Subjective and Objective Evaluation of the Prosody of English Spoken by French Speakers: the Contribution of Computer Assisted Learning

Nadine Herry, Daniel Hirst

CNRS Laboratoire Parole et Langage, Université de Provence, Aix-en-Provence, France
nadine.herry@wanadoo.fr, daniel.hirst@lpl.univ-aix.fr

Abstract

This paper describes preliminary results from an ongoing project on the subjective and objective evaluation of the prosody of English spoken by French speakers making use of a system of computer-assisted learning for prosody. The system was tested for 6 months and data was analysed with two objectives: on the one hand the subjective evaluation of the prosody of English spoken by French speakers in order to determine the system's efficiency; on the other hand an objective evaluation attempting to establish a correlation between the level of a French speaker and a number of automatically extracted prosodic parameters. Although the critical statistical interactions we sought did not reach the level of significance, a number of effects suggest that both aspects of the project merit further investigation.

1. Introduction

There is a considerable literature on the problem of the acquisition of prosody in a second language and the methods used for aiding this. There has, however, been far less investigation of the possibilities of computer-assisted learning of prosody.

Lane and Buiten [12] studied the acquisition of prosody and imagined an automatic system to evaluate it. The system was apparently not very efficient and students made no progress in their oral capacity. Vardanian [16] tried something similar for teaching English intonation to Brazilian students but with a better visualiser. The students had the possibility of comparing their own production with the model. For three weeks, a control group tried to learn 6 intonation patterns just by imitation, while the experimental group used both imitation and visualisation. Despite the improved visualiser no significant difference was observed between the two groups. James used Ph. Martin's pitch visualiser [13] to test the effect of visual feedback on the acquisition of prosodic patterns for English students learning French, and he concluded that "one fact that did merge clearly was the efficacity of visualisation patterns in the field of applied phonetics and the teaching of intonation" (242).

[8] describes the development of *Prosodia*, a computer-assisted system for teaching English prosody to French students. It has been tested for 6 months and was evaluated in two ways. A subjective evaluation of the prosody of English spoken was carried out on an experimental group and a control group in order to evaluate the efficiency of the system. At the same time the relation between the subjective evaluation and a number of objective acoustic parameters was examined with a view to establishing an objective evaluation.

The method was developed in collaboration with CNRS and University of Provence, was financed by the Ministère Français de l'Education Nationale, de la Recherche et de la Technologie. The method is based on a simplified version [6].of the "tune" approach to British English intonation patterns developed by O'Connor and Arnold [14] see also [9].

2. Experiment

2.1. Data

The corpus (500 sentences) was recorded by two native speakers (one male and one female). Exercices were built combining both segmental and accentual problems (taken from [4][5][6][7] in the form of minimal pairs such as

Why choose! *vs* . White shoes!

Look at that blackbird *vs*. Look at that black bird

These were pronounced with one of 5 different intonation patterns (High Jump, Glide Up, Dive, Take Off, Glide Down), together with an indication of the attitude intended (annoyed, reassuring, contradicting etc.). The position of the nucleus, the length of the sentence and its segmental difficulties were varied systematically.

2.2. Subjects

20 second year students of English were trained with this material. Half of the students constituting the experimental group used a prototype of the Prosodia software. The other half of the students constituting the control group worked with the same material in a traditional language laboratory. All the students were volunteers.

2.3. Procedure

All the students trained for 6 months. Two tests (December and April) were organised in a language laboratory for the two groups. They were shown the test material 10 minutes before the beginning of the test. For the first test, the students had no formal training and had not studied intonation at all. The sentences were presented preceded by a few sentences providing an appropriate context. For the second test in April, when the students had had explicit training in producing specific intonation patterns, these were simply identified with what were to them, by then, familiar labels (High Jump, Glide Up etc.).

The 20 students were evaluated by an expert. Each student received 4 marks (on a twenty point scale) corresponding to:

(1) the quality of the vowels,
(2) the quality of consonants
(3) the quality of production
(4) the quality of repetition

To these we added:

(5) the average of mark 3 and 4
(6) the average of mark1,2,3 and 4.

We also took into account the marks the students received for their phonetics exam in June (one global mark on a twenty point scale).

2.4. Acoustic analysis.

The students' productions and the models were digitised and manually labeled using the Praat software [1]. A comparison between the students' productions and the models. [8] brought to light a number of acoustic parameters which appeared to be systematically different. These concerned differences in rhythm and in pitch.

