Odyssey 2010
The Speaker and Language Recognition Workshop
28 June – 1 July 2010, Brno, Czech Republic

Suprasegmental Acoustic Cues of Foreignness in Czech English

Jan Volín [*] *and Radek Skarnitzl* [**]

[*] Department of Anglophone Studies, Metropolitan University Prague
[**] Institute of Phonetics, Faculty of Arts, Charles University in Prague
volin@mup.cz; radek.skarnitzl@ff.cuni.cz

Abstract

English as the lingua franca of the modern world is spoken by increasing numbers of individuals from various nations and of different linguistic backgrounds. Detection of foreign accents can potentially lead to improvement in the development of ASR systems which have to cope with vast, but, at a certain level of generalization, finite variation of English accents. Samples of Czech English have been parameterized in terms of linguistically explicable suprasegmental variables and subjected to multiple regression analyses with foreignness scores as the dependent variable. The results are remarkably consistent and confirm that the chosen parameters contain cues of accentedness strength and might be used in detection and possibly explanation of the Czech accent in English.

1. Introduction

Foreign accents in speech have been studied for centuries from various perspectives. Quite probably, there has always been some aspect of applicability or even profit present in the accent considerations. People often seem to care about the impression their accents make and are willing to exert considerable efforts and means to improve this impression. At times, certain individuals offer miraculous methods of accent improvement for cash while carefully hiding the fact that the methodology of pronunciation training leaves much to be desired: our current knowledge is too imperfect to design exercises which guarantee fast progress to everyone. Several studies are listed in [1], showing that formal instruction in general does not particularly improve pronunciation [1: 200]. Many of the remarks on foreign accent understanding made more than half a century ago by Abercrombie [2] are certainly true today: despite the advances in foreign accent research, it is not even known to what extent foreign accentedness really matters in everyday communication. It is known, however, that it *does* matter.

The applicational demands are also definitely not diminishing at present. Eskenazi, in her overview of the field of computer assisted language learning [3], provides numbers of examples of goal-directed effort to improve technology which is clearly needed. And just to mention a few other fields of application, military interests include foreign accentedness since intelligence collection may be crucially dependent on linguistic disguise. Similarly, the areas of banking, security and forensics require systems that identify individual speakers and possibly provide some information about them.

There is also a serious psychological concern connected with foreign accents. Some studies have demonstrated that accented speech can bias the perception of the speaker's personality. For instance, social reactions were elicited from respondents who heard foreign-accented English in [4]. Various measures of foreignness were correlated with the impressions of the personality of an unknown speaker. It was found that eight features of foreignness induced the perception of lower social status of the speaker with correlation coefficients ranging from $r = -0.67$ to $r = -0.78$ [4: 498].

Still a grosser example of social prejudice is provided by [5] who found that the processing of the speech signal in the listener's brain is affected even by the visual attributes of the speaker, namely by his or her overt ethnicity. The experimenter instructed over 60 white American students to follow a 500- to 600-word-long text of a lecture type. Half of the students heard the speech while looking at a photograph of a lecturer of their own ethnicity, while the other half watched a photograph of a lecturer of Asian origin. Although they were listening to the same material, the two groups of respondents differed significantly in comprehension. Those who believed they were listening to an Asian lecturer achieved worse comprehension scores after the presentation [5: 516].

Some more examples of circumstantial changes in accented speech perception are provided by [6]. It follows that if the perception of speech is affected so deeply even on subconscious levels, we should take into account the possibility that foreign accentedness may have a bearing on the quality of people's lives and it should be studied seriously and thoroughly.

On the technological side of the problem, there is an idea of ASR systems switching between models according to the speech accent they have to recognize. Naturally, the precondition for this improvement is the correct recognition of the speech accent in a fast and effective pre-processing stage [7]. The probabilities of the target units could be increased by removing the differences between the "much too general" standard of most of the current ASR applications. The HMMs could move closer to a system in which an actual speaker is structuring his or her linguistic messages.

The question of what codes the foreign accent in the speech signal has been answered differently by different authors. Interestingly, seven of the eight features mentioned above in connection with [4] were defined segmentally, i.e., based on phone alterations, while only one was characterized suprasegmentally. Similarly, [8] gives an overview of mutual influences between the mother tongue and the target language of immigrants, and mentions exclusively vowels and consonants, while no explicit statement concerning prosodic dimension of speech is made. In an earlier study by the same author, foreign accent seems to be equaled with incorrect pronunciation of vowels and consonants [9: 171].

Etymologically, the word "accent" itself (in many languages other than English as well) acknowledges that foreignness is somehow coded in suprasegmentals. One of the reasons why our present study focuses on prosodic correlates in the speech signal is the apparent bias in research towards the segmental dimension of the foreign accent (mentioned also in [10]). Another and from the practical point of view more

important motivation of our focus is the fact that while certain vowel or consonant cues of foreign accent may be absent in an individual sentence whose accentedness is being assessed, there is always some rhythm, intonation and some tempo present in every utterance. For instance, the velar nasal [ŋ] is often pronounced for 'ng' spelling in standard English. Czech speakers of English often pronounce it as [ŋk] which is felt as a marker of foreignness. However, the frequency of occurrence in texts is about 1 % in all contexts, including those where it actually should be pronounced [ŋk] [11 in 12: 232]. This means that many sentences may not contain this marker at all. There is no suitable item to observe it in within the first 44 words of our introduction in this paper. Some prosodic markers, on the other hand, should always be present in a regular utterance.

