INTERSPEECH 2009 BRIGHTON

# Prosodic Analysis of Foreign-Accented English

*Hansjörg Mixdorff* [1] *and John Ingram*[2]

[1] Department of Informatics and Media, BHT University of Applied Sciences, Berlin, Germany
[2] School of EMSAH, University of Queensland, Brisbane, Australia

mixdorff@beuth-hochschule.de, j.ingram@uq.edu.au

## Abstract

This study compares utterances by Vietnamese learners of Australian English with those of native subjects. In a previous study the utterances had been rated for foreign accent and intelligibility. We aim to find measurable prosodic differences accounting for the perceptual results. Our outcomes indicate, inter alia, that unaccented syllables are relatively longer compared with accented ones in the Vietnamese corpus than those in the Australian English corpus. Furthermore, the correlations of syllabic durations in utterances of one and the same sentence are much higher for Australian English subjects than for Vietnamese learners of English. Vietnamese speakers use a larger range of f0 and produce more pitch-accents than Australian speakers.

**Index Terms**: foreign accent, prosodic analysis

10.21437/Interspeech.2009-666

## 1. Introduction

Although foreign accent is most readily associated with segmental deviations from the native norm, prosodic differences certainly account for many difficulties in understanding accented speech (see, for instance, [1][2]). In the current study we examine speech collected from Vietnamese learners of Australian English. In previous work the data have been assessed by native listeners for intelligibility and strength of foreign accent on a scale from 1 to 5 [3]. We now attempt to perform a prosodic analysis of the recordings and compare them with corresponding utterances by native Australian subjects in order to establish objective parameters that best reflect foreign accent, as well as are correlated with the subjective measures of foreign accent and intelligibility. Whereas English is often classified as a stress-timed language, Vietnamese is a syllable-timed tone language, a contrast which obviously poses a number of prosodic problems for learners of the other language.

## 2. Speech Material and Method of Analysis

The original corpus consists of recordings from a rephrasing task leading to 23 target sentences. The rephrasing was intended to deflect subjects' attention from their own speech to the linguistic aspects of the task, in order to obtain unselfconscious L2 pronunciation, while at the same time yielding content controlled target sentences for purposes of phonetic comparison between subjects. This elicitation strategy yielded more natural pronunciation, but a proportion of speakers failed to produce sentences that matched the intended target. The sentences were uttered by a total of 17 Vietnamese learners of English, 2 males, 15 females. In addition, the corpus contains recordings by three native Australian English speakers (all female) and two second-generation Vietnamese Australians (all female). These utterances were employed in a perceptual experiment with Australian English listeners yielding ratings with respect to accentedness and intelligibility. The listeners also had to note down in English the words they had understood. The part of this data uttered by Vietnamese speakers shall henceforth be referred to as *VIET*.

As a reference, a set of recordings produced by 14 Australian English speakers (all female) was prepared. This data set, plus the Australian utterances from the data set just mentioned, shall be referred to as *AUS*. All data had been collected through a web interface at a sampling rate of 8 kHz at 16 bits using the students' personal headsets and computers.

In a first step, all recordings were forced-aligned on the word and phone-levels using a trial version of the *LINGWAVES* UK English Forced Aligner [4]. The text targets were the expected results of the rephrasing task. Examination of alignment results, however, showed that the quality of most recordings was relatively poor due to low signal levels, strong low frequency drift, background noise (thermal and environmental) and distortion. Hence, the forced alignment procedure frequently produced errors and mismatches, as well as pause insertions. Other problems included signal truncations at the end (missing sounds or even words), wrong textual content (other than the intended targets), as well as hesitations and repairs.

Of the *VIET* data (total of 481 sentences available) 25 utterances were truncated, 81 were distorted, 7 exhibited repairs and 54 contained wrong textual content. Despite these flaws 461 utterances had actually been presented in the previous perception study [1] and yielded ratings for accentedness and intelligibility. Ultimately 398 utterances (including distorted ones that were otherwise correct), that is, 83% of the original total were used for further analysis. Of the *AUS* data original 298 utterances 259 were found usable; all others had one of the above-mentioned problems and were discarded from further processing.