The following parameters for each student at each date were subsequently extracted from the data by means of a Praat script:

Rhythm:

- percentage duration of vowels
- average consonant duration
- standard deviation of consonant duration
- coefficient of variation of consonant duration
- average vowel duration
- standard deviation vowel duration
- coefficient of variation of vowel duration
- percentage of number of vowels
- difference (in percentage) between the average sentence duration of the students and the average duration of the 2 native speakers
- difference (in percentage) between the standard deviation of intensity of the students and the average standard deviation of the 2 native speakers
- difference (in percentage) between the coefficient of variation of intensity of the students and the average coefficient de variation of the 2 native speakers

Pitch:

- difference (in percentage) between the range of F0 of the students and that of the 2 native speakers,
- the difference between the standard deviation of F0 of the students and that of the 2 native speakers,
- the difference between the coefficient of variation of F0 of the students and that of the 2 native speakers,
- the difference (in percentage) between the F0 variation in slope of F0 of the students and that of the 2 native speakers.

The rhythm parameters were chosen so that the typological differences between so-called 'stress-timed' languages like English and 'syllable-timed' language like French might be characterised using the parameters analysed in [15] together with a number of other parameters which have been shown [3] to be correlated with these distinctions. For the pitch characteristics we chose those parameters which seemed most typical of the differences between the students' productions and those of the models.

3. Statistical Analysis

In order to compare the subjective evaluation of the experimental group with that of the test group we carried out an ANOVA test using the student's marks as dependent variables and the group and date of the tests as independent variables..

For the objective evaluation we carried out a CART analysis of the prosodic parameters To do so we used a statistical software CRUISE (Classification Rule with Unbiased Interaction Selection and Estimation) [11] to predict the 6 marks of the April test according to the 15 prosodic parameters described above.

3.1. Results of the subjective evaluation

The ANOVA test showed that the experimental and control groups improved their marks for all four categories: the quality of vowels, the quality of consonants, the quality of production and the quality of repetition.

For mark 1, the quality of vowels, the group effect was significant $F(1,36) = 10.762$, $p = 0.0023$. The date effect showed a tendancy but did not quite reach significance $F(1,36) = 3.256$, $p = 0.0796$ (table 1).

Table 1: Anova table for mark 1 (quality of vowels)

Tableau d'ANOVA pour NOTE1V
Critère d'inclusion : Critère 1 de preparationdonne10.4.01.txt (importé).svd

	DDL	Somme des carrés	Carré moyen	Valeur de F	Valeur de p
GP	1	40,000	40,000	10,762	,0023
DATE	1	12,100	12,100	3,256	,0796
GP * DATE	1	2,500	2,500	,673	,4175
Résidus	36	133,800	3,717		

4 cas omis (manquants).

Figure 1 shows the average improvement in the quality of vowels between December and April for the two groups.

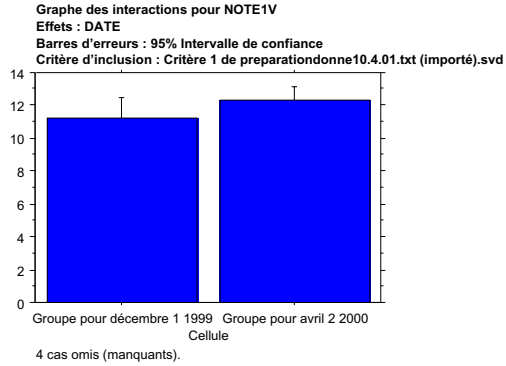

Figure 1: Date effect for mark 1 (quality of vowels)

For mark 2 (the quality of consonants) the group effect was significant $F(1,36) = 12,.764$, $p = 0.0010$ (table 2)

Table 2: Anova table for mark 2 (quality of consonants)

Tableau d'ANOVA pour NOTE2C
Critère d'inclusion : Critère 1 de preparationdonne10.4.01.txt (importé).svd

	DDL	Somme des carrés	Carré moyen	Valeur de F	Valeur de p
GP	1	60,025	60,025	12,764	,0010
DATE	1	7,225	7,225	1,536	,2232
GP * DATE	1	1,225	1,225	,260	,6129
Résidus	36	169,300	4,703		

4 cas omis (manquants).

It appeared from this that the experimental group (with an average of 11.8/20) was globally significantly better than the control group (9.3/20).

Table 3: Averages of the experimental and control groups for mark 2 (quality of consonants)

Tableau des Moy. pour NOTE2C
Effets : GP
Critère d'inclusion : Critère 1 de preparationdonne10.4.01.txt (importé).svd

	Nombre	Moy.	Dév. Std	Err. Std
exp	20	11,800	2,587	,579
temoin	20	9,350	1,631	,365

4 cas omis (manquants).

