Diphthongization as a cue for the automatic identification of British English dialects

Emmanuel Ferragne, François Pellegrino

Laboratoire Dynamique Du Langage, UMR CNRS 5596

Emmanuel.Ferragne@univ-lyon2.fr

François.Pellegrino@univ-lyon2.fr

Introduction

☞The vowels in the words ***deeds*** and ***food*** have been phonologically analyzed as monophthongs. However, formant stability in these vowels varies across dialects in the British Isles.

☞Standard Southern British English **(sse)** is known to exhibit rather diphthongized realizations of *deeds* and *food* whereas Scottish Highlands English **(shl)** has true monophthongs (see Fig. 1 and Fig. 2)

✓**Goal:**

☞Can the degree of monophthong diphthongization alone constitue a reliable cue to dialect classification?

☞Comparison of classification based on formant trajectories alone with classification performed by a trained phonetician

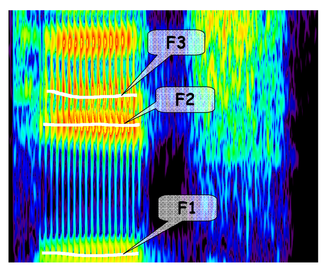


*Figure 1: Broad-band spectrogram (300 Hz filters; 4000 Hz diplayed) of the vowel in **deeds** spoken by a male speaker from the **Scottish Highlands***

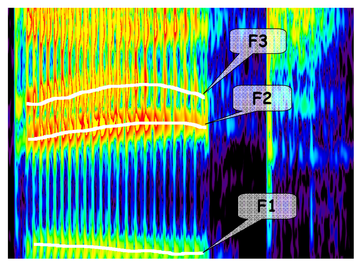
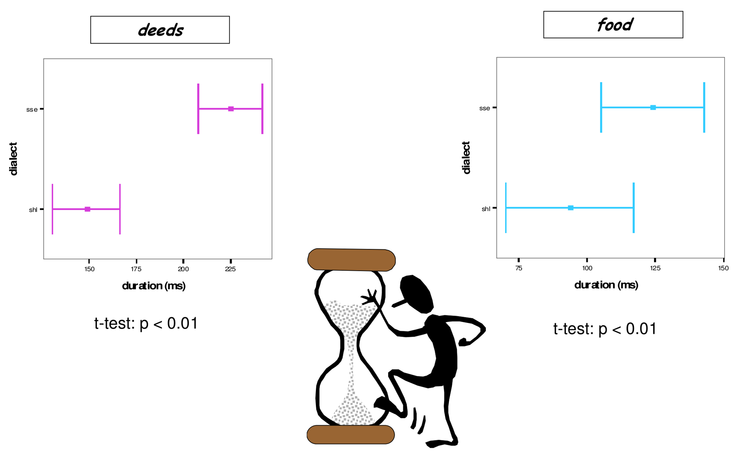


*Figure 2: Broad-band spectrogram (300 Hz filters; 4000 Hz diplayed) of the vowel in **deeds** spoken by a male speaker of **Standard Southern British English***

Caution:

Duration is the most obvious spurious factor here:



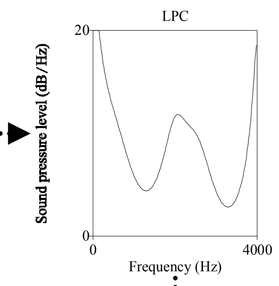
Perceptual experiment

Identification task: an expert native English phonetician was asked to listen to the vocalic portion of *deeds* and *food* and tell whether the stimulus was uttered by a speaker of the **Scottish Highlands** or a speaker of some **other** British English dialect (sse) known to diphthongize these vowels to a certain extent.

Three types of stimuli:

natural vowels (nat): the vowels of *deeds* and *food* were segmented (boundary placed at the inception of the formant structure and at the end of it), extracted and normalized for amplitude → 15 tokens for each vowel

FLATTENED F0 :
men: 120 Hz
women: 200 Hz



neutralized duration (nd): mean duration for the vowel in *deeds* and for the vowel in *food* was computed over the vowel tokens from the two dialects.

☞mean deeds: 180 ms

☞mean food: 106 ms

☞the duration of each *deeds* vowel was modified (using PSOLA) to equal 180 ms, and the duration of each *food* vowel: 106 ms.amplitude normalized →15 tokens for each vowel

resynthesized vowels (syn):

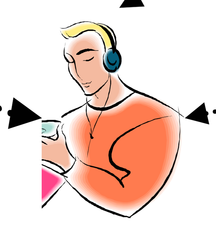
☞vowel resampled: 8000 Hz

☞lengthened (PSOLA): deeds: 275 ms; food: 175 ms.

☞LPC: 25 ms window, 5 ms steps, 8 coefficients, pre-emphasis above 50 Hz

☞flat F0: men: 120 Hz; women: 200 Hz

☞amplitude normalized →15 tokens for each vowel

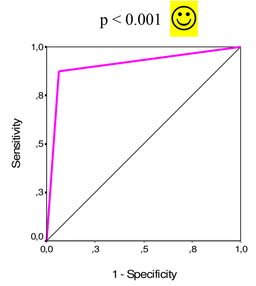


Trained phonetician

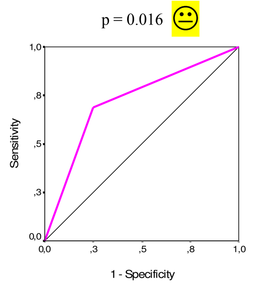
Results

word	condition	hit (%)	false alarm (%)	identification (%)
deed	nat	43.75	3.125	90.625
deed	nd	34.375	12.5	71.875
deed	syn	18.75	9.375	59.375
food	nat	31.25	18.75	62.5
food	nd	28.125	37.5	40.625
food	syn	9.375	3.125	56.25