Of the three functional variables defined in [13], i.e., comprehensibility, intelligibility, accentedness, we have focused our attention on the last one. We have previously established that listeners can detect Czech accented English quite reliably and that both listeners whose native language is English and listeners whose mother tongue is Czech achieve considerable agreement in the evaluation of the strength of Czech accent in English utterances [14]. What is still unanswered is the question of typical features of Czech English. Although we have informally observed many of the characteristics of the Czech accent, we have only recently found a few directly measurable candidates of accentedness [15].

Apart from the technological applicability we are predominantly interested in the conceptualization of foreign accent. Therefore, we will not use some of the otherwise effective methods like MFCC parameterization along with HMM-based search since these do not provide linguistically interpretable features. The parameters used in our study have more straightforward correlates in the phonetic structure of speech (see below, section 2).

Most of the real life speech communication acts take place under the so-called adverse speech conditions. The acoustic environment of our speech interactions is often full of other noises (maskers), or the source of the speech signal is in such a position relative to the percipient that the signal arrives incomplete. Therefore, in addition to clean speech signal we will also work with the signal that has been degraded in two ways: masked with the so-called coffee shop noise and bandpass-filtered so as to neutralize a substantial amount of segmental information. While intelligibility of such a signal understandably decreases, the acoustic cues that indicate foreignness might remain intact. Such robust markers would be useful to identify.

The questions asked in our study can be summarized as follows. Do any combinations of the candidate acoustic measures explain the variation in foreign accent scores of Czech English utterances? Which of the successful markers of foreignness (if any) are still useful in a degraded speech signal? Is there any substantial difference between the two forms of speech signal degradation used in our study?

2. Method

2.1. Material

Czech accented speech was produced by 11 female speakers. They were university students selected from the pool of speakers used in [14]. The women were reading texts of BBC news-bulletins. Four of the speakers had been assessed as having very little Czech Accent (henceforth Accent Group A), another four as having a strong Czech Accent (henceforth Accent Group C) and three had received evaluation corresponding roughly to the mid-point between the two extreme groups (henceforth Accent Group B).

The original evaluation of eight of the speakers was carried out by 10 listeners whose mother tongue was English and 13 native speakers of Czech. Both groups of listeners were linguistically trained and their scores displayed remarkable agreement. The scores of foreignness were later confirmed again by another group of listeners: 26 undergraduate students of English with lesser linguistic training than the previous group and of Czech origin. The new ratings delivered by less linguistically sophisticated listeners correlated again with the previous ones very highly ($r = 0.96$).

Assessments were elicited on a 5-point scale, which is the one that has been used most commonly in the foreign accent research [1: 194]. It is also our impression that this scale is the most convenient for listeners to mentally embrace. A score of an individual utterance is an arithmetic mean of scores awarded by all listeners.

Three more speakers of the same background were added for preliminary testing of the results. These speakers were not evaluated by groups of listeners and were only impressionistically assessed by the experimenters. A native English BBC newsreader was also added to this set.

For the current experiment, three utterances were chosen per speaker, each 12 to 14 words long. The utterance had to be free of dysfluences. Some of the utterances were also manipulated by PSOLA to change the sex of the speaker artificially since we wanted to avoid over-learning during the perception tasks and we also wanted to find out whether the artificial change of the sex of the speaker bears on the foreignness scores. Since there were eight speakers used in the training set (and three plus one in the testing set) and 14 items were also assessed in the 'male voice' form, the whole training set consisted of 38 items. The testing set comprised 12 items: 9 from Czech speakers of English and 3 from a native BBC news reader who was assumed to have a low foreignness score (It has been shown in the past that even native speakers can be judged to have some foreign accent when listeners only hear short speech samples without a context [e.g., 16, 17]).

The coffee shop noise was recorded in a regular cafeteria dining room with five to ten competing talkers distant from the microphone and noises caused by movements of chairs, cutlery, plates and other regular appliances. The noise was relatively steady in its amplitude and was mixed to the speech signal at SNR = – 6dB.

The filtered speech samples had the band between 400 and 1400 Hz preserved, and everything below and above this band removed. This band was chosen based on trial-and-error experiments in [18] where this band was found to be most suitable for automatic detection of energy distribution differences between regular Czech and English news reading.

2.2. Variables

The dependent variable in all the current analyses is always the foreignness score (*FS*), although we differentiate among the scores produced by listeners when hearing the clean speech signal and scores resulting from the assessment of masked or

filtered speech (see above). These scores are kept separately under all conditions.

The independent variables are listed in Table 1. We measured seven parameters in the time domain, three parameters in the energy domain and four features of the frequency domain (fundamental frequency – F0 only). In the time domain, two of the measures concerned articulation rate, while the remaining five were speech rhythm correlates. Articulation rate was measured in syllables per second (AR-syll) and in phones per second (AR-pho). As explained, e.g., in [19], these two measures are not identical and human perception should be apparently somehow associated with both of them.

