Learning the prosodic structure of a foreign language with a pitch visualizer

Philippe Martin

EA 3967, UFR Linguistique, Université Paris Diderot, Paris, France

philippe.martin@linguist.jussieu.fr

Abstract

This paper presents a new version of the pronunciation software program WinPitch LTL. This completely redesigned version is much more easier to use by the teacher and learner alike, and contains the regular function of prosodic real time display, variable speed playback, prosodic morphing, on screen teacher comment display, etc. New features include video capabilities (the program can process multimedia files in most formats) as well as automatic alignment of the learner's imitation on the teacher's models. This allows for an automated comparison and explanation on the differences analyzed on the segmental and suprasegmental levels.

The new version of the software has a companion program for the preparation of lessons in any language (the program is Unicode compliant) allowing easy navigation by the learner between examples contained in each unit. Segments of prosodic curves can be highlighted in any color and text easily added for on screen explanation of specific melodic or rhythmic properties of the model.

1. Introduction

The history of computer aided learning of pronunciation of a foreign learned language traces back at least to 1964 (Vardanian, 1964), with the first attempt to display the learner voice pitch on a screen in order to improve the perception and the realization of sentence prosody. In these heroic days, a real time display of the learner's fundamental frequency was obtained through a sophisticated mechanism involving a rotating radar screen…

The main underlying idea of these realizations is based of the fact that learners are supposed to be "deaf" to the phonological system to be acquired (as their perception is mainly driven by their mother tongue phonological grid), and that their pronunciation could be tremendously improved if the prosody of their realization was approaching the target language prosody.

If learners were considered "deaf" to certain phonological or phonetic features, giving them supplementary graphical information was felt at the times to be highly beneficial. These assumptions reveal the paramount importance given to sentence prosody in the pronunciation of a foreign language as well as the benefit of graphical input to supplement auditory perception. Real time visual feedback pertaining to the three main parameters of prosody, fundamental frequency, intensity and syllable duration, became an essential part of CAL development to acquire the prosodic features of a learned language.

Many developments of hardware and software implementing these ideas appeared in the last 30 years, along the progress made in computer technology. Among the most notable, we can cite Madsen (1973), VisiPitch (1975), Pitch Computer (1978), Speech Viewer (1985), for hardware devices, and WinPitch (1996), Speech Tutor (2003) for software packages.

2. Learning prosody

All these realizations proposed to the learner an imitation of some model intonation curve, without reference to any phonological or phonetic fact (with the possible exception of sentence modality supposedly encoded by the sentence final intonation contour). This drill method approach (O'Connor and Arnold, 1973) does not pertain to any linguistic knowledge of intonation, which may explain why the use of pitch visualizers did not become very popular in the course of three decades of development, as their effectiveness was felt often debatable (James, 1976, De Bot, 1983).

Another aspect that hampered a large use of these devices is the apparent complexity of the so called pitch curve, an acoustical estimation of the time evolution of the laryngeal frequency. Not only these curves reflect a particular phonetic realization of some intonation model, but graphic details displayed on screen are often not pertinent linguistically or simply not perceived by the listeners. Better ways of displaying the prosodic information have been sought (e.g. Spaai and Hermes, 1993), but only in integrating perception effects in displaying the prosodic information.

Another problem in the development of teaching applications is linked to the graphic emphasis done on pitch at the expense of rhythm, which is another important component of prosody. Speech synthesis experiments have shown that in some cases rhythm is more important for comprehension than pitch. An example of application that puts emphasis on rhythm rather than pitch can be found in Delmonte, Petrea and Bacalu (1997).

3. Principles for implementation

The development of WinPitch LTL as a tool to teach the prosody (and possibly other phonetic features) of a foreign language was conducted along the following well known pedagogical line going to passive to active learning.:

1. I hear and I forget.
2. I see and I remember.
3. I do and I understand.

3.1. I hear and I forget

"I hear and I forget" corresponds to the situation found in early language laboratories: the learner simply repeats the model heard, and sentences can be organized in sequences

reflecting the acquisition of a particular point of pronunciation (e.g. the mute e or vowel-vowel linking in French). Prosodic aspects were often limited to the location of lexical stress and the correct pronunciation of stress groups. Only in tone languages such as Mandarin would the realizations of pitch acquire some phonological significance in the drills offered to the learner.

3.2. I see and I remember

"I see and I remember" is reflected by the advent of pitch visualizers, displaying, sometimes in real time, a pitch curve model to be imitated by the learner. The advantage on the simple "listen and repeat" approach pertains to the assumed phonological deafness of learners, who can now see what they may not hear. The possibility to slow down the speech rate of the model constitutes a further improvement as to allow the correlation between auditory and visual perception: the learner can now link perceptively the graphic movement of pitch on the screen with the perception of it (through the synchronous displacement of a screen cursor for instance). In real time, the visual correlation between the perceived sound and a moving cursor is almost impossible to achieve.

