TOOLS DEVOTED TO THE ACQUISITION OF THE PROSODY OF A FOREIGN LANGUAGE

Guillaume Henry, Anne Bonneau and Vincent Colotte

LORIA, Speech Group
{guillaume.henry, anne.bonneau, vincent.colotte}@loria.fr

ABSTRACT

The work presented here is developed within a project devoted to the acquisition of English prosody by French learners. Our goal is to improve both production and perception. To that purpose, we develop speech signal transformations (auditory correction) and propose a real diagnosis of the learner's production, exploiting knowledge about L1 and L2 prosody as well as acoustical analyses. We present our strategy and a simple example.

Keywords: CALL, prosody, signal transformations

1. INTRODUCTION

This work aims at improving the production and the perception of English prosody by French learners, thanks to speech signal transformations and knowledge about the prosody of the mother language (L1) and the target language (L2).

Since the early nineties, several works devoted to Computer Assisted Language Learning (CALL) have resulted in the development of pronunciation training systems. Most of these systems, as Better Accent [13], offer visual feedbacks (melodic patterns especially) about the realizations of the learner and that of a reference native speaker. A real-time visualization of the melodic pattern is provided by Winpitch LTL II [16]. This software, devoted to language teachers, enables the user to modify prosodic cues (fundamental frequency, intensity, segment durations, and pauses) and annotate prosodic displays. The Prosodic Module of SLIM (Multimedia Interactive Linguistic Software) [8], which aims at improving the realization of the English lexical accent by Italian speakers, offers an automatic diagnosis of the learner's mastering of duration cues (in particular the correct lengthening of the stressed syllable). Feedbacks about duration cues are also given by the Virtual Language Tutor (VTL) [11], a 3D talking head focusing on learners' articulation.

Whereas Winpitch is intended to language teachers, SLIM and VTL carry out automatic diagnoses. This work aims at offering an evaluation of the learner's realization based upon F0 contour, phone duration as well as knowledge about the prosody of L1 and L2. Thanks to the comparison between the learner's realization and that of a reference speaker, made possible by automatic alignment tools, we propose a real analysis of the learner's production. This allow us to deliver relevant feedbacks, including a diagnosis, visual feedbacks as well as an auditory correction intended to make learners aware of the prosody of the foreign language.

We will present the tools developed for the project in the second section, as well as, in the third section, a simple example.

2. TOOLS

Acoustic and prosodic analyses are performed with Winsnoori [15], a software devoted to visualization, analysis and processing of speech signals. The user can annotate speech signals phonetically and orthographically, edit F0 contours, intensity curves, and calculate phones and syllables duration (if the speech signal is labelled). The visualization of prosodic cues, represented onto the spectrogram, gives important information to users. Indeed, the effectiveness of visual feedbacks in the domain of CALL has not to be proved anymore [5].

Signal modification functions have been included in Winsnoori. These functions are based on an improved version of TD-PSOLA method [6] and allow users to manually modify contours, speech rates as well as syllable durations. Then the modified signals are resynthesized, and the users can save the modifications. If the modification consists of imitating the prosodic cues of a model, learners can appreciate the differences between their realization and what they are expected to realize. In this work, an automatic version of these modifications has been developed.

2.1. Corpus

A corpus, associated to a set of exercises conceived especially for the acquisition of English prosody, was created with the help of English teachers from the University of Nancy 2 and secondary school teachers. This corpus is composed of transparent isolated words (for example: difference, ministry...),

a set of sentences (a few hundreds), and small texts (a few tens). The corpus focuses on specific points of prosody acquisition (for example the acquisition of lexical accent, focus accent, or the intonation patterns of a declarative sentence or a question). It was recorded by two English teachers (a male and a female), who are native speakers of English and by a French learner (a female non native speaker).

2.2. Automatic Alignment

Prosodic cues generally appear on well determined linguistic and phonetic entities. So a preliminary segmentation into words, phones and syllables is necessary to localize the prosodic events and to compare the learner's realization with that of a reference. Our approach is described in the next paragraph.