Table 1: Independent variables in the original experimental design of the study

Variable	*Domain*	*Unit*
Articulation Rate - syll	time	syllable·sec^{-1}
Articulation Rate II - pho	time	phone·sec^{-1}
PVI-Consonant	time	ratio
PVI-Vowel	time	ratio
Vowel proportion	time	ratio
Cons-dur variation I	time	millisecond
Cons-dur variation II	time	ratio
Stress-unstress SPL	energy	decibel
Stress-mean SPL	energy	decibel
Unstress-mean SPL	energy	decibel
F0 standard deviation	frequency	semitone
F0 range	frequency	semitone
F0 percentile range	frequency	semitone
F0 decl. gradient	frequency	ST·sec^{-1}

The pairwise variability index (PVI) was measured in a labelled chain of phones in which all the immediately neighbouring consonants were merged into a consonantal interval and all the neigbouring vowels were merged into a vocalic interval (e.g., [20]). Durations of the successive consonantal and vocalic intervals were measured and durational variation for each of them was calculated from:

$$PVI = 100 \times \sum_{j=1}^{n=1} \left| \frac{d_j - d_{j+1}}{d_j + d_{j+1}} \right| /(n-1), \quad (1)$$

where n is the number of consonantal or vocalic intervals (depending on which PVI is measured) in the investigated stretch of speech, and d is the duration of a consonantal or vocalic interval. This index can hypothetically vary between 0 and 100 but not including these limits. In realistic speech tasks, the values close to these limits also do not occur.

Vowel proportion (V%) as a correlate of a rhythmic type was suggested by [21] and since then has been used with greater or smaller success in a number of studies. It is simply

$$V\% = 100 \times \frac{\sum_{k=1}^{m=1} d_{Vk}}{d_{Utt}}, \quad (2)$$

where m is the number of vocalic intervals in the given stretch of speech, and d_V is the duration of a vocalic interval while d_{Utt} is the duration of the investigated stretch of speech (usually an utterance).

Variation in consonantal interval durations was also shown by [21] as relevant in speech rhythm consideration. It is expressed as the standard deviation from the mean of consonantal interval durations:

$$s_{C-dur} = \sqrt{\frac{\sum_{i=1}^{n=1} \left| d_i - \bar{d} \right|}{n-1}} \quad (3)$$

It is obvious that this manner of measuring variation is vulnerable to overall speech tempo – slower speakers will produce a higher standard deviation. Dellwo [22] therefore recommends normalizing the measure relative to the mean, which is a common statistic procedure leading to the coefficient of variation:

$$C_{Var} = 100 \times \frac{s_{C-dur}}{\bar{d}}, \quad (4)$$

where $s_{C\text{-}dur}$ is the standard deviation from the mean of consonantal interval durations from equation (3) and it is divided by the mean duration of consonantal intervals in the stretch of speech that is being investigated.

Sound pressure levels were measured in the Praat software [23]. The effective window length was set to about 20 milliseconds. The intensity measures were acquired in the mid 30 milliseconds of all stressed and unstressed syllabic nuclei in the investigated stretch of speech: we calculated the arithmetic mean of five measurements 3 ms apart, centred around the mid of the syllabic nucleus. The SPL values were used in three descriptors: (a) in the mean difference between stressed and unstressed syllables, (b) the difference between mean SPL of all the stressed syllables and the overall SPL mean of the utterance, and (c) the difference between mean SPL of all the unstressed syllables and the overall SPL mean of the utterance.

Fundamental frequency characteristics were also collected with the help of [23]. Autocorrelation method was used taking F0 values every 10 milliseconds. The resulting F0 tracks were manually corrected against errors like octave jumps or accidental voicing in voiceless obstruents. F0 variation was represented by standard deviation of all voiced points from the mean (henceforth F0-s). Another variation descriptor was F0 range calculated as the difference between the maximum and minimum F0 value in the stretch of speech under scrutiny (F0-rg). Similarly, the percentile range was taken from the 10th percentile to the 90th percentile: it is sometimes suggested that this measure better approximates human perception (F0-percrg)). The last F0 descriptor was the gradient of the regression line through all the measured voiced points (F0-grad). This descriptor suggested in [24] has been successfully used to approximate the general intonation downtrend in speech (e.g., [25], [26], [27]).

2.3. Analyses

Multiple regression analyses were performed with the software STATISTICA 7 and the following parameters were taken into account in the progress. In all combinations of independent variables, it was the adjusted coefficient of determination ($Adj.R^2$) which was taken most seriously since it reflects the proportion of variance explained by the given model with correction for the number of cases per number of explanatory

variables. Unlike the ordinary coefficient of determination (R^2) it does not overestimate the power of the model. In all cases it was also observed whether the growth of the $Adj.R^2$ is not at the expense of the test criterion F, which would indicate that the model could become less effective for further generalization and prediction and also less economical.

Naturally, both b- and β- coefficients were of interest: the former for the prediction of *FS* values of unknown cases, the latter for the comparison of the importance of individual variables in the generated model.

3. Results

3.1. Clean speech signal

The first series of multiple regression analyses (MRA) concerned search for an optimal set of independent variables (i.e., explanatory factors) with regard to the foreignness score based on clean speech signal. With 38 cases and the general requirement to have about five to fifteen cases per factor, it was obvious that from the original number of 14 independent variables only about three or four can be ultimately retained in a realistic model.

A series of plain correlation analyses and an inspection of 2-D scatterplots showed that in the temporal domain, the measures of consonantal variation had no explanatory power whatsoever. Similarly, the difference between the SPL of either stressed or unstressed syllables from the mean in the energy domain did not exhibit any meaningful link with the accentedness scores. All four fundamental frequency domain measures proved quite promising, although three of them were conveying the same feature: variation of the F0 values in the F0 contours.