3.3. I do and I understand

Through the modification of the prosodic parameters (fundamental frequency, intensity, syllable duration, rhythm and pauses) of their own voice with re-synthesis techniques such as PSOLA, the learner is now able to manipulate graphically his/her production and therefore get a direct understanding of how to achieve the prosodic movements required on their own voice, without actually performing it themselves. This corresponds to a "learning by doing" learning process. It is thus possible with this approach to design lessons to acquire the proper realizations of pitch events linked to various prosodic structures of a foreign language.

3.4. Automated feedback

The use of pitch visualizers can be in presentia, with the possible comments of an instructor who can directly intervene to guide the learner in the process, or in absencia with the software automatically bringing its own comments on a particular learner performance. In commercial realization, this feedback is more than often limited to a score, with no or very little comment of the type and place of errors made in the learner imitation. Obviously this feedback should reflect specific points of an (underlying) intonation theory, describing prosodic structures for instance.

4. Phonetics and phonology

What lacks in most if not all these realizations is a clear reference to the linguistic content (phonological and/or phonetic). Learners were put in front of a computer screen displaying information that was hard or impossible to link to any coherent intonation theory. The acquisition of Mandarin tones is a good counter example where the role of pitch is clearly phonological. Therefore its description and integration in a systematic set of examples can be more easily established. The automatic feedback given to the learner can therefore be phonological as pertaining to the system of contrasts existing between the 4 tones of the language, and phonetic in commenting the particular details of realization in term of syllable length and pitch contour.

Examples in languages where the phonological functions of intonation is not so well established such as English or French require the author to put sets of model sentences together to make sure the phonological functions are clearly explained to the learner and properly shown in the progression of examples and presentation of data. Although some attempts have been made along these lines, the lack of consensus among researchers in the theory of sentence prosody constitutes a serious problem. In our authoring effort for instance, only perceptually prominent syllables of the sentence defining stress groups should be imitated by the learner in their characteristics of duration and pitch movement.

5. Implementing principles and observations

The new version of WinPitch LTL attempts to implement these principles: it gives the learner the opportunity to listen to models at a reduced, programmable speech rate to enhance comprehension, to see the prosodic parameters on screen either in real time or at reduced speed, to listen to his/her performance and to learn by doing by modifying the prosodic parameters of his/her own voice graphically according to the prosodic features to acquire. Phonologically pertinent speech segments are displayed highlighted in a colour chosen by the author-teacher, responsible for putting model sentences together.

5.1 Navigation

WinPitch navigation toolbar allows an easy selection of sentences in a lesson, to listen to the model at variable speech rate (programmable in the 15% to 200 % range), to replay learner imitations in any order, to listen and repeat in sequence all the models of a lesson, and to graphically enter prosodic modifications of any learner repetition.

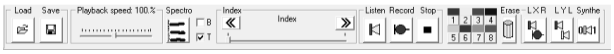

Fig. 1: WinPitch LTL toolbar

A zoom function displays the model on the whole screen with or without the corresponding spectrographic display of both model and sentence.

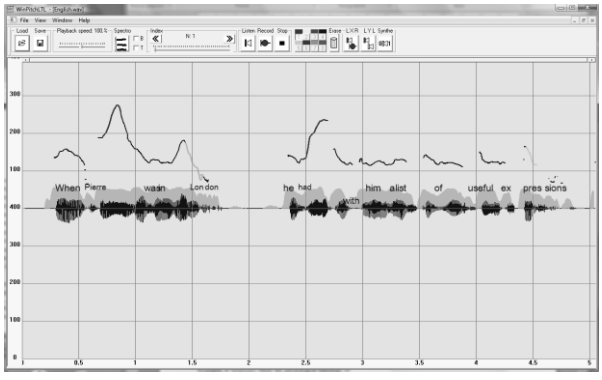

Fig. 2: Screen layout

The screen is divided in a top section reserved for the model, the aligned text, the model highlighted pitch and intensity curves (a spectrogram can also be displayed), and a bottom section allocated to learner imitation**s.**