Marks 1 and 2 show a higher level for the experimental group. For mark 3 (quality of production) the date effect was significant F (2,54) = 28.929, p< 0.0001 (table 4,figure 2).

Table 4: Anova table for mark 3 (quality of production)

Tableau d'ANOVA pour NOTE3prod

	DDL	Somme des carrés	Carré moyen	Valeur de F	Valeur de p
GP	1	9,204	9,204	1,200	,2781
DATE	2	443,658	221,829	28,929	<,0001
GP * DATE	2	27,708	13,854	1,807	,1740
Résidus	54	414,075	7,668		

6 cas omis (manquants).

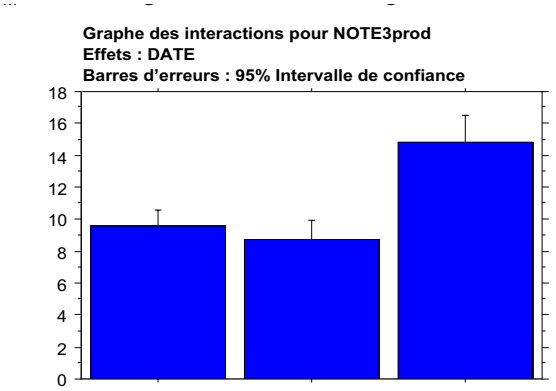

Figure 2: Effect of the date factor on mark 3 (quality of production).

There was no difference between December 1999 and April 2000 but a considerable difference between these two dates and the marks of June, that is between the training period and the final exam. We note a difference of five points (10/20: 15/20) in the quality of production.

Mark 4 (quality of repetition) shows that date and group effects are significant (table 5)

Table 5: Anova table for mark 4 (quality of repetition)

Tableau d'ANOVA pour NOTE4rep
Critère d'inclusion : Critère 1 de preparationdonne10.4.01.txt (importé).svd

	DDL	Somme des carrés	Carré moyen	Valeur de F	Valeur de p
GP	1	42,025	42,025	6,207	,0175
DATE	1	291,600	291,600	43,067	<,0001
GP * DATE	1	9,025	9,025	1,333	,2559
Résidus	36	243,750	6,771		

4 cas omis (manquants).

The date effect F(1,36) = 43.067, p < 0.0001 showed that there was improvement in the quality of repetition between December and April.

The group effect F (1,36) = 6.207, p = 0.0175 showed that the experimental group achived higher marks at repetition. The method (with and without the software) seem to have a positive effect.

For mark 5 (average of mark 3 and 4), the date effect was significant F (1,36) = 10.932, p = 0.0021 (table 6)

Table 6: Anova table for mark 5 (average of mark 3 and 4)

Tableau d'ANOVA pour NOTE5moyinto
Critère d'inclusion : Critère 1 de preparationdonne10.4.01.txt (importé).svd

	DDL	Somme des carrés	Carré moyen	Valeur de F	Valeur de p
GP	1	17,227	17,227	3,762	,0603
DATE	1	50,064	50,064	10,932	,0021
GP * DATE	1	16,577	16,577	3,620	,0651
Résidus	36	164,869	4,580		

4 cas omis (manquants).

Both groups seem to have improved their capacity for repetition and production.

The group effect showed a similar tendancy but this again did not reach significance F (1,36) = 3.762, p = 0.0603.

For mark 6 (average of mark 1,2,3,4) both the the date effect F (1,36) = 8.313, p = 0.0066 (table 7) and the group effect F (1,36) = 9.351, p = 0.0042) were significant.

Table 7: Anova table for mark 6 (average of mark 1,2,3,4)

Tableau d'ANOVA pour NOTE6moygle
Critère d'inclusion : Critère 1 de preparationdonne10.4.01.txt (importé).svd

	DDL	Somme des carrés	Carré moyen	Valeur de F	Valeur de p
GP	1	30,625	30,625	9,351	,0042
DATE	1	27,225	27,225	8,313	,0066
GP * DATE	1	5,625	5,625	1,718	,1983
Résidus	36	117,900	3,275		

4 cas omis (manquants).

These results show that the training method used by both groups produced a positive effect. The experimental group was globally better than the control group and their was no significant interaction between the two effects which would have allowed us to conclude that the computer-assisted training was more efficient than the classical training.

3.2. Results for the objective evaluation

The statistical analysis with CRUISE gives a regression tree with intervals for each mark. However, Cruise enabled us to

predict students' marks only between 9 and 17 out of twenty. It determined the prosodic parameters, their degree of importance and their intervals for the 6 marks. For example for mark 3 (quality of production) (figure 3) Cruise predicts two levels i.e. students between 5 –8/20and 9-12/20.