For the sake of completeness we first provide the results of multiple regression analyses with only two independent variables in the model. Table 2 presents those in which both variables proved significant at least at the level of $\alpha = 0.05$. Although these simple models do not represent our main concern, they are indicative of the role of individual factors in the general framework and will also be useful later on to elucidate certain speech masking effects in adverse conditions. The stronger of the two independent variables (i.e., more significant in terms of the β-coefficient) is always mentioned first in each line of Table 2.

It is interesting to notice that if an F0 measure is introduced into the equation, it is always stronger than the other variable. Within the F0 domain, the variability indicators are more important than the declination gradient. The SPL difference between stressed and unstressed syllables makes a significant contribution with both temporal and frequency domain factors and varies in its importance quite substantially. Articulation rate measures, on the other hand, occur in the lower part of Table 2 where the more significant models are located.

At this stage, it is difficult to decide whether the articulation rate in syllables per second (AR-syll) is a more effective measure than the articulation rate in phones per second (AR-pho). They correlate with each other highly (r = 0.96) and in each case in Table 2 they can be replaced with one another without changing the situation substantially.

Table 2 always reports only the instance with higher explanatory power (i.e., higher coefficient of determination - $Adj.R^2$). Naturally, AR-syll and AR-pho do not occur in Table 2 together in one line because when put into an analysis concurrently, one of them becomes redundant.

Table 2: Results of multiple regression analyses with two independent variables ($p < 0.0001$).

Variables	*R*	*Adj.R²*	*F (2, 35)*
SPL; PVI	0.64	0.387	12.70
F0-grad; SPL	0.68	0.435	15.26
F0-grad; PVI	0.71	0.469	17.33
F0-rg; PVI	0.71	0.483	18.31
F0-s; V%	0.72	0.484	18.38
F0-percrg; V%	0.72	0.484	18.38
PVI; AR-syll	0.72	0.486	18.47
F0-rg; SPL	0.73	0.500	19.49
F0-rg; AR-pho	0.74	0.527	21.57
F0-s; AR-pho	0.74	0.527	21.63
SPL; AR-syll	0.74	0.528	21.70

Interestingly, this is not the case when any of them is entered together with another temporal domain measure: the pairwise variability index of vocalic intervals (traditionally PVI-V, henceforth only PVI). Both overall tempo as such and rhythmicity of the speech can apparently jointly contribute to the detection of Czech accent in English.

It is clear from Table 2 that the highest explained variance in the data was slightly over 50 percent. The next step, then, was to find out whether addition of another variable into the model can improve the performance. Table 3 presents those cases where only one candidate measure from each domain (frequency, time, energy) was entered and where the combination yielded significance for all of the three independent variables used. There were only four such combinations (Table 3): in other combinations of three variables, there was always at least one with a standard error of estimate which prevented it from reaching significance.

Table 3: Results of the first set of MRA with three independent variables, one chosen from each domain ($p < 0.0001$).

Variables	*R*	*Adj.R²*	*F (3, 34)*
F0-rg; SPL; PVI	0.79	0.584	18.31
SPL; AR-syll; F0-s	0.79	0.591	18.82
F0-s; AR-pho; SPL	0.80	0.601	19.62
F0-rg; SPL; AR-pho	0.81	0.625	21.53

Again, the variables in individual lines are ordered according to the magnitude of the β-coefficients, with the highest one coming first. Clearly, the gain in explained variance in comparison with the best model in Table 2 is about 10 percent. It has to be noted that neither the gradient of F0 declination nor the vocalic proportion of an utterance (V%) take part in any of the combinations.

Table 4 presents results of regression analyses with three independent variables again, but this time the variables did not have to be one from each domain. The requirement that all three variables be significant was obeyed.

It was established that for all three variables to be significant, two of them had to be from the time domain: one representing the tempo of speech, the other rhythmicity. The

trend from all previous models for F0 parameters to have the highest β-coefficient if present in the model is maintained. Also similarly to the previous set of analyses, the articulation rate in syllables per second (AR-syll) caused a slight decrease in the power of the model compared with the articulation rate expressed in phones per second (AR-pho). However, as we are computing patterns on 38 cases only, these differences are actually negligible.

Table 4: Results of the second set of MRA with three independent variables, not representing each domain ($p < 0.0001$).

Variables	*R*	*Adj.*R^2	*F (3, 34)*
F0-grad; AR-pho; PVI	0.77	0.561	16.76
AR-pho; PVI; SPL	0.79	0.586	18.45
F0-rg; AR-pho; PVI	0.79	0.593	19.01

This can be observed in Table 5 where four independent variables are entered into the analysis and the fluctuations of the adjusted R^2 and F go in the opposite direction when AR-pho is replaced with AR-syll. What is more important is the fact that the explained variance, as captured by $Adj.R^2$, has still risen by about two to five percent compared with the best models with three independent variables.

Table 5: Results of multiple regression analyses with four independent variables ($p < 0.0001$).

