Automatic Feedback for L2 Prosody Learning

Anne Bonneau, Vincent Colotte

Chapter Number

Automatic Feedback for L2 Prosody Learning

Anne Bonneau and Vincent Colotte
LORIA/CNRS, LORIA/UHP (Université Henri Poincaré), France

1. Introduction

The emergence of speech signal processing functions has allowed speech scientists to analyse and modify speech with the aim of improving the perception and/or the production of speech in adverse conditions and language learning. In most cases, these tools are all the more efficient that they take into account the phonological system of each language.
The applications of these aids are numerous and include the improvement of speech intelligibility in noise (Hazan & Simpson, 1998), and for hearing aids (Loizou, 1998), computer assisted-aids for language learning devoted to speech therapists or learners of a foreign language. They are concerned either with speech intelligibility (comprehension), or with the perception and production of speech sounds and prosody. One of the best known works in this domain, the study by Tallal (Tallal et al., 1996), published in Science, proposed speech modifications for hearing impaired children.
Among signal modifications, the enhancement and slowing down of specific regions of speech signals has been the object of numerous studies and applications. Ortega & Hazan (1999) applied these techniques to the improvement of speech intelligibility in second language learning. Colotte et al. (2001) also tackle speech intelligibility in second language learning. By means of an entirely automatic method, they enhanced unvoiced stops and fricatives and slowed down transitions.
This paper deals with the elaboration of advanced feedback devoted to aid learners in the acquisition of the prosody of a foreign language, and presents a pilot experiment investigating the immediate impact of such feedback on learners. Note that, from now on, we will use L2 for second (non-native) language and L1 for the learners' first (native) language. A computer-assisted aid in language learning (a CALL system) and more precisely in prosody can offer at least three kinds of feedback: (1) visual feedback, such as visual displays of the learners' melodic curves, often associated with those of reference speakers (2) automatic diagnoses, based upon acoustical analyses of learners' realisations and (3) "advanced" perceptual feedback, through speech manipulations.
Visualisation of melodic curves has been proposed since the early 60s, not only in second language learning but also in other domains such as hearing deficiencies. Vardanian (1964) was one of the first scientists to use melodic curve visualisation in second language learning and test its impact on learners (Brasilian students learning English). Her results, may be due to the poor quality of visualisation (oscilloscope displays) were far from convincing. Nevertheless, after a first period of scepticism, speech specialists, like James (1977), who used Philippe Martin's melodic curve detection (Germain-Rutherford & Martin, 2000), or De Bot (1983), agreed on the efficacy of visual patterns. Current commercial computer-assisted aid in

language learning systems, such as "Tell me more" of Auralog, LangMaster or Better Accent (Kommissarchik & Kommissarchik, 2004) propose visual display of learners' melodic curves.
The simple visualisation of melodic curves, although interesting, is not sufficient if one wants to provide learners with clear feedback about their production. Indeed, as Chun (1998) noted, if no further feedback is provided, the learners have to "extrapolate" their deviations themselves. A more efficient aid would provide learners with indications about their deviations and the way to improve their realisations. But the elaboration of automatic diagnoses encounters at least two major problems. Firstly, an automatic diagnosis of a learner's intonation presupposes the segmentation and labelling of the signal in syllables and in speech sounds. Yet this automatic segmentation, made by alignment if the text is known, is risky for non-native speech. (See 2.1.). A second major problem stems from the notion of "error" or "deviation". Indeed, we face issues such as "what are the links between acoustic cues and categories?" and "what are the accepted deviations?" For the moment the way to deal with this problem consists in proposing an evaluation of the degree of difference between prosodic patterns produced by a reference (or references) and those of the learner. Impact of feedback (diagnosis) on English intonation for French advanced learners (students in English at university) has been elaborated and tested by Herry and Hirst (Herry & Hirst, 2010). Learners did not improve their realisations due to diagnosis. The authors underlined that students participating in the test were volunteers, which might have influence results.
Another interesting way of helping learners in language learning relies upon speech modifications. Winpitch LPL (Martin, 2004), a speech signal editor, proposes functions especially designed for L2 teachers and learners, which enable the user to modify by hand fundamental frequency[1] and duration and annotate prosodic displays. Manipulations of prosodic cues are intended to make learners aware of prosodic patterns that do not exist in their first language; they are in general realized through PSOLA (Pitch Synchronous Over-Lap and Add) resynthesis. An interesting exploitation of speech manipulation consists in replacing the learner's prosodic cues by those of a reference without modifying the learner's timbre. WinSnoori (Laprie, 1999), software for speech analysis, enables users to realize this substitution by hand. An automatic version of this substitution has been realized (Henry et al., 2007) and implemented in WinSnoori. Other speech transformations based on an accent morphing technique have been recently proposed in the domain of foreign language intelligibility (Ingram et al., 2009; Yanagisawa & Huckvale, 2007). In particular, the last authors modified the accent of the speaker whilst maintaining his identity to improve the intelligibility of foreign-accented speech.
In this paper, we will present the various kinds of feedback (visual displays, diagnosis, and perceptual feedback based upon prosodic cues substitution)[2] we elaborate for L2 prosody learning (section 2), and a pilot experiment devoted to test their immediate impact (section 3). Results are given in section 4.

