Design and development of computer assisted learning of prosody

Anne Bonneau†, Koray Balci‡, Yves Laprie† and Vincent Colotte†
† LORIA/CNRS and ‡ INRIA

E-mail: Anne.Bonneau@loria.fr, Koray.Balic@itc.it, Yves.Laprie@loria.fr, Vincent.Colotte@loria.fr

ABSTRACT

This paper describes a toolkit that provides teachers of foreign languages with tools intended to facilitate learning of foreign language prosody. This toolkit can be used simultaneously to give teachers interactive illustrations of the target language prosody with respect to that of the mother tongue, and to help learners being aware of the acoustic correlates expected. For that purpose we developed signal transformations (F0, duration and energy) based on TD-PSOLA and automatic alignment adapted to English uttered by French speakers. Considering that teachers should be given the possibility of customizing their courses, these tools have been developed in the form of ActiveX controls that teachers can easily incorporate in any MS Office document.

1. INTRODUCTION

Whereas visual feedback on supra-segmentals has proved its efficiency in the domain of language learning [1], speech visualization tools are not yet widely used in schools. The reasons, as A. Germain pointed out [2], probably come from the lack of communication between speech specialists and teachers of foreign languages, a lack of formation of these teachers, as well as the relative lack of accessibility of speech software. Furthermore, much of speech software intended to improve the oral production of a second language, gives very poor feedback on the errors made by learners and let them extrapolate their difficulties [3]. Winpitch [2] represents a new generation of speech software, more convivial, which includes speech transformation tools (modifications of prosodic parameters), and is intended to be used, at least during the first steps of learning, with the help of teachers. In this vein, we present a new software devoted to the improvement of the production of prosody in a second language. This toolkit includes a large set of functions, such as speech transformations, but also automatic alignment, based upon ASR. These functions have been ported in the form of Active X components, which can be inserted in any MS Office documents, and are accessible to all.

Indeed, we are working with a view of giving teachers of foreign languages knowledge about prosody of French compared against the prosody of the target language (i.e. English in our case) to allow them, in a second stage, to exploit these signal transforms tools to assist learners.

Thus our toolkit is made up of three parts:

- an interactive course (initiation) on prosody, devoted to teachers, based upon our set of speech tools,
- a database of references, i.e. sentences uttered by native speakers, that the speaker will repeat and which will serve to judge his/her realization,
- a set of speech tools, including speech signal edition, transformation, automatic alignment, now available from any MS Word or PowerPoint files.

We present the set of tools and its usefulness for language learning in chapter 2, and give a concrete example in chapter 3.

2. SPEECH TOOLS FOR LANGUAGE LEARNING.

The user has at his disposal a set of tools to edit, modify and label automatically speech signals [4], which are available from any MS Word or PowerPoint document. Each sound

ISBN 1-876346-48-5 © 2003 UAB

and each part of a sentence can be edited, played back, zoomed and modified. In this chapter we will show how these tools can improve language learning.

Signal edition tools. The functions include the edition of the signal waveform, the display of the spectrogram, the F0 contour and the intensity curve, as well as speech play back and labelling. Signal transformations have been developed using an improved version of TD-PSOLA. It exploits a pitch marking algorithm together with a dynamic resampling algorithm that allows high quality stimuli to be generated[7]. During learning (lessons and exercises), the user plays back and visualizes its own utterances, as well as those of the references. If spectrogram reading requires a phonetic initiation, intensity curves and melodic contours, superimposed onto the spectrogram, are accessible to all. Segmentation and labelling allow the user to make the right associations between the different visual representations and the speech units, and to determine the duration of speech segments.

The association of visual and auditory feedbacks makes the learner aware of the differences between the correct pronunciation and his/her own, which improves his perception of the second language. Then the user can try to bring his realization closer to the target and to improve his pronunciation.

It is important to note that, if visual feedback gives objective measures to compare the learner's utterances with that of the references, the interpretation of these measures is far from trivial. The help of the teacher is necessary to decide what is pertinent or not and to guide the beginner in his/her initiation. This raises the issue of developing a simplified model of the target prosody that could give some information about the expected acoustic. Even if it is hard to imagine a general model it should be possible within the framework of single words or simple sentences. This is an objective for our future work.