Variables	*R*	*Adj.*R^2	*F (4, 33)*
AR-syll; PVI; SPL; F0-s	0.83	0.644	17.71
AR-pho; PVI; SPL; F0-rg	0.84	0.668	19.58
AR-syll; PVI; SPL; F0-rg	0.84	0.678	20.49

Although the significance of the model is always very high ($p < 0.0001$), the first analysis with F0 standard deviation (F0-s) had this particular measure (i.e., F0-s) insignificant in the model. The third model proved to be the strongest, and no other addition of any of the remaining variables yielded a significant improvement. In terms of b-coefficients then, the model occurs as follows:

$$FS = 8.54 - 0.53x_1 - 0.25x_2 - 0.1x_3 - 0.05x_4 \quad (5)$$

where x_1 is the articulation rate in syll·sec^{-1}, x_2 is the SPL difference between stressed and unstressed syllables, x_3 is the F0 range in semitones, and x_4 is the pairwise variability index of vocalic intervals expressed as a ratio hypothetically between 0 and 100, but in our material ranging only between 20.1 and 43.2.

In terms of β-coefficients, which show the relative importance of the variables in the model, the regression equation is:

$$FS_{norm} = -0.36x_1 - 0.34x_2 - 0.29x_3 - 0.28x_4 \quad (6)$$

Figures 1 and 2 capture the situation decomposed into two three-dimensional scatterplots. The regression lines were added manually and should only serve for illustration. The figures demonstrate why all the coefficients are negative. To reach higher foreignness score (i.e., stronger Czech accent in English) the speech has to be slower, with smaller differences between stressed and unstressed syllables, smaller pitch range, and smaller variation in durations of vocalic intervals in-between consonantal intervals.

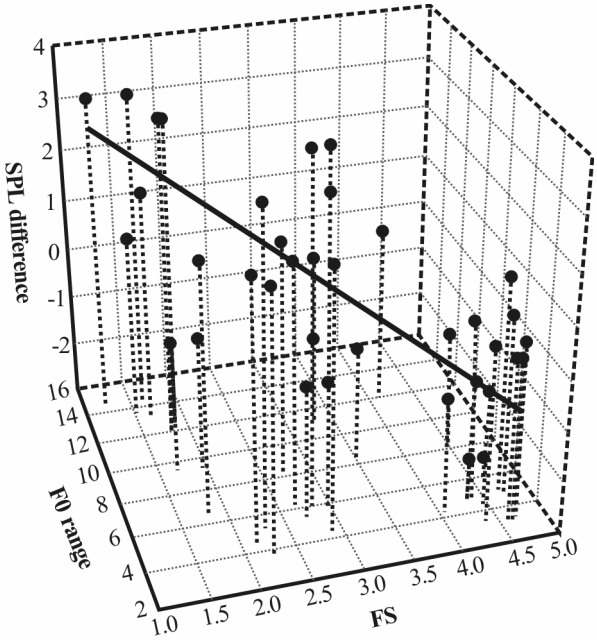

Figure 1: 3-D scatterplot of the relationship among foreignness scores (*FS*), F0 range and SPL differences from equations (5) and (6).

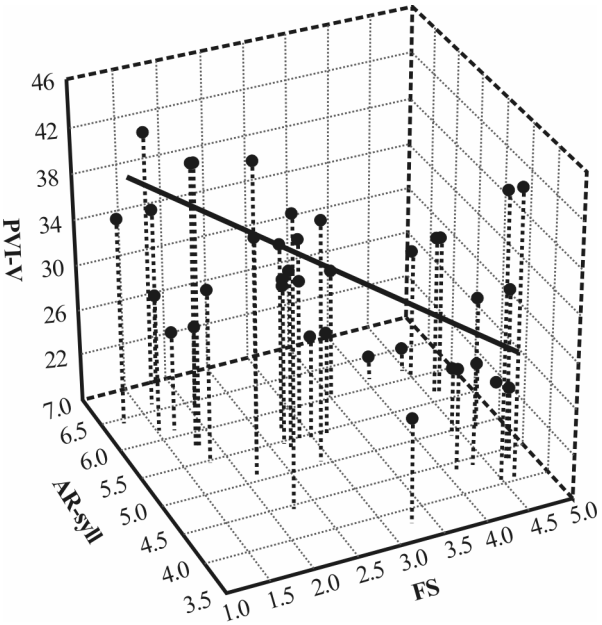

Figure 2: 3-D scatterplot for foreignness scores (*FS*), articulation rate in syllables per second and pairwise variability index from equations (5) and (6).

3.2. Degraded speech signal

The second major question of our study was the behaviour of the explanatory variables under adverse speech conditions. Tables 6 and 7 present results of analyses of simple situations

with two independent variables only, for filtered and masked speech respectively.

Table 6: Results of multiple regression analyses with filtered speech and two independent variables ($p < 0.0001$).

Variables	*R*	*Adj.R^2*	*F (2, 35)*
PVI; AR-pho	0.63	0.363	11.55
PVI; C-dur-s	0.66	0.400	13.32
PVI; V%	0.68	0.426	14.76
SPL; AR-pho	0.74	0.515	20.63
F0-grad; F0-rg	0.77	0.563	24.87
SPL; PVI	0.78	0.579	26.40
F0-rg; PVI	0.79	0.605	29.28
F0-grad; SPL	0.80	0.613	30.27
F0-grad; PVI	0.82	0.647	34.90
F0-s; PVI	0.82	0.653	35.80
F0-rg; SPL	0.85	0.704	44.91
F0-percrg; SPL	0.87	0.736	52.52
F0-s; SPL	0.87	0.748	55.97

Table 7: Results of multiple regression analyses with masked speech and two independent variables ($p < 0.0001$).