Since we were primarily interested in the intonational and rhythmic properties of the accented speech, the forced alignment was re-run on the syllabic level by editing the phonetic transcriptions yielded from the text-based alignment in the first step to employ syllabic subdivisions.
Subsequently, the label files from the alignment procedure were converted to *PRAAT* TextGrid format [5] and combined in a single TextGrid containing syllable and phone labels. The syllabic boundaries were then hand-corrected and phone labels automatically adjusted in proportion to the syllable. At this stage we were not interested in the identity and exact boundaries of phones actually realized, but the rhythmic structure of the utterances.

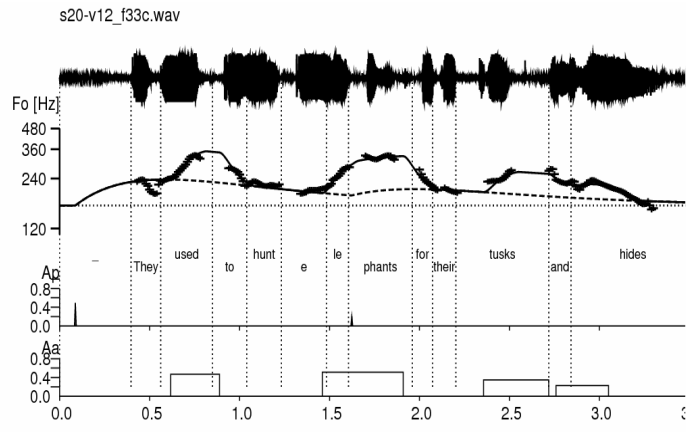


Figure 1: *Example of analysis of sentence 20, uttered by a Vietnamese speaker "They used to hunt elephants for their tusks and hides".*

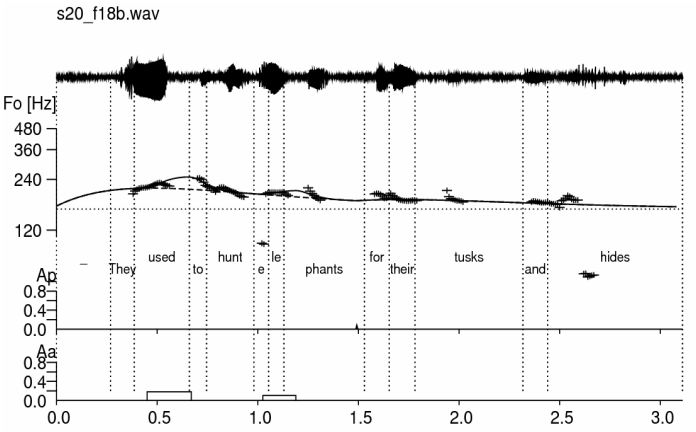


Figure 2: *Example of analysis of sentence 20, uttered by an Australian speaker of English.*

In order to analyse the intonational properties of the two corpora, *F0* values were extracted at a step of 10ms using the *PRAAT* default pitch extraction settings. Due to the strong low frequency drift in most of the data, we high-pass filtered at a cut-off frequency of 150 Hz which improved the performance of the *F0* extraction considerably.

A sub-corpus of four sentences was subjected to Fujisaki model [6] parameter extraction [7] as shown in Figure 1 (sentence 20, produced by Vietnamese speaker 12, a female). The figure displays from the top to the bottom: The speech wave form, the *F0* contour (+signs: extracted, solid line: model-based), the text, the underlying phrase and accent commands.

## 3. Prosodic Parameters and Results of Analyis

The objective of the analysis was to examine the relationships between Australian listeners' sentence-wise judgments on accentedness and intelligibility of the *VIET* data and objective prosodic speech parameters. Although it could be argued that the percept of foreign accent is mainly associated with errors on the segmental, that is, the phone level, other investigations indicate that prosody also plays an important role [1][2].

One type of error found in L2 utterances is the wrong placement of lexical stress. Therefore, a major design criterion guiding the development of the sentence set was the incorporation of stress-contrasts. Learners of English often fail to correctly mark the primary stress in compound words/word groups such as "**blue**-bell" (a type of flower) and "blue **bell**", as well as "**strong**-box" (safe) versus "strong **box**" (correct primary stress in bold type)[8].
Perceptual examination of sentences from the *AUS* as well as the *VIET* data, however, revealed that even the Australian speakers did not assign correct primary stress, producing "*strong-**box**" as well as "***blue** bell", for instance. An error exclusively found with the Vietnamese speakers was the realization "*news**pa**per". Otherwise there were not any errors regarding stress placement in the *VIET* data.