After users utter a linguistic entity (a word, a group of words, or a sentence) from the corpus, a segmentation of their realization is performed. First, a phonetization of the text is carried out using the CMU dictionary. Then, the segmentation is computed with a text-to-speech alignment, which establishes the correspondence between phonetic units and parts of the speech signal. Text-to-speech alignment is achieved using Hidden Markov Models [10]. Two different kinds of model have to be used: one for native speakers and another one for non-native speakers (learners). Indeed, learners of a foreign language tend to replace the sounds they do not know by sounds from their mother language. That's why models used for native speakers (learned on the TIMIT database and developed for ASR purposes) should be adapted to non-native speakers [4]. Nevertheless the modelling of non-native speech is still under development in laboratories. So we decided to keep the native learner models for the moment. Finally the syllabification program of NIST was applied to the CMU dictionary in order to obtain a database of syllabified words.

2.3. Evaluation of the learner's production

First, let's say that the native speaker's and the learner's realizations are automatically displayed with their segmentation into phones and syllables. Visual displays are shown on the spectrogram of the learner's realization. For example, the red curve in Fig.1 represents the F0 contour. Syllables and phones relative durations of the reference and that of the learner are shown on the learner's realization.

We wish to provide an evaluation of the learner's production, based upon an acoustical analysis of his/her realization. Note that this approach deals with the realization of established prosodic categories. This analysis relies upon a comparison with a reference (a native speaker realization for the moment), which implies the localization of the prosodic entity under consideration, on the reference realization. This localization can be either automatic for simple prosodic phenomena appearing in short sentences (such as the acoustic manifestation of the lexical accent in isolated words) or provided by the prosodic annotation of a database or even by a prosodic model. For our first application (production of the lexical accent), we use an automatic detection of the stressed syllable. Once the prosodic event is localized on the reference's realization, the learner's production is analysed and an evaluation of his/her production is provided. This evaluation is not easy to perform and can vary from a simple distance from the target, which is not a satisfactory solution, to a more precise judgment based upon perceptual experiments. When possible, we choose to rely on the results from perceptual experiments. The evaluation is provided both in the form of a short text and visual displays such as arrows (see Fig.1). We will show in section 3 an example of our method in a simple case (realization of the lexical accent in isolated words) as well as the exploitation of knowledge of L1 and L2 prosody.

The reduction phenomenon (especially when a syllable is not pronounced) requires a specific treatment. If learners do not reduced unstressed syllables, this absence of reduction cannot be considered as a mistake. But in this case, the comparison is no longer possible and the learner is invited to repeat his or her realization and to reduce the appropriate syllable. In addition, this process make him or her aware of this phenomenon.

2.4. Automatic auditory correction

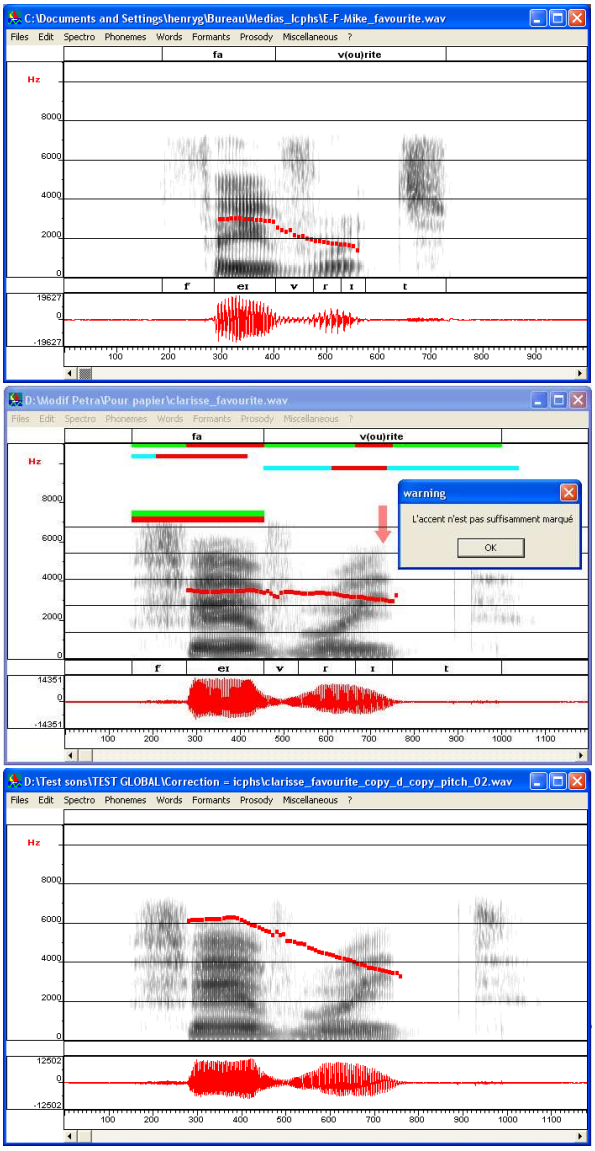
Speech signals modification can be done manually by users as shown in [3]. In this paper, an automatic auditory feedback is carried out. We proposed two correction solutions to make learners aware of their deviations.