Signal transformation tools. During exercises, the teacher shows the student what he expects by modifying his/her own voice. The user can modify prosodic cues such as duration, fundamental frequency and intensity (independently or at the same time) at each instant of the signal. These modifications can be designed by hand in a very simple way: the user monitors curves with the mouse directly on the spectrogram. After this operation, the signal is resynthesized and can be saved. Then it is possible to correct only the prosodic parameters that have been wrongly realized, on a segment where the error is the most interesting. As examples, the teacher can modify the intensity and/or the duration of an accentuated vowel, or change the final melodic contour of a question. Since other acoustic cues (segmental or supra-segmental), are not affected by the manipulation, the student can compare his/her original realization with the locally corrected one, and better apprehend what drives it apart from the target. Another way of making the learner aware of his/her error is to exaggerate it.

Among other useful modification functions, the user can cut, copy and filter a segment and listen to the modified signal.

Automatic alignment. Annotating, which can be done in sounds and in words under WinSnoori, localizes the segments on their different visual representations. The user can annotate the signal or visualize a annotation file. When this file doesn't exist, and this is generally the case for student's realizations, the sentence can be annotated by automatic alignment software our team has developed. First, some probable phonetic transcriptions are generated from the orthographic transcription, then the software try to determine the exact emplacement of each sound in the signal. This last task is performed with an automatic speech recognition system (ASR), and is facilitated by the knowledge of the segments to be found and of their phonetic context. Since ASR systems are trained with native speakers, we had to adapt our system to French learners speaking English. We are presently collecting the Timit corpus uttered by young French speakers in one secondary school and two universities of Nancy for that purpose.

Indeed, it is of great importance to have at one's disposal a good alignment since a precise detection of segments is necessary to compare the learner and the reference sentences and to modify speech prosody. This alignment cannot be realized by using native American-english speakers' acoustic models because phones uttered by French speakers may deviate substantially from those expected. The English phoneme /b/, for instance, is closer to the French /p/ than the French /b/ that is voiced along the whole closure, and French speakers almost systematically produce nasal vowels that do not exist in English. Therefore, the alignment procedure must take these specificities into account. Note that this alignment problem become much more difficult when the mother tongue of learners is unknown.

Snorri Active X components. Software must be user friendly and simple enough so that teachers and students can adapt themselves to it easily. For that purpose we have ported the main functions of our speech analysis software in the form of ActiveX controls that can be easily inserted in any MS Office document (such as Word or PowerPoint), or tutoring tools such as Authorware, as well as web pages. In addition to the labelling tools, notes linked to the signal or spectrogram can be used to focus the attention of the reader and explain some salient phenomenon about various domains (acoustics, prosody, coarticulation effects...). Then it is very simple to prepare speech tutorials for students. Snoori ActiveX Components will be available from our website at http://www.loria.fr/equipes/parole

3. EXAMPLE

In France, teaching of English prosody for non-specialists insists on the place and the strength of English lexical accent as well as on the main intonation patterns. One of its goals is also to make learners aware of French prosody. In this chapter, we show how our toolkit improves prosody learning, taking the acquisition of the English lexical accent as an example.

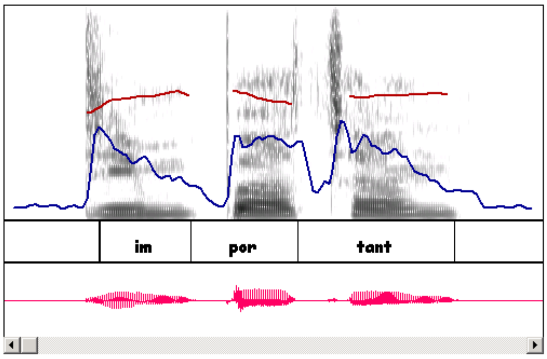

Figure 1: Word “important” uttered by a French speaker

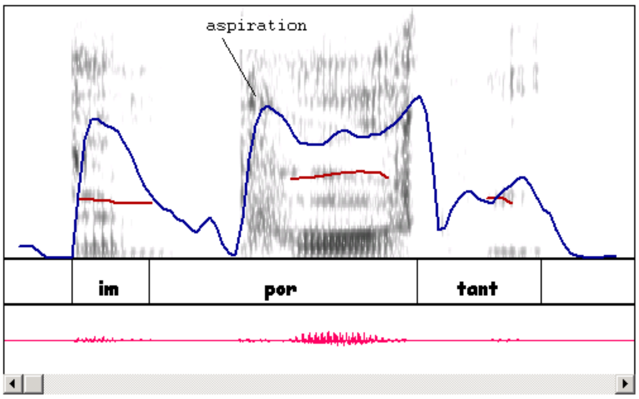

Figure 2: Word "importance" uttered by a native English speaker. The F0 contour (intonation) is in red, the intensity curve in blue.