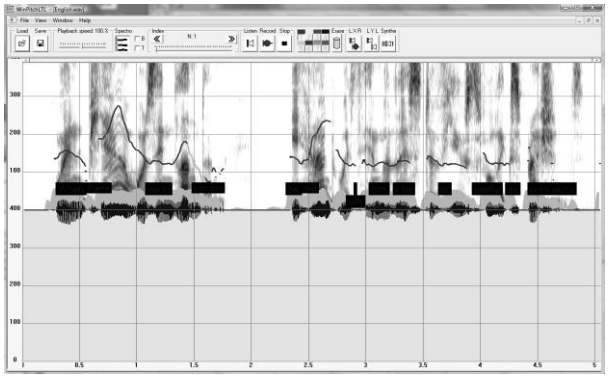

Fig. 3: Spectrographic display

5.2 Slow rate playback

To ensure a better perception of the model by the learner and allow the visual and auditory coordination between a screen cursor and speech played back at a slower rate, the model and imitation playback rate can be adjusted continuously between 15% and 200%. Slow playback is performed by a PSOLA (Moulines and Charpentier, 1990) engine, based on the pitch synchronous insertion or deletion of pitch period. Thanks to the use of a reliable pitch tracking algorithm (the spectral comb method), the slowed speech is usually of excellent quality. Conversely, the model can be played back at a higher speed to test the learner comprehension in various conditions.

5.3 Speech segment highlighting

Speech segments can be highlighted by the author in any colour, giving the possibility to the author to define the linguistically or otherwise pertinent sections of the model to be imitated by the learner.

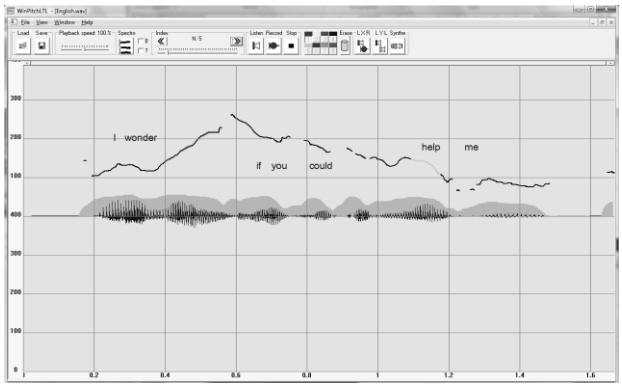

Fig. 4: Stressed syllables highlighting, the pertinent stressed segments are highlighted in color.

In the example of figure 2, stressed syllables of the sentence *I wonder if you could help me* are highlighted in red, showing the contrast of pitch slope rising falling typical in English in sentences with two stress groups.

WinPitch LTL is Unicode compliant, which means text in any font available in the Unicode standard can be entered in the notepad or on screen aligned with the pitch curve as shown in figure 1. Text can be entered directly using an appropriate keyboard driver available in Windows based systems, or by clicking on cells of a characters table (see figure 5), a special set can also eventually be defined by the user).

5.4 Prosodic morphing

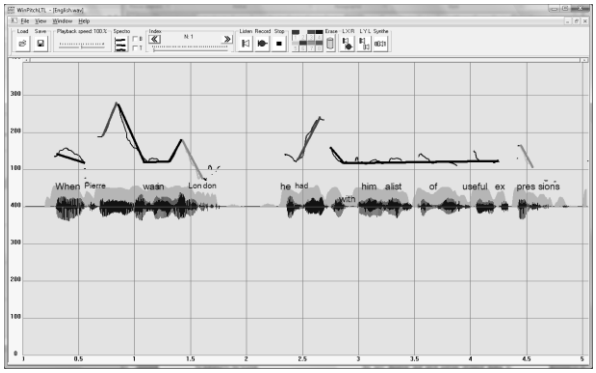

Fig. 5: Prosodic morphing

All four prosodic parameters (fundamental frequency, intensity, syllable duration and pauses) can be modified on either the model or the learner sentence through simple graphic commands. Using the spectrographic display, the user can easily change specific syllable durations or pitch movement according to the model presented to the left of the screen. Propositioned duration (in red) and pitch (in white) lines are placed on the learner part of the screen. These lines can be dragged, cut, and its vertices placed easily with the mouse on an appropriate spot on the screen in order to define the new duration and pitch pattern obtained during re-synthesis (activated by clicking on a single button).

