different non-English L1 languages, as well as native English from various regions. The L1 language of the non-native speakers for these different accents are listed in Table 1 below.

Table 1. L1 languages and their respective abbreviations (following ISO 639.2 [6] where possible).

Language (abbreviation)	Language (abbreviation)
Arabic (ara)	Mandarin (mrn)
Bengali (ben)	Polish (pol)
Bulgarian (bul)	Portuguese (por)
Cantonese (can)	Romanian (ron)
Dutch (nld)	Russian (rus)
Farsi (fas)	Serbian (srp)
French (fra)	Spanish (spa)
German (deu)	Swahili (swa)
Hindi (hin)	Swedish (swe)
Italian (ita)	Thai (tha)
Japanese (jpn)	Turkish (tur)
Korean (kor)	Vietnamese (vie)

3. Methodology

The data was re-sampled to 16 kHz (to match our pre-existing Hidden Markov Model (HMM) phoneme alignment system [7]), and analysed to produce Mel-Frequency Cepstral Coefficients (MFCCs) [8] with a 25 ms time window for each frame, and one frame every 12.5 ms. They were then segmented using HMMs by forced alignment with HTK [9]. The pronunciations were specified with a single British English lexicon compiled from a number of sources. The same lexicon was used regardless of the speaker's accent. The phoneme inventory was a subset of the Standard Southern British English variant of SAMPA notation [10]. Optional "short-pause" models were allowed between words.

3.1. HMM topology

The data used for *training* the HMMs was an *ad hoc* corpus from the Phonetics Laboratory, University of Oxford, originally recorded for a different purpose. The speakers were all of Standard Southern British English. The utterances consist of a mixture of complete sentences, single words and phrases; all of varying lengths. The recordings were made with different equipment and at different sampling rates, digitally re-sampled to 16 kHz. The database consists of over 23,000 utterances, making a total of 48,000 spoken words taken from a vocabulary of 16,000.

The models for all phonemes had the same structure: 4 emitting states, with 1-state skips, and 4 Gaussian mixtures per state. They were trained using embedded re-estimation via 4 iterations of the Baum-Welch algorithm [11], in each of three phases:

- Training from flat-start HMMs, initialised to the global means of all the training data, to produce single-mixture phoneme, silence, and short-pause models.
- Disambiguation of alternative pronunciations (including presence or absence of inter-word pauses) followed by re-training of the models.
- Disambiguation as before, and an increase in the number of mixtures in appropriate states (using a randomised duplication of each existing mixture), followed by final re-training of the full models.

This system has been shown to work well in previous work [7], and although the data being aligned here had a much wider range of accents, informal evaluation of the alignment accuracy indicated better than expected reliability.

3.2. Data-Driven Discrimination

Four 'prosodic' features were calculated for each labelled segment. These are described in Section 3.2.1, below.

For computational simplicity other researchers have used fixed analysis lengths to capture the prosody. For example, Shriberg and Stolcke [12] extracted prosodic features using a fixed window duration of 200 ms either side of the analysis point, excluding any silent intervals.

To make our features more accurately reflect the conventional definition of a prosodic unit, yet retain the computational efficiency and simplicity of Shriberg and Stolcke's approach, we evaluated each feature over the whole of a phoneme, taking into account it's neighbours (4 phonemes either side) rather than using a fixed time-window. This suppresses non-prosodic and speaker-dependent variations. Like them, we also excluded any silent intervals.

3.2.1. Prosodic features:

- *Phoneme duration* – the duration of each labelled phoneme in seconds, divided by the average of the 9 labelled phonemes centred on the one in question.
- *Intensity* – the power of each labelled phoneme in arbitrary units, divided by the average of the 9 labelled phonemes centred on the one in question.
- *Pitch* – the pitch of each labelled phoneme in Hertz, divided by the average of 9 voiced labelled phonemes centred on the one in question. The pitch was estimated using the Praat 'pda' command [13]. This is one of the best publicly available pitch estimation programs, at least for 'clean' speech, such as that used here [14].
- *Voicing* – the proportion of frames within each segment which were deemed voiced by the 'pda' program. This was not normalised in any way, since the values it produces can be very approximate, especially for short segments.