Variables	*R*	*Adj.R^2*	*F (2, 35)*
F0-rg; V%	0.67	0.417	14.21
AR-pho; SPL	0.67	0.418	14.31
F0-rg; AR-pho	0.68	0.438	15.39
F0-rg; SPL	0.69	0.441	15.57
PVI; SPL	0.69	0.442	15.63
F0-grad; AR-pho	0.69	0.442	15.68
F0-rg; PVI	0.74	0.521	21.15
PVI; C-dur-s	0.74	0.526	21.56
F0-s; V%	0.74	0.528	21.71
PVI; V%	0.75	0.534	22.23
F0-s; AR-pho	0.75	0.543	22.95
PVI; AR-syll	0.76	0.549	23.51
F0-percrg; V%	0.77	0.567	25.25
F0-s; PVI	0.77	0.570	25.48
F0-percrg; AR-pho	0.78	0.579	26.41
F0-grad; PVI	0.78	0.579	26.42
F0-percrg; PVI	0.78	0.584	27.00

Surprisingly, there were more significant results than in the case of clean speech, although the general sensitivity to foreignness seems to be lower: the foreignness scores occupy a smaller range. There are some new significant combinations compared with clean speech results. A new variable which did not bring significant results before is the variation in durations of consonantal intervals (C-dur-s). However, the normalized version of the same measure is not significant, which signals that it is actually a reflection of the articulation rate that causes the result: slower speakers have a greater standard deviation of consonantal interval durations.

It can be observed again that if there is an F0 parameter present in the combination then its β-coefficient is greater than that of the other variable. Despite the degradation of the speech signal, the explained variance is generally greater (*cf.* Table 2). Especially the cases of F0 variation in combination with SPL for filtered speech are very successful.

An addition of the third variable into the model brought about results summarized in Tables 8 and 9. Articulation rate did not qualify into any combination in the case of filtered speech and into one combination in the case of masked speech. Generally, the only significant combinations when all three domains were represented are those with F0 features (always the most important in the model), together with the SPL difference between stressed and unstressed syllables and the pairwise variability index of vocalic intervals. Masked speech again leads to a weaker model than filtered speech.

Table 8: Results of MRA with filtered speech and three independent variables (IV), one chosen from each domain ($p < 0.0001$).

Variables	*R*	*Adj.R^2*	*F (3, 34)*
F0-grad; SPL; PVI	0.86	0.719	32.50
F0-percrg; SPL; PVI	0.88	0.761	40.27
F0-rg; SPL; PVI	0.89	0.774	43.17
F0-s; SPL; PVI	0.89	0.779	44.53

Table 9: Results of MRA with masked speech and three IV, one chosen from each domain ($p < 0.0001$).

Variables	*R*	*Adj.R^2*	*F (3, 34)*
F0-rg; PVI; AR-pho	0.74	0.505	13.59
F0-rg; PVI; SPL	0.78	0.576	17.77

If we do not insist on representing each of the domains, the combinations of three variables that bring about significant results are still possible. We present them in Tables 10 and 11. The adjusted R^2 for filtered speech is in all cases lower than in Table 8 while the opposite holds for the masked speech (*cf.* Tables 9 and 11).

Table 10: Results of MRA with filtered speech and three IV, not representing each domain ($p < 0.0001$).

Variables	*R*	*Adj.R^2*	*F (3, 34)*
SPL; PVI; AR-pho	0.81	0.620	21.10
SPL; PVI; AR-syll	0.82	0.640	22.89
F0-rg; PVI; V%	0.83	0.662	25.18
F0-percrg; PVI; V%	0.84	0.680	27.26
F0-s; PVI; V%	0.85	0.690	28.44

Table 11: Results of MRA with masked speech and three IV, not representing each domain ($p < 0.0001$).

Variables	*R*	*Adj.R^2*	*F (3, 34)*
PVI; F0-rg; AR-pho	0.80	0.606	19.96
PVI; F0-grad; V%	0.81	0.624	21.44
PVI; AR-syll; SPL	0.81	0.630	22.04
PVI; F0-grad; AR-pho	0.81	0.632	22.19
F0-s; PVI; AR-pho	0.82	0.640	22.96
PVI; V%; F0-rg	0.82	0.642	23.10
F0-percrg; PVI; AR-pho	0.83	0.658	24.76
F0-s; PVI; V%	0.83	0.664	25.32
F0-percrg; PVI; V%	0.84	0.681	27.38

None of the four-variable models significant for clean speech was successful in the case of filtered speech and neither was any other combination of four variables. For masked speech

there were only two models with all four variables reaching significance. These are displayed in Table 12. Interestingly, the best model has the same amount of explained variance as the one generated for clean speech (cf. Table 5, last line) and the variables involved are also basically the same apart from the swap between F0 range and F0 standard deviation, which both express variation in F0 contours.

Table 12: Results of MRA with masked speech and four independent variables ($p < 0.0001$).

Variables	*R*	*Adj.R^2*	*F (4, 33)*
PVI; AR-pho; F0-rg; SPL	0.83	0.643	17.67
PVI; AR-syll; F0-s; SPL	0.84	0.673	20.08

In terms of b-coefficients then, the model occurs as follows:

$$FS = 6.41 - 0.04x_1 - 0.3x_2 - 0.36x_3 - 0.1x_4 \quad (7)$$

where x_1 is the pairwise variability index of vocalic interval duration, x_2 is the articulation rate in syllables per second, x_3 is the F0 standard deviation and x_4 is the SPL difference between stressed and unstressed syllables.