We looked at how the number of speech pauses in an utterance had influenced the percepts of accentedness and intelligibility and found that they are both significantly correlated, but more with the judgment of accentedness than intelligibility as shown in Table 1.

Table 1*: Correlation between number of pauses in sentence, intellegibility and perceived accentedness.*

| | | intelligibility | accentedness |
|---|---|---|---|
| number of pauses | correlation | .171** | .301** |
| | significance | .001 | .000 |
| | N | 387 | 387 |

For further quantitative analysis, the syllabic labels from the *VIET* and *AUS* corpora were compared with respect to mean and standard deviations as well as rhythmic properties of the utterances. Analysis showed a considerably higher syllable rate of 4.8 syllables/second for the Australian English speakers against 3.5 syllables/second for the Vietnamese learners of English. Within the *VIET* data, however, syllabic rate in a single sentence does not seem to be significantly correlated with judgments of accentedness and intelligibility.

We then investigated whether the syllable-timed property of Vietnamese as opposed to the stress-timed properties of English also affected the realizations of the Vietnamese learners.
First we divided the text sentences into metric feet and compared the durations of syllables which were the head of a foot with all others. For the *VIET* data we yielded mean durations for head and non-head syllables of 326 ms and 250 ms, whereas for the *AUS* data values were 249 and 181 ms, respectively. The resulting mean duration ratios (non-head/head) were of .767 and .728, hence the foot-based distinction did not yield a significant difference between Vietnamese and Australian English realizations.

Based on the text underlying the utterances we categorized all syllables as belonging to one of three classes: *Unstressed*, *potentially stressed* and *stressed*. This classification was mainly based on the super-ordinate part-of-speech properties. Whereas lexically stressed syllables in nouns, for instance, were classified as *stressed*, lexically stressed syllables of verbs were classified as *potentially stressed*, and all others as

*unstressed.* Although individual realizations could possibly be different, results showed that stressed and potentially stressed syllables in the *AUS* corpus were relatively longer than unstressed syllables than in the *VIET* corpus. The following table shows the results of comparison.

Table 2: *Mean and standard deviation of syllabic durations for unstressed, potentially stressed and stressed syllables.*

| group | syllable type | mean[ms] | s.d.[ms] | N |
|---|---|---|---|---|
| Vietnamese | unstressed | 226 | 119 | 1925 |
| | potentially stressed | 325 | 122 | 293 |
| | stressed | 381 | 138 | 944 |
| | total | 281 | 144 | 3162 |
| Australian | unstressed | 157 | 081 | 2256 |
| | potentially stressed | 246 | 065 | 352 |
| | stressed | 306 | 112 | 1087 |
| | total | 209 | 112 | 3695 |

The ratio of mean durations unstressed/stressed is .513 for the Australian English speakers whereas it is .594 for the Vietnamese speakers. The values are .707 versus .637 for the potentially stressed syllables.
This suggests that although Vietnamese speakers apply the rules of the English stress system they tend to produce syllables of more uniform lengths than the Australian English speakers.

Looking more closely at the rhythmic patterns of individual sentences we correlated the syllabic durations in one realization of a sentence with the syllabic durations in all the other realizations of the same sentence. The advantage of this approach is that the effect of the speech rate on this measure is rather small. This measure was previously used for evaluating the quality of a duration-predicting model in text-to-speech synthesis [9].
For easier comparison the duration of occasional short intra-utterance pauses is added to the duration of the syllable preceding that pause. This strategy can be justified by the fact that a pause in principle is an extreme case of the final lengthening usually observed in syllables preceding prosodic boundaries.