3.2.2. Acoustic segmental difference:

A single value measuring non-prosodic acoustic dissimilarity, was calculated as follows: divide the signals, S_1 and S_2, into overlapping frames, and calculate a modified version of the Itakura-Saito distance [4] between all the frames, $f_1(n)$, of S_1 and all those, $f_2(m)$ of S_2. This modified function, $d(f_1(n), f_2(m))$, is symmetrical, and defined as:

$$d(f_1(n), f_2(m)) = \sqrt{\frac{E_{12} \cdot E_{21}}{E_{11} \cdot E_{22}} - 1}$$

where E_{xy} is the residual power after filtering f_x with an optimal linear prediction filter calculated for f_y, and x and y are 1 or 2, as appropriate. These values are then squared and used as 'local distances' in dynamic time warping (DTW) between the signals. The resultant cumulative distance is then normalised to account for the signals' durations and is used as the measure of acoustic similarity:

$$D(S_1, S_2) = \overline{d(f_1(n), f_2(m))^2}$$

This measure is similar to the Itakura-Saito distance, but applies to a whole phoneme-level unit of speech (even if it is non-stationary) and is independent of its duration.

4. Results

For each 'prosodic' feature (power, pitch, duration and voicing) we define a distance as the square of the difference between respective values. The non-prosodic (acoustic) analysis yields a distance directly. For each feature, we calculate the ratio of distances between every labelled segment from each of the different accents, compared to those from the same accent. The discrimination score is defined as the ratio of mean inter-accent distances to mean intra-accent distances:

$$Disc.\ Score = \frac{Mean(inter\ accent\ distance)}{Mean(intra\ accent\ distance)}$$

The results for the most discriminating phoneme of each feature, averaging distances over every accent pair, are shown in Figure 1.

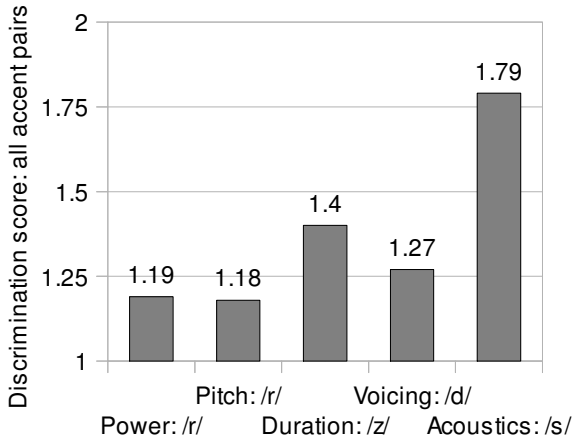

Figure 1: Score for the most discriminating phoneme for each feature.

It seems that the best overall discrimination is provided by the acoustic differences between /s/ phonemes, with the durations of /z/ phonemes in 'second place'. Although these results are informative, they do not give a complete picture of the importance of the various features for each accent; they