5.5 Authoring program

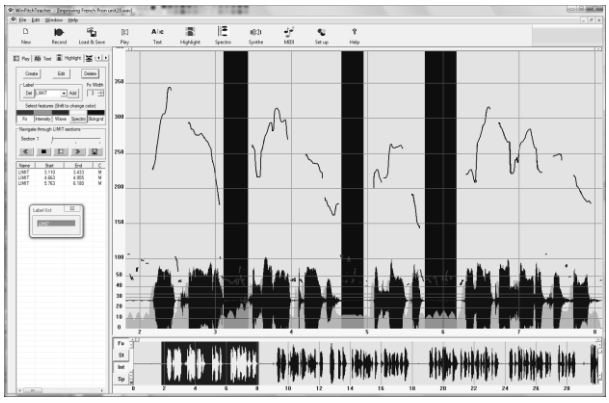

Fig. 6: Authoring program

A companion authoring program (WinPitch Teacher) gives total control of all functions to elaborate sets of model sentences. From a pre-recorded file (in wav or mp3 format) or after a direct recording of model sentences, the author can with simple commands insert the segments reserved after each model for the learner production, add text on the notepad and on screen, pre-place the duration and pitch morphing lines, highlight pertinent segments with any colour, define segments

for automatic mapping on the learner imitation, add comment in HTML format, etc.

6. Conclusion

WinPitch LTL is an innovative software program derived from older versions developed in the last 8 years (Germain and Martin, 2000). To the learner, it offers a user friendly interface, allowing easy navigation in the sets of model sentences included in a lesson. The learner can listen to the models proposed at variable speech rate to obtain a convincing correlation between visual and auditory perception of pitch movements and syllable durations related to pertinent segment highlighted in colour, and repeat and observe his/her own realization in terms of prosodic parameters displayed in real time. Model pertinent highlighted segments are automatically mapped on the imitation, and the learner can correct if necessary the corresponding pitch, intensity and duration parameters through simple graphic command driving re-synthesis of his/her own voice.

The teacher, thanks to the use of a complete set of authoring functions, can easily build lessons by defining graphically time spaces reserved to the learner, by adding text in any language (using Unicode fonts) on screen and in the notepad, write notes in HTML format.

A set of lessons have been developed for English and Mandrain, using the recordings og the well known method Assimil “L’anglais sans peine” and “Le chinois sans peine”.

References

Abberton, E., & Fourcin, A. 1975. Visual feedback and the acquisition of intonation. In E. H. Lenneberg, & E. Lenneberg (Eds.), *Foundations of language development* (2nd ed., 157-165). New York: Academic Press.

Adriaen M. 1983. 'A New Approach to the Teaching of French Intonation'. *Abstracts of the Tenth International Congress of Phonetics Sciences*, Dordecht, Holland, 731-735.

Anderson-Hsieh, J. 1994. Interpreting visual feedback on suprasegmentals in computer assisted pronunciation instruction. *The CALICO Journal*, *11*(4), 5-21.

de Bot, K. 1983. Visual feedback of intonation I: Effectiveness and induced practice behavior. *Language and Speech, 26*(4), 331-350.

Delmonte R., M.Petrea, C.Bacalu 1997, SLIM Prosodic Module for Learning Activities in a Foreign Language, *Proc. ESCA, Eurospeech97*, Rhodes, Vol.2, 669-672.

Germain, A. et Martin, Ph. 2000. Présentation d’un logiciel de visualisation pour l’apprentissage de l’oral en langue seconde », *www.alsic.org*, 3, No 1, 61-76.

James, E. 1976. The acquisition of prosodic features of speech using a speech visualizer. *International Review of Applied Linguistics 14*, 227-243.

Lane, H., & Buiten, R. 1969. A self-instructional device for conditioning accurate prosody. In A. Valdman (Ed.), *Trends in language teaching*, 159-174. New York.

Léon, P., & Martin, Ph. 1972. Applied linguistics and the teaching of intonation. *Modern Language Journal, 56*, 139-44.

Lepetit, D. 1990. The F0 learner: A phonologically deaf. *The Bulletin of the Phonetic Society of Japan* **195**, 11-17.

Malfrère, F., Deroo, O., Dutoit, T. and Ris, C., 2003. Phonetic alignment : speech synthesis-bases vs. Viterbi-based, *Speech Communication*, 40, 503-515.

Moulines, E. & Charpentier, M., 1990. "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", *Speech Communication*, 9, 453-467.

O'Connor, J. D., and G. F. Arnold. 1973. *Intonation of Colloquial English*, London: Longman.

Spaai G.W.G. and Hermes D.J. 1993,A visual display for the teaching of intonation, *CALICO Journal* 10, 19-30.

Stibbard, R. 2000. Teaching English Intonation with a Visual Display of Fundamental Frequency, *http://iteslj.org/Articles/Stibbard-Intonation/*

Vardanian, R. M. 1964. Teaching English through oscilloscope displays. *Language Learning*, *3/4*, 109-118.

Weltens, B., & de Bot, K. 1984. Visual feedback of intonation II: Feedback delay and quality of feedback. *Language and Speech*, *27*(1), 79-88.

WinPitch LTL, 2009. http://www.winpitch.com