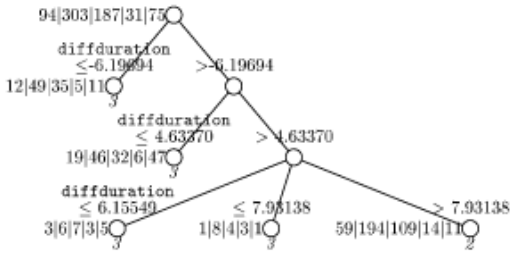

Figure 3: Regression tree for predicting mark 3 (quality of production).

64 per cent were predicted in level 2 (marks from 5 to 8) and 42 per cent in level 3 (9 to 12). If the difference of duration is ≤ 8% they are assigned to level 3 and if >8% to 2 . This suggest that the rate of articulation of the students is directly correlated with their score. For more discussion see [8] chapter 5.].

For all the marks except mark 4 (repetition) the sentence duration appeared to be the most highly correlated factor. For mark 4 the most significant parameter was the range of F0., For marks 1,2,5, and 6 parameters such as the coefficient of variation of consonant duration, range of F0, slope of F0, and standard deviation of F0 were selected by the software as the most discriminant parameters.

4. Discussion

Despite our conviction that a method such as Prosodia, based on perception and visualisation simultaneously, can prove an effective aid for the improvement of a student's prosodic capacity, we have not been able to demonstrate objectively that this is the case. One of the reasons for this was a global difference in level between the test group and the control group, which can possibly be attributed to a higher level of motivation on the part of those who volunteered for the experimental group. It is also possible that the experimental group was too small for the expected interaction to appear.

The results of the objective evaluation show that articulation rate by itself accounts for quite a large proportion of the difference between students' marks. Such a parameter is obviously a cover term for a large quantity of other features which characterise these productions. It is altogether quite probable that a number of other parameters will need to be tested.

The search for a clear demonstration of the efficiency of computer assisted tools needs consequently to be pursued actively, as does the attempt to carry out an objective evaluation of the students' productions.

Acknowledgements

Part of this research was financed by the Ministère Français de l'Education Nationale, de la Recherche et de la Technologie.

References

[1] Boersma, P.; Weenick, D. 1992-2002. Praat. A system for doing phonetics by computer. http://www.praat.org.

[2] Canto, H., 1988. Une expérience d'enseignement de l'intonation du Français à des Anglophones. *Toronto Working Papers* 9, 48-70.

[3] Cruz, R., 2000. Analyse acoustique et phonologique du portugais brésilien. Doctoral thesis, Université de Provence.

[4] Duchet, J.L., 1991. *Code de l'anglais oral*. Paris: Ophrys.

[5] Duchet, J.L.; Deschamps, A.; Fournier, J.M.; O'Neil, M., 2000 *Manuel de phonologie de l'anglais.* Paris: Didier-Erudition, CNED.

[6] Ginésy, M., 1995 *Mémento de Phonétique anglaise.* Paris: Nathan.

[7] Ginésy, M., 2000 *Phonétique et phonologie de l'anglais*. Paris: Ellipses.

[8] Herry, N., 2001. *Evaluation subjective et objective de la prosodie anglaise parlée par des français: Apport de l'enseignement assisté par ordinateur*. Doctoral thesis, Université de Provence.

[9] Hirst, D.J., 1998. Intonation in British English. In D.J. Hirst & A. Di Cristo (eds) *Intonation Systems. A Survey of Twenty Languages.* Cambridge: Cambridge University Press.

[10] James, E., 1977. The Acquisition of a Second-Language intonation Using a Visualizer. *Canadian Modern Language Review*, 33, 4, 503-506.

[11] Kim, H.; Loh, W.-Y., 2001. CRUISE User Manual. (Technical Report 989) March 3 1998, revised November 10, 2001. Department of Statistics, University of Wisconsin, Madison. http://www.wpi.edu/~hkin/cruise/

[12] Lane, H; Buiten, R., 1965. *A preliminary manual for the speech auto-instructional device*. (Progress report N° 5). Behavior Analysis Laboratory, University of Michigan.

[13] Martin, P., 1973 Les problèmes de l'intonation: recherches et applications. *Langue française*, 19, 4-32.

[14] O'Connor, J.D.; Arnold, 1961. *Intonation of colloquial English.*. London: Longman. (2nd edition 1973).

[15] Ramus, F.; Nespor, M.; Mehler, J., 1999. Correlates of linguistic rhythm in the speech signal. *Cognition*, 72, 1-28.

[16] Vardanian, R., 1964. Teaching English through oscilloscope displays. *Language Learning* 3-4, 109-117.