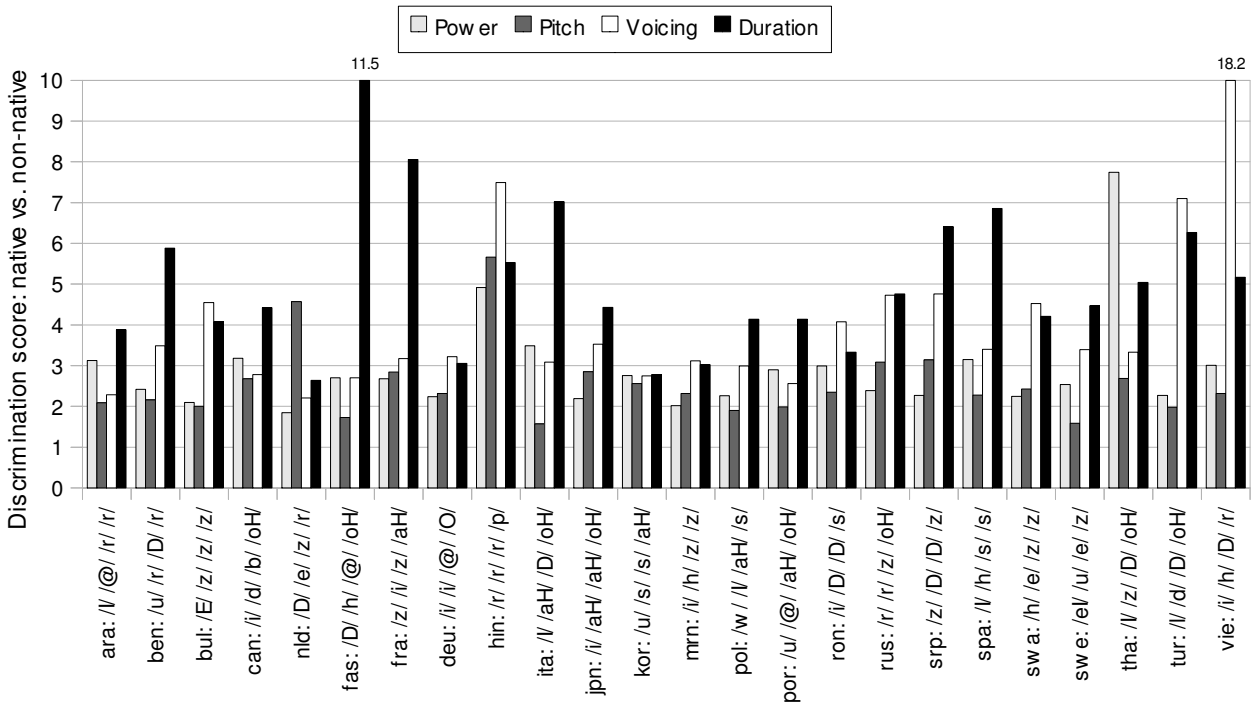

Figure 2: Discrimination score between native and non-native accents for the most discriminative phonemes, for each prosodic feature (Section 3.2.1), vs. language abbreviation (Table 1). For each L1, 4 bars are plotted. The first (light grey) bar represents the discriminative ability of the most discriminative phoneme based on power; the second (dark grey) bar represents that of the most discriminative phoneme based on pitch; the third, based on voicing, and the last based on duration. Identities of the 'most discriminative' phonemes are listed in the same order below the x-axis.