In the first correction, the prosodic cues of the learner's realization are directly replaced by the prosodic cues of the model while keeping the timbre of the learner's voice. In a first time, the relative durations of the learner's phones are aligned with that of the reference. In a second time, a new F0 contour for the learner's utterance is computed using a linear interpolation of the model's normalized F0 contour. Then the learner's realization is resynthesized and learners can appreciate the resulting speech signal. An example of such a correction is given on Fig.1 (third spectrogram).

The second possible correction consists of enhancing the deviations of the learner. As an exam-

ple, the French lexical accent is essentially correlated to a lengthening of the last syllable of the word. Thus French learners will tend to keep this lengthening to English realizations even on unstressed syllables. Thus it is possible to lengthen this last syllable and to let learners listen to the exaggerated version as well as the reference in order to make them aware of the expected realization.

Figure 1: Comparison of the word "favourite"



3. EXAMPLE : evaluation of the lexical accent production by a French speaker

We take the example of the word "favourite" (British English) uttered in isolation. The native speaker's realization and learner's realization are presented in Fig. 1. The evaluation is based upon the acoustic cues and exploits knowledge of the prosody of L1 and L2.

3.1. Using the knowledge of the prosody of L1 and L2

It is commonly said that learners of a foreign language are "deaf" to the prosodic system they are studying. So the first language highly influences realizations in the second language [2]. This influence is particularly perceptible for French people learning English because these two languages do not belong to the same prosodic category [1, 14]. Indeed French is considered as "syllable-timed" and English as "stress-timed" [7]. The evaluation we provide is based on the knowledge of the specific problems encountered by French speakers learning English prosody. The first problem comes from the place of the lexical accent: fixed in French and free in English [17]. In addition, the English lexical accent is strongly marked on an acoustical point of view whereas the French one is relatively weak [9]. In fact the French accent just consists of a lengthening of the last syllable of the word (or the group of words). English lexical accent is characterized by a pitch modification, an increase of intensity and a lengthening of the vocalic nucleus of the stressed syllable. In addition English unstressed syllables are frequently reduced whereas French syllables keep their original and clear timbre.

As we previously said, French learners will tend to use prosodic features of their mother language instead of the prosodic features of the target language. For example, they may lengthen the last syllable of a word, even when this syllable is unstressed in English. A particular attention will then be put on duration cues all over the word.

3.2. Analysis

The adjective "favourite" is analysed. In that particular case, the second vowel ("ə") is reduced by the reference (cf. audio file 1) and by the learner (cf. audio file 2).

The automatic alignment in phones and in syllables is shown respectively at the bottom and at the top of the interface. The phonetic transcription in IPA of "favourite" is the following: ['f eɪ v (ə) r ɪ t].

For isolated words, we relied on the fundamental frequency peak to identify the place of the lexical accent in the word. The F0 pattern is analysed using a linear stylisation, which seems sufficient for this first application [12]. Besides F0 values are converted to a semi-tones scale using the relation (1) defined by Bagshaw.

$$F0_{Semi-Tones} = 12.log_2(\frac{F0_{Hz}}{55}) \quad (1)$$

In a first time, a syllable is considered as stressed if the F0-peak belongs to this syllable. This parameter (F0) is efficient for words uttered in isolation.