In terms of β-coefficients, the regression equation is:

$$FS_{norm} = -0.39x_1 - 0.32x_2 - 0.3x_3 - 0.22x_4 \quad (8)$$

3.3. Importance of explanatory variables

Individual variables occurred in successful models with unequal frequency. We summed up for each variable the number of occurrences in significant models. The result is displayed in Figure 3. This count is only illustrative since the two articulation rate measures used (AR-pho and AR-syll) are underrepresented in it: when they produced results that were too similar, only one of them was charted. Nevertheless, it is still a picture consistent with the rest of the analyses. It shows that amongst the F0 measures it is the standard deviation and range that work best. The difference between mean SPL of stressed and unstressed syllables was also a relatively successful measure but it was the only one of the three originally designed SPL measures that functioned. In the temporal domain the pairwise variability index of vocalic intervals and articulation rate seem to be most useful.

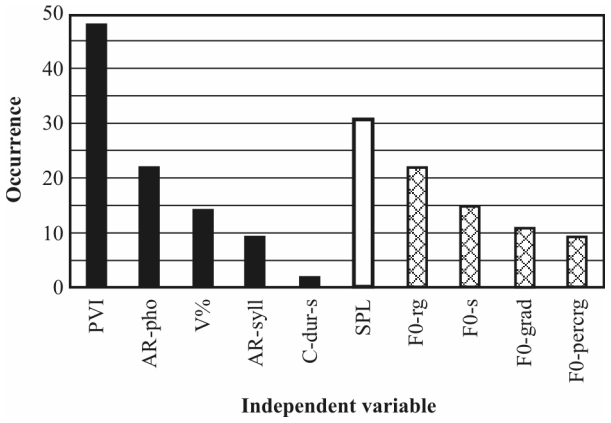

Figure 3: Numbers of occurrences of explanatory variables in models that reached significance.

At the most general level, the best models are those that explain most variance in the dependent variable. It is obvious from Tables 2 to 12 that the best models achieve 70 to 78 % explained variance (100*adjusted R^2). Interestingly, the first eight models with the highest explained variance relate to filtered speech. All of them contain an F0 descriptor and seven of them include an SPL measure. Five of them involve a PVI variable and none of them relies on articulation rate.

For preliminary testing, we took one native speaker (a professional BBC news reader) and three Czech speakers who seemed to have a different degree of foreign accent in their speech when we informally listened to them. Using the best five models we predicted the following scores for their speech:

$$1.59 < 2.38 < 2.44 < 3.10$$

The lowest score was computed for the native speaker, and the remaining three for Czech speakers in the order parallel to our pre-test impression. Given that the range of scores in the training set was 1.23 to 4.33, this preliminary result seems realistic and encouraging.

4. Discussion

We have shown that b-coefficients produced by multiple regression analyses can be used for the prediction of human evaluation of foreign accentedness concerning Czech English. As the Czech language is by no means exceptional or linguistically unique, we can expect that some, or perhaps all, of our findings can be applicable in other types of foreign accentedness as well.

There are several issues that should still be addressed. One of them is the reliability of the foreignness scores. Although we had three different sets of listeners (native English, native Czech experts and native Czech naive) and their scoring correlated very highly, it is reported in literature that the degree of foreign accent is influenced by a number of factors and is unstable across changing conditions. For instance, Flege and Fletcher [6] showed that listeners are stricter with the same foreign accented samples if items with native speakers are added to the experimental set. Moreover, when listeners familiarize themselves with the sentences to be assessed, they also become stricter. It would be useful for further research to have a better understanding of what individual levels of foreignness scores practically mean.

As to the predictors, F0 variation seems to be quite powerful. However, the importance of F0 range is a bit worrying since it hinges on two values only (the minimum and maximum), and these two values can be easily anomalous as a result of a measurement error or an unusual speech event. It would be better to find a way which strengthens the role of the F0 standard deviation instead.

In the area of rhythmic features, the PVI behaviour is in line with [18] who also found that for automatic discrimination between Czech and English the pairwise variability index of high-energy regions (corresponding to vowels or highly sonorous consonants) was lower in Czech than in English. Our current finding can be interpreted as a rhythmic interference of Czech in Czech English.

In future research, it might be useful to look at individual cases, especially those that are distant from regression lines, and to see if there is a possibility of improving the measures so that they match the human perception more accurately. For example, it seems that the articulation rate in either syllables

per second or phones per second could be refined with respect to prosodic boundaries in the sample: phrase-final lengthening distorts the overall articulation rate for the phrase and can be misleading. Another solution might be to calculate the tempo for a much larger stretch of speech as a global personal characteristic of the speaker.

Similarly, the SPL measure (i.e., the mean difference between the stressed and unstressed syllables) could be refined and possibly made more sensitive if the unstressed syllables with full vowels were left out and only reduced syllables (the so-called weak syllables) were measured. The usefulness of this step also remains to be determined in the follow-up research.

Filtered speech produced better results than both masked and clean speech. This fact remains unexplained. We can only speculate that as we were searching for suprasegmental cues and filtering suppressed the segmental ones the foreignness scores are more consistent with the input. Masked speech may have worked more poorly due to the competing speakers, who add some extra cognitive load, which in turn makes the foreignness scores less clear cut. Again, future research should clarify these issues.