2. The L2 prosody learning platform

A platform designed for manual and automatic feedback of learners' prosody has been implemented in a research version of WinSnoori software. The platform contains classical

[1] The variations of the fundamental frequency (F0) generate the melodic curve.
[2] Parts of these tools have been presented in (Henry et al. 2007)

functions such as spectrographic displays, speech recording, real-time F0 (fundamental frequency) display, and integrates three particularly interesting functions with respect to language learning: an automatic text-to-speech alignment for native and non-native speech, a module for modifying the speech signal manually (the fundamental frequency and duration can be modified at the same time), and another module displaying automatic feedback on learners' (non-native) productions and exploiting most of the above-mentioned functions.
Besides classical F0 displays, two kinds of feedback are provided to learners, each of them based upon a comparison between a reference and the learner's production. The first feedback, a diagnosis, provided both in the form of a short text and visual displays such as arrows, comes from an acoustic evaluation of the learner's realisation; it deals with two prosodic cues: the melodic curve, and phoneme duration. The second feedback is perceptual and consists in a replacement of the learner's prosodic cues (duration and F0) by those of the reference. We will first describe the phonetic alignment, necessary for automatic diagnoses and speech modifications, then the method developed to modify speech signals, and, finally, perceptual feedback, diagnosis and the modus operandi.

2.1 Automatic text-to-speech 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. After users produce 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 (Fohr et al., 1996).
Two different kinds of model have to be used: one for native speakers and another one for non-native speakers (learners). Indeed, since learners of a foreign language tend to replace the sounds they do not know by sounds of their first language, models used for native speakers (learned on the TIMIT database and developed for automatic speech recognition - ASR- purposes) should be adapted to non-native speakers. Although we have already at our disposal an ASR system designed for non-native speech recognition (Bouselmi et al., 2005) we are still working on non-native alignment. Indeed, if ASR systems do not need precise segment boundaries for typical applications such as automatic speech transcription or translation, this is not the case for language learning applications where the production of a subject is analysed and corrected. The need for precise boundaries is obvious for diagnosis about segment durations, but is also important for speech modification functions such as the ones we design. Since the detection of very precise segment boundaries is not always possible, we need to associate confidence levels with detections. Then feedback can be avoided in the cases where detections are not sufficiently reliable. We are presently working on these two aspects of ASR for non-native speech: the improvement of segment boundary precision and the elaboration of confidence levels. The syllabification program of NIST was applied to the CMU dictionary in order to obtain a database of syllabified words.

2.2 Speech signal modification algorithm

Signal modification functions have been included in the platform. These functions are based on an improved version of well known TD-PSOLA method (Colotte & Laprie, 2002;

Moulines & Charpentier, 1990) and allow users to manually modify F0 contours, speech rates as well as phoneme durations. It means that we can apply a global or local modification factor. For instance, we can slow down a particular part of the signal and speed up another one of the same signal.

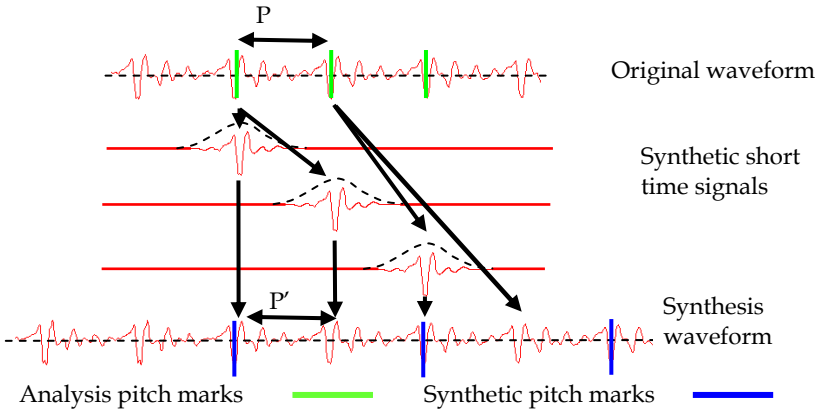

Fig. 1. Example of slowing down: the duration of the original signal is lengthened by a factor 2 (P′=P).