If $d_{A1}\ d_{A2}\ d_{A3}\ldots\ d_{AN}$ are the durations of syllables in a given N-syllable sentence produced by speaker A and $d_{B1}\ d_{B2}\ d_{B3}\ldots\ d_{BN}$ the durations of syllables produced by speaker B the correlation can be calculated as follows. The mean syllable durations for each utterance are given by

$$\bar{d}_A = \frac{1}{N}\sum_{i=1}^{N} d_{Ai} \qquad \bar{d}_B = \frac{1}{N}\sum_{i=1}^{N} d_{Bi} \tag{1}$$

Then the covariance *cv* is defined as

$$cv_{AB} = \frac{1}{N-1}\sum_{i=1}^{N}(d_{Ai} - \bar{d}_A)\times(d_{Bi} - \bar{d}_B) \tag{2}$$

The respective standard deviations are

$$sd_A = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(d_{Ai} - \bar{d}_A)^2} \tag{3}$$

$$sd_B = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(d_{Bi} - \bar{d}_B)^2} \tag{4}$$

The resulting correlation is then given as

$$corr_{AB} = \frac{cv_{AB}}{sd_A \times sd_B} \tag{5}$$

which expresses the correlation between two particular utterances. By averaging over all M utterances of a given sentence for each group (*VIET, AUS*) we can yield a measure of similarity within that group, for instance,

$$\overline{corr}_{AUS} = \frac{1}{M}\sum_{i=1}^{M} corr_i \tag{6}$$

as well as the mean inter-group correlation. Results indicate that the Australian English realizations (mean ρ=.883) are much more similar in their rhythmic structure (more highly correlated) than the Vietnamese ones (mean ρ=.731). Also the cross-correlation between the two groups is rather low (mean ρ=.741).
In order to test whether the sentence-based correlations we had found were valid indicators of foreign accent we calculated the centroid of all Australian utterances for each sentence. That is for each syllable in a given sentence we averaged over all observed instances in the Australian data set, yielding prototypical syllabic durations for each sentence. Subsequently we calculated the correlations between each of the Vietnamese utterances and their corresponding Australian duration norm. Statistical analysis showed that this rhythmic correlation was significantly (ρ=-.173 $p < .01$) correlated with the Australian listeners' judgment of foreign accent, however, not correlated with their judgments of intelligibility. This is a fairly high value considering that it only concerns durational characteristics of the foreign-accented utterance.
Another measure of rhythmic differences between languages widely used is the so-called normalized pair-wise inter-variability index *npvi* [10] which examines the duration difference between consecutive syllables or syllable parts. Since we do not yet have a complete phonetic annotation of the speech data we only calculated the syllable-based scores defined as:

$$npvi = 100\times\frac{1}{N-1}\sum_{i=1}^{N-1}\frac{\left|(d_i - d_{i+1})\right|}{0.5\times(d_i + d_{i+1})} \tag{7}$$

The denominator corrects the term for speech rate variations henceforth the index is called "normalized". The *npvi* for the *VIET* data of 61.54 is only slightly different from the 63.99 in the *AUS* data and therefore does not present a consistent parameter for separating the two data sets. Its tendency suggests less inter-syllabic variability in the Vietnamese data. We have to take into account however that the Vietnamese data spans a range of proficiency levels so that some speakers might already attain a very high rhythmic proficiency whereas others exhibit almost "Vietnamese" rhythm.

The extracted *F0* contours were modelled using the Fujisaki model in order to establish the differences between the *AUS* and *VIET* data set. To this effect automatic parameter extraction was performed on utterances of sentences 20-23 [7].

Then the analysis results were inspected and if necessary corrected using the interactive *FujiParaEditor* [11].
The numerical results of the analysis are displayed in Table Table 3. It shows means and standard deviations of accent command amplitude and duration for the Vietnamese and Australian data. As can be seen - though mean durations are quite similar - the Vietnamese employ *F0* much more for marking accented syllables than the Australian English speakers. This is reflected by the higher values of accent command amplitudes *Aa*.

Table 3: *Mean and standard deviation of accent command amplitude Aa and accent command duration.*

| Group | | *Aa* | duration [ms] |
|---|---|---|---|
| Australian | Mean | .16 | 272 |
| | s.d. | 07 | 136 |
| | N | 175 | 175 |
| Vietnamese | Mean | .24 | 262 |
| | s.d. | .15 | 138 |
| | N | 250 | 250 |

If we look at the frequency of accent commands there are 1.09 commands per second in the Australian group but 1.43 for the Vietnamese group. The syllable-based frequency is one command every 2.62 syllables in the *VIET* group but one command every 4.47 syllables in the *AUS* data.