Table 1: *F0 statistics of the word "favourite".*

	Learner	Native Speaker
Mean	25,95	22,27
SD	**0,6**	**4,89**
Min.	24,67	14,49
Max.	26,81	27,27

In this example, the F0-peak of the French speaker is on the right syllable. Nevertheless the F0-standard deviation for the learner is less than 1 semi-tone whereas it is more than 4 semi-tones for the native speaker (cf. Table 1). In addition, a significant difference of pitch can be observed between the first and the second syllable on the reference whereas a very weak difference can be noticed on the learner's realization. In that case, the learner is informed that its realization is deviant (warning window, see second spectrogram in Fig.1). These statistics confirm the problems presented in the previous section: prosodic cues coding here the lexical accent are less salient on the learner's realizations. The F0 pattern of a French speaker pronouncing English utterances (or words) tend to be relatively flat. Generally speaking, this is a recurrent problem of learners of a foreign language, which is all the more marked for French learners who do not use pitch variations in the French lexical accent's realization.

In addition to the visualization of his or her realization, the learner is provided with a small text, summing up the evaluation that has been performed as well as the auditory correction (cf. audio file 3). An analysis of the syllables and phones duration is planed.

4. CONCLUSION

We have presented in this paper tools to help French learners of English prosody to improve both their production and perception. Learners are provided with an automatic evaluation as well as an auditory correction. These feedbacks rely on signal processing tools and knowledge of the prosody of L1 and L2. We will now focus on an evaluation of our method in Language Learning.

5. REFERENCES

[1] Abercrombie, D. 1967. *Elements of General Phonetics*. Chicago: Aldine.

[2] Bagshaw, P. C. 1994. Automatic prosodic analysis for computer-aided pronunciation teaching. PhD Thesis University of Edinburgh.

[3] Bonneau, A., Camus, M., Laprie, Y. and Colotte, V. 2004. A computer-assisted learning of English prosody for French Students. *Proc of InSTIL/ICALL* Venice.

[4] Bouselmi G., Fohr D., Illina I., Haton J.-P.2005 Fully Automated Non-Native Speech Recognition Using Confusion-Based Acoustic Model Integration *Proc of Interspeech* Lisboa.

[5] Bot De, K. 1991. Visual feedback on intonation I: Effectiveness and induced practice behaviour. *Language and Speech* 6 (4), 331–350.

[6] Colotte, V. and Laprie, Y. 2002. Higher pitch marking precision for TD-PSOLA. *Proc of European Signal Processing Conference (EUSIPCO)* Toulouse.

[7] Dauer, R. M. 1983 Stress-timing and syllable timing reanalysed. *Journal of Phonetics* 11, 51–62.

[8] Delmonte, R. 2002. A prosodic module for self-learning activities *Speech Prosody* Aix-en-Provence.

[9] Faure, G. 1962. *Recherches sur les Caractères et le Rôle des Eléments Musicaux dans la Prononciation Anglaise*. Paris: Didier.

[10] Fohr, D., Mari, J. F. and Haton, J. P. 1996. Utilisation de modèles de Markov pour l'étiquetage automatique et la reconnaissance de BREF80. *JEP* Avignon.

[11] Granström, B. 2005 Speech Technology for Language Training and e-Inclusion. *Proc of Interspeech* Lisboa.

[12] Hart t', J. 1952. F0 stylisation in speech: straight lines versus parabolas. *J. Acoust. Soc. Am.* 6, 3368–3370.

[13] Kommissarchik, J. and Kommissarchik, E. 2000. BetterAccent Tutor- Analysis and Visualization of Speech Prosody. *Proc of InSTIL/ICALL* Dundee.

[14] Ladefoged, P. 1975. *A Course in Phonetics*. New York: Harcourt Brace Jovanovich.

[15] Laprie, Y. 1999. Snoori, a software for speech sciences MATISSE.

[16] Martin, P. 2004. WinPitch LTL II, a Multimodal Pronunciation Software. *Proc InSTIL/ICALL* Venice.

[17] Vaissière, J. 2002. Cross-linguistic prosodic transcription: French versus English. *In honour of the 70th anniversary of Prof. L.V. Bondarko,. N. B. Volslkaya, N. D. Svetozarova and P. A. Skrelin* St.-Petersburg, 147–164.