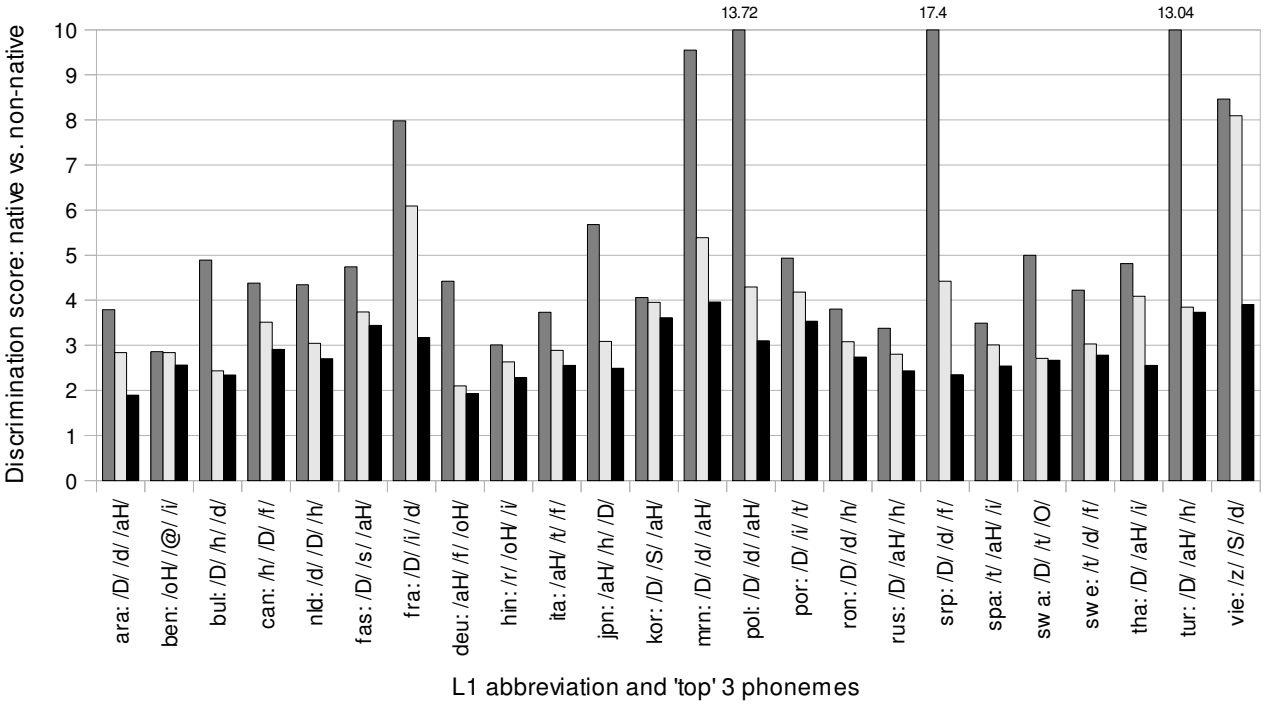

Figure 3: Discrimination score between native and non-native accents for the three most discriminative phonemes based on non-prosodic acoustic difference (Section 3.2.2), vs. language abbreviation (Table 1). For each L1, 3 bars are plotted. The first (dark grey) bar represents the discriminative ability of the most discriminative phoneme based on non-prosodic acoustic difference; the second (light grey) bar shows that of the second most discriminative phoneme; the third, the third most discriminative. Identities of the 'most discriminative' phonemes are listed below the x-axis.

simply indicate which phonemes and which features are most generally useful in discriminating between arbitrary pairs of accents.

There is not enough space available here to present all our results for all accent pairs, so we will now focus on comparing the native English accent with non-native ones. Thus in Figures 2 and 3, the discrimination score is calculated with intra-accent distances based only on native English utterances, and the inter-accent distances are calculated between native English and each of the other accents individually.

The results for 'prosodic' features are shown in Figure 2. The x-axis labels show the abbreviations for the speakers' L1 languages followed by the phonemes which gave the highest score for power, pitch, duration, and voicing features respectively.

The non-prosodic acoustic feature, as described in section 3.2.2, is presented in Figure 3 using the same ratio. Here, the x-axis labels show the speakers' L1 language followed by the 3 most discriminative phonemes for each accent.

5. Discussion

Comparing every English accent pair, it can be seen from Figure 1 that in general, the most distinctive 'prosodic' features are the duration of /z/, the voicing of /D/, and the power and pitch of /r/. Overall, the most acoustically distinctive phoneme is /s/.

The 'prosodic' features for each non-native English accent, in comparison to the native English accent, are illustrated in Figure 2. This shows that the duration and voicing of the segments play the major part in differentiating between the native and non-native accents. The phonemes /D/, /z/, /s/, /r/, /aH/ and /oH/ seem to hold some of the most important prosodic information. In particular, /D/ and /z/ are the most often distinctive feature with respect to voicing, and /oH/ with respect to duration, but there is no clear 'winner' for power or pitch.