Table 4 shows means and standard deviations for the phrase command magnitude *Ap* which indicates the amount of *F0* reset taking place at the onset of a new phrase.

Table 4: *Mean and standard deviation of phrase command magnitude Ap.*

| group | mean | standard deviation | N |
|---|---|---|---|
| Australian | .268 | .164 | 90 |
| Vietnamese | .326 | .173 | 97 |
| Total | .298 | .171 | 187 |

It is obvious that the Vietnamese speakers adjust their declination line more strongly which is an indication that they employ a larger *F0* range when they talk than the Australian English speakers. They also rephrase more frequently on the average once every 6.75 syllables compared to 8.69 syllables for the Australian speakers. This result however might also be partly due to the higher speech rate of the Australians. Figure 2 shows an utterance of sentence 20 produced by an Australian speaker (compare with Figure 1). Note also the extremely low *F0* (creaky-voice) on the syllables ‘e’ and ‘hides’. This phenomenon is very frequent in the Australian data but almost non-existent in the Vietnamese.

## 4. Discussion and Conclusions

The current study concerned the prosodic analysis of accented English speech data produced by Vietnamese learners. We found that the number of pauses in an utterance, a typical indicator of disfluency, is strongly correlated with the percept of foreign accent, but less so with perceived intelligibility. On the rhythmic level Vietnamese learners of English produce relatively longer unaccented syllables than Australian English speakers, which suggests, that their rhythm is influenced by the syllable-timed structure of Vietnamese. The syllabic durations in the Australian English groups are more uniform than within the Vietnamese group expressed by the durational correlations between individual productions of the same sentence. The pair-wise variability index of syllabic durations, however, does not differ considerably between the two groups.
On the intonational level Vietnamese speakers produce stronger excursions of *F0* and use a wider range of *F0* than the Australian English controls. They place pitch-accents more frequently and exhibit much less vocal fry than their Australian English counterparts.
Future work will concern perceptual experiments with segmentally and prosodically manipulated stimuli in order to examine which factors contribute most to the percepts of strong foreign accent and reduced intelligibility. Furthermore, we will test whether our findings can be applied to enhance computer-aided pronunciation training.

## 5. Acknowledgements

This work was supported by DFG/ARC collaboration grant no. 447 AUS-113/28/0-1, as well as a UQ travel grant.

## 6. References

[1] Anderson-Hisieh, J., Johnson, R. and Koehler, K. “The relationship between native speakers judgements of nonnative pronunciation and deviance in segmentals, prosody and syllable structure”, Language Learning 42: 4 529-555, 1992.
[2] Magen, H.S., “The perception of foreign-accented speech”, Journal of Phonetics, vol. 26, 381-400, 1998.
[3] Nguyen, T. and Ingram, J., “A corpus-based analysis of transfer effects and connected speech processes in Vietnamese English”, Proceedings of the Tenth Australian International Conference on Speech Science & Technology, Sydney, Australia, 2004.
[4] www.wevosys.com
[5] www.praat.org.
[6] Fujisaki, H. and Hirose, K., “Analysis of voice fundamental frequency contours for declarative sentences of Japanese”, Journal of the Acoustical Society of Japan (E) 5(4) 233-241, 1984.
[7] Mixdorff H., “A novel approach to the fully automatic extraction of Fujisaki model parameters”, Proceedings of ICASSP 2000, vol. 3, 1281-1284, Istanbul Turkey, 2000.
[8] Pittam J. and Ingram J., “Accuracy of perception and production of compound and phrasal stress by Vietnamese-Australians”, Applied Psycholinguistics, 13(1), 1-12**,** 1992.
[9] Mixdorff H. and Jokisch, O., “Evaluating the quality of an integrated model of German prosody”, International Journal of Speech Technology 6(1): 45-55, 2003.
[10] White, L., & Mattys, S.L., “Calibrating rhythm: First language and second language studies”, Journal of Phonetics, 35*,* 501-522*,* 2007.
[11] http://public.tfh-berlin.de/~mixdorff/thesis/fujisaki.html