5. Conclusions

Multiple regression coefficients reflecting suprasegmental properties of speech in the frequency, time and intensity domains are capable of reflecting the strength of Czech accent in English with respect to human evaluation of speech. The variation in F0 tracks, SPL difference between stressed and unstressed syllables, and PVI of vocalic intervals in speech appear to be the best predictors. Further refinement of the measures might produce even more robust models.

6. Acknowledgements

This work was supported by the European Union grant MRTN-CT-2006-035561 – *Sound to Sense*, and by the Czech Ministry of Education grant VZ MSM0021620825.

7. References

[1] Piske, T., MacKay, I.R.A. and Flege, J.E., "Factors affecting degree of foreign accent in an L2: a review", *Journal of Phonetics 29*, 2001, pp. 191-215.

[2] Abercrombie, D., "Teaching pronunciation", *English Language Teaching III/5*, 1949, pp. 113-122.

[3] Eskenazi, M., "An overview of spoken language technology for education", *Speech Communication 51*, 2009, pp. 832-844.

[4] Brennan, E.M. and Brennan, J.S., "Measurements of accent and attitude toward Mexican-American speech", *Journal of Psycholinguistic Research 10/5*, 1981, pp. 487-501.

[5] Rubin, D.L., "Non-language factors affecting undergraduate's judgments of non-native English speaking teaching assistants", *Research in Higher Education 33/4*, 1992, pp. 511-531.

[6] Flege, J.E. and Fletcher, K.L., "Talker and listener effects on degree of perceived foreign accent", *J. Acoust. Soc. Amer*., *Vol. 91/1,* 1992, pp. 370-389.

[7] Wu, T., Dochateau, J., Martens, J-P., and Van Compernolle, D., "Feature subset selection for improved native accent identification", *Speech Communication 52*, 2010, pp. 83-98.

[8] Flege, J.E., "Language contact in bilingualism: Phonetic system interactions", *Laboratory Phonology 9*, 353-382, Mouton de Gruyter, Berlin, 2007.

[9] Flege, J.E., Frieda, E.M. and Nozawa, T., "Amount of native-language (L1) use affects the pronunciation of an L2", *Journal of Phonetics 25*, 1997, pp. 169-186.

[10] Boula de Mareüil, P. and Vieru-Dimulescu, B., "The contribution of prosody to the perception of foreign accent", *Phonetica 63*, 2006, pp. 247-267.

[11] Fry, D.B., "Frequency of occurrence of speech sounds in Southern English". *Archives Neélandaises de Phonétique Expérimentale 20*: 103-106. 1947.

[12] Cruttenden, A., "Gimson's Pronunciation of English", 7th Edition. Hodder Education, London, 2008.

[13] Munro, M.J., and Derwing, T.M., "Foreign accent, comprehensibility and intelligibility in the speech of second language learners", *Language Learning 45/1*, 1995. pp. 73-97.

[14] Skarnitzl, R., Volín, J. and Drenková, L. "Tangibility of foreign accents in speech: the case of Czech English", *2nd Prague Conf. on Linguistics & Literary Studies: Proc.*: 11-20. 2005.

[15] Volín, J. and Skarnitzl, R., "The strength of foreign accent under adverse listening conditions" (submitted).

[16] Flege, J.E., Munro, M.J., MacKay, I.R.A., "Factors affecting strength of perceived foreign accent in a second language", *JASA 97/5*, 1995, pp. 3125-3134.

[17] Flege, J.E. "Second-language learning: The role of subject and phonetic variables", in *STiLL 98 – Proceedings*. ESCA, Stockholm: 1-8. 1998.

[18] Volín, J., Pollák, P., "The dynamic dimension of the global speech-rhythm attributes", *Proceedings of 10th Interspeech*: 1543-1546. 2009.

[19] Pfitzinger, H.R., "Local speech rate as a combination of syllable and phone rate", *Proceedings of 5th ICSLP, ISCA – Sydney*: 1087-1090. 1998.

[20] Grabe, E., and Low, E.L., "Durational variability in speech and the rhythm class", in: Gussenhoven, C., Warner, N. (Eds.), *Papers in Laboratory Phonology 7*, Mouton de Gruyter, Berlin: 515-546. 2002.

[21] Ramus, F., Nespor, M. and Mehler, J., "Correlates of linguistic rhythm in the speech signal", *Cognition 73*, 1999, pp. 265-292.

[22] Dellwo, V., "Rhythm and speech rate: a variation coefficient for ΔC", *Language and language-processing: Proc. of the 38th Linguistics Colloquium, Piliscsaba 2003.* Peter Lang, Frankfurt am Main: 231-241. 2006.

[23] Boersma, P., Weenink, D., *Praat: doing phonetics by computer* (Version 5.1). Retrieved Feb 10, 2009, from http://www.praat.org/.

[24] Lieberman, Ph. et al., "Measures of the sentence intonation of read and spontaneous speech in American English", *JASA 77/2*, 1985, pp. 649-657.

[25] Swerts, M., Strangert, E. and Heldner, M. "F0 declination in read-aloud and spontaneous speech", *Proceedings of 4th ICSLP*: 1501-1504. 1996.

[26] Streefkerk, B.M., Pols, L.C.W., Bosch, L., "Towards finding optimal features of perceived prominence", *Proceedings of 14th ICPhS*: 1769-1772. 1999.

[27] Volín, J., "Downtrends in standard British English intonation". Hector, Frankfurt am Main, 2008.