Looking at the ratio of the non-prosodic differentiation of the native and non-native accents in Figure 3, in general, the phonemes /D/, /aH/ and /i/ seems to hold the most significant information. At least one of these is in the top 3 most discriminating phonemes in 22 out of 24 non-native accents, with /D/ is present in 17, /aH/ in 12, out of 24.

6. Conclusion

Although each accent pair has its own characteristics and should be examined individually for more detailed analysis, this study has shown some more general and informative features for differentiating between accents of English.

The results in Section 4 show that both 'prosodic' and non-prosodic features can be strong indicators of differences between native English and non-native accents. Both prosodic and non-prosodic features of /D/ and /aH/ appear to be particularly good indicators for differentiation between native and non-native English. However, for specific accent pairs there can be other phonemes and features which provide a much stronger indication. For example, when comparing Farsi to English L1, the duration of the /oH/ phoneme is by far the strongest cue.

The statistics presented have all been derived automatically and without reference to any human labels or interpretation of phonetic features. Consequently the differences between accents, as identified, are objective.

This study has already produced an extremely large amount of data describing the differences between accents. Further studies using non-native accents as the baseline, and looking at more than just the 'top' results, should prove extremely valuable.

Acknowledgements

The author would like to thank Prof. John Coleman and Dr. Greg Kochanski (both of the Phonetics Laboratory, University of Oxford) for their helpful advice and comments.

References

[1] Schaden, S., “Generating Non-Native Pronunciation Lexicons by Phonological Rules”, in Proc. ICSLP, 2004.

[2] Fung, P., and Liu, W. K., “Fast Accent Identification and Accented Speech Recognition”, in Proc. ICASSP, 1999.

[3] Teixeira, C., Trancoso, I., and Serralheiro, A., “Accent Identification”, in Proc. ICSLP, Philadelphia, PA, 1996.

[4] Itakura, F., and Saito, S., “A statistical method for estimation of speech spectral density and formant frequencies”, Electronics & Communications in Japan, 53A: 36-43, 1970.

[5] “The Speech Accent Archive”, George Mason University, http://accent.gmu.edu/

[6] “ISO 639-2 Registration Authority”, available online at http://www.loc.gov/standards/iso639-2/

[7] Baghai-Ravary, L., Kochanski, G., and Coleman, J., “Objective Optimisation of Automatic Speech-to-Phoneme Alignment Systems”, in “Human Language Technologies as a Challenge for Computer Science and Linguistics”, Vetulani, Z., (ed.), Poznan, 2009.

[8] Bridle, J. S., and Brown, M. D., “An experimental automatic word recognition system”, JSRU Report No. 1003, Joint Speech Research Unit, Ruislip, England. 1974.

[9] Young, S. J., et al., “The HTK Book (for HTK Version 3.4)”. Cambridge University Engineering Department, 2006.

[10] Wells, J. C., “SAMPA computer readable phonetic alphabet”. In Gibbon, D., Moore, R. and Winski, R. (eds.), 1997. Handbook of Standards and Resources for Spoken Language Systems. Berlin and New York: Mouton de Gruyter. Part IV, section B.

[11] Baum, L. E., Petrie, T., Soules, G., and Weiss, N., “A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains”, Ann. Math. Statist., vol. 41, no. 1, pp. 164-171, 1970.

[12] Shriberg, E., and Stolcke, A., “Prosody-Based Automatic Segmentation of Speech into Sentences and Topics”, Speech Communication 32(12), Special Issue on Accessing Information in Spoken Audio, September 2000.

[13] Boersma, P., and Weenink, D., “Praat: doing phonetics by computer”, http://www.praat.org/

[14] Kotnika, B., Högeb, H., and Kačič, Z., “Noise robust F0 determination and epoch-marking algorithms”, Signal Processing vol 89, issue 12, pp 2555-2569, December 2009.