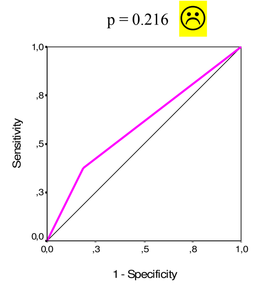
deeds nat — p < 0.001



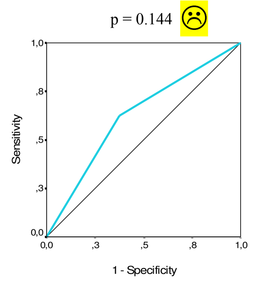
deeds nd — p = 0.016



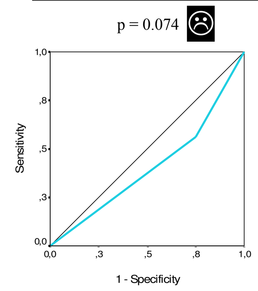
deeds syn — p = 0.216



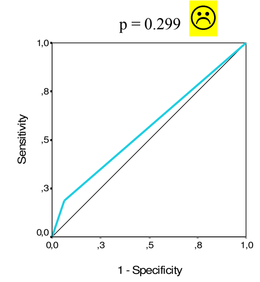
food nat — p = 0.144



food nd — p = 0.074



food syn — p = 0.299



ROC curves caption: Fisher's exact test: H_0: no association between phonetician's response and stimulus

Acoustic measurements

Three measures of diphthongization

(Central frequency for the first three formants in Bark measured with the Praat program)

1. **dif**: difference between value at 80% of vowel duration and value at 20% of duration
2. **sd**: standard deviation of 9 values extracted from each formant: every 10% of the duration
3. Δ: 9 values extracted from each formant (same as above)

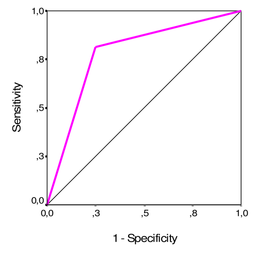
$$\Delta = \frac{1}{n-1}.\frac{1}{d}\sum_{k=1}^{n}\left|F_{k+1}-F_k\right|$$

F_k: frequency of a given formant (F1, F2 or F3) at point *k* (*n*=9); *d*: duration between measurements: 1/10 of total vowel duration

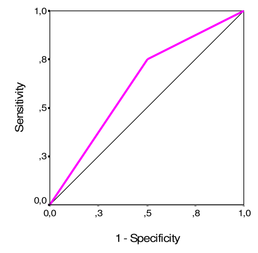
Classification: linear discriminant analysis

ROC curves caption: % correct identification and Fisher's exact test: H_0: no association between actual dialect and dialect-membership predicted by linear discriminant analysis; **plus formant with highest correlation with discriminant function**

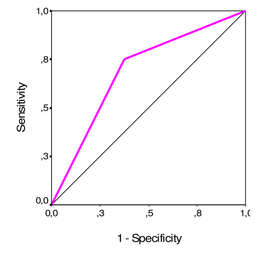
dif *deeds* (F1,F2,F3) — 78.1% - p < 0.001 - F1



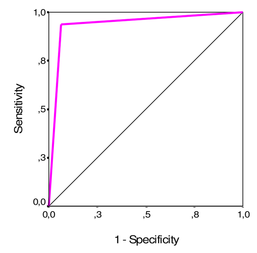
sd *deeds* (F1,F2,F3) — 62.5% - p = 0.137 - F2



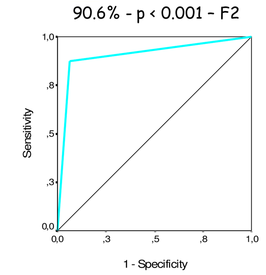
Δ *deeds* (F1,F2,F3) — 71.9% - p = 0.037 - F1



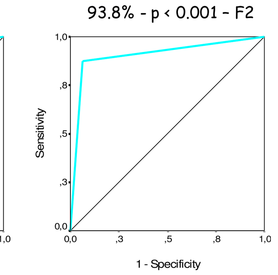
duration *deeds* — 93.8% - p < 0.001



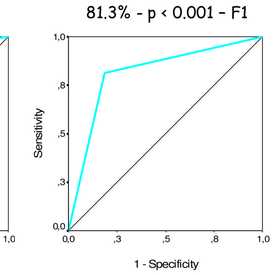
dif *food* (F1,F2,F3) — 90.6% - p < 0.001 - F2



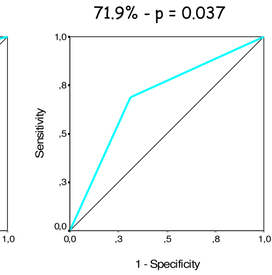
sd *food* (F1,F2,F3) — 93.8% - p < 0.001 - F2



Δ *food* (F1,F2,F3) — 81.3% - p < 0.001 - F1



duration *food* — 71.9% - p = 0.037



Conclusion

☞In the perceptual experiment, the phonetician managed to correctly classify above chance level only the *deeds* vowel, specifically in the **nat** condition. Although diphthongization must have helped, it seems that duration was the main cue used in this task. The poor results for *food* suggest a ceiling effect perhaps due to lack of within-dialect homogeneity.

☞The best classification score with linear discriminant analysis for *deeds* using formants is achieved with **dif**; and F1 trajectory is the most relevant dimension. However, duration constitutes the most reliable cue. As for *food*, **sd** is the best metric, with F2 showing the highest correlation with the discriminant function. Duration here only plays a marginal role.