English lexical accent is very different from the French one, first because its place is free, whereas the French one is fixed, but also because it is very well marked on an acoustical point of view. Indeed, the stressed syllable is more intense, higher and longer than the unstressed ones. Furthermore, unstressed syllables are sometimes reduced: its vocalic timbre comes close to that of a neutral vowel. The French accent is essentially characterized by a lengthening of the last syllable of a group of words.

In order to show how our toolkit put in light mispronunciations, we will consider the example of the word « important », uttered by a 12 years old French girl, with two years of English lessons. The acoustical characteristics of the stressed syllable “por”, uttered by an English speaker, its high pitch, long duration and great intensity, are very well marked and visible (Fig. 2). We also note that the stop /p/ is aspirated, -there is a weak noise just before the vowel /o/. If the user select (by a simple mouse click) the last syllable “tant” and listen to it separately many times, he could realize that this syllable is just a nasal murmur. The French student is not aware of this reduction and pronounces the syllable with its full timbre (Fig. 1). The French learner realisation is characterized by an attenuation of the English accent (the syllable “por” is less long, less high and less intense, with respect to other syllables, than in the reference) and the persistence of a French accent (long duration) on the last syllable (the stressed syllable in French). We have corrected the pitch and the durations on the learner’s realisation (Fig.3).

ISBN 1-876346-48-5 © 2003 UAB

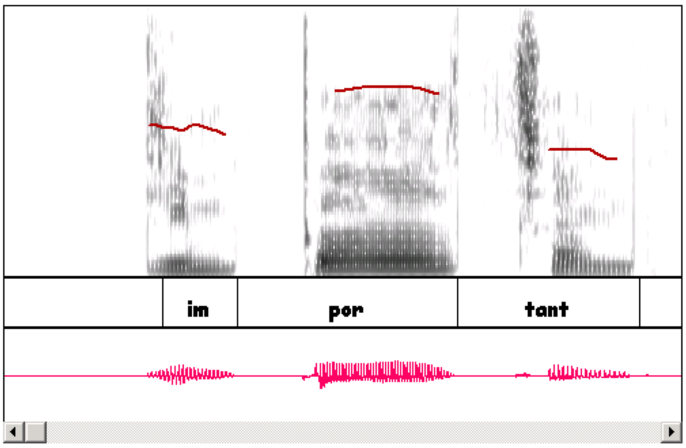

Figure 3: Word "important" after corrections on energy, duration and F0

4. PERSPECTIVES

Although these kinds of modification can be implemented automatically to a certain extent (see [5] for instance), a crucial issue is how to make learners aware of the deviations with respect to the expected acoustic correlates. This issue is all the more difficult since signal transformations presented before must be used carefully to prevent any acoustic and/or prosodic artefact. The work of Makarova [6] can be used to predict the strength of modifications to guarantee that learners perceive exaggerated correct or incorrect acoustic correlates in the case of Russian and Japanese. We have the project of conducting the same kind of perception experiment in the case of French and English.

We are convinced that the possibility of evaluating the impact of his own modifications allows the learner to have a better idea of the foreign language prosody he is learning. However, this requires an efficient interaction with a tutor, human or virtual. A virtual tutor should rely on a set of two prosody models for both mother tongue and foreign language, together with a comparison tool focusing on critical differences. The difficulty is to build models that are sufficiently general to cover most of the sentences encountered by learners.

REFERENCES

[1] J. Anderson-Hsieh, “Interpreting visual feedback on suprasegmentals in computer assister pronunciation instruction”, *CALICO,*Vol 11.4, 1994.

[2] A. Germain and P. Martin, “Presentation d’un logiciel de visualisation pour l’apprentissage de l’oral d’une langue seconde*”* , *ALSIC*, Vol. 3. 1., 2000

[3] D.M. Chun, “Signal analysis software for teaching discourse intonation” *Language Learning and Technology*, Vol. 2, 1. 1998.

[4] Y. Laprie “Snorri, a software for speech sciences”, *MATISSE,* 1999.

[5] Y. Meron and H. Keikichi, “Language training system utilizing speech modification”, Proceedings of the fourth International Conference on Spoken Language Processing, pp 1449-1452, Philadelphia, 1996

[6] V. Makarova, “Perceptual correlates of sentence-type intonation in Russian and Japanese”, Journal of Phonetics, Vol. 29.2, 2001.

[7] V. Colotte and Y Laprie, “Higher pitch marking precision for TD-PSOLA”, Proceedings of XI European Signal Processing Conference (EUSIPCO), Toulouse, 2002