The modification method is based on the decomposition of the speech signal into overlapping pitch synchronous frames and the modification of pitch or duration is obtained by duplication/decimation of some frames.
Firstly, this method supposes the detection of the pitch marks: the signal periods are marked at the maximum or minimum relevant peaks (for the voiced parts[3]). These marks are spaced every pitch period and indicate the centre of the (analysis) frames (see fig.1 and 2).
Secondly, we need to compute the new position of these marked frames in the modified signal. The main requirement is to maintain the consistency of mark location between frames in order to preserve the original temporal structure of the signal under analysis. This marking directly influences the quality of the resulting signal. In (Colotte & Laprie, 2002), we have proposed a high precision algorithm for pitch marking at two levels: analysis and synthesis marks. In one hand, dynamic programming selects peaks in the signal for marking periods. Through correlation and pruning strategies, the algorithm overcomes errors which may appear with other algorithms. In addition, the algorithm is very fast in computation, which is very suitable for TD-PSOLA method. In the other hand, the combination of our pitch marking with a fast re-sampling method (to obtain the true synthesis frame) during the synthesis step increases the signal quality. This gain in accuracy avoids the reduction of quality between original and synthetic signal observed with the classical TD-PSOLA method: the level of noise between harmonics is reduced with our method.
Thirdly, each mark in the future signal is associated with a mark of the original signal. If we want to slow down (resp. speed up) the signal, the space between frames needs to be preserved (to keep the same pitch) and an original frame could be duplicated (resp. removed) to obtain the good length of the signal (see fig.1). If we want to modify the pitch, preserved, as for the slowing down, an original frame could be duplicated (or removed) to

[3] For unvoiced part, by definition without period, an arbitrary spacing is used (for instance 5ms).

obtain the good length with the good spacing (see fig.2). The pitch and duration modifications can be merged together.

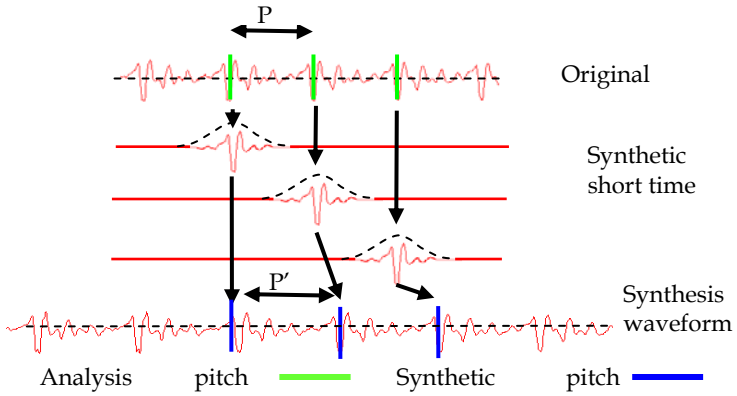

Fig. 2. Example of pitch modification (here a lengthening of the period/a lowering of the F0): the original pitch is lengthened by a factor 1.2 (P'=1.2P).

Finally, the signal is rebuilt from these frames (removed and/or duplicated) thanks to the principle of overlapping.
The advantages of this method are that we work in the temporal domain without any computation of transformation into the frequency space and the principle of using frames synchronized with the pitch does not destroy the coherence of the signal for each frame (notion of shape-invariance). We obtain a high quality of resynthesis without modification of the timbre of the voice. The method allows us to lengthen the pitch period until 3 times the initial period (i.e the F0 can be lowered by 3). For the duration, the slow down is only limited by the fact that a numerous duplication of the same part can create a new period (for unvoiced sound) and a noise of voicing can appear where there was no voicing (the factor > 4-5). But the slowdown is sufficiently local and generally small enough to avoid this drawback.
These modification functions can be use independently of the diagnosis process. The learner or teacher can (manually) modify the signal as he wants, to become aware of the link between prosodic cue variations and perception.

2.3 Automatic perceptual feedback

Such functions can be exploited to imitate the prosodic cues of a model, so that learners can appreciate the differences between their realization and what they are expected to realize. The signal is resynthesized with the required modifications, and the users can listen to the modified signals as well as visualize their spectrographic representations and the new melodic curves.
We have developed a module which realizes this imitation automatically. In a first step, the relative durations of the learner's phones are aligned with that of the reference. In a second step, 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 then he/she can appreciate the resulting speech signal.
Note that the used resynthesis algorithm keeps the timbre of the learner's voice, since only F0 and phoneme durations are modified. The copy of the F0 takes into account the mean F0 (pitch) of the speaker.

2.4 Diagnosis

The diagnosis is based upon a comparison between the realisation of the learner and that of the reference speaker, as well as phonetic knowledge about prosodic patterns of L1 and L2. Let us take the example of the realisation of English lexical stress in isolated words by French speakers (the object of the experiment presented in section 3). The syllable which should be stressed is assumed to be the one exhibiting the higher F0 in the reference's production. Thus the system evaluates, in semi-tones, the peak height of this syllable with respect to other syllables, in both realisations (the native and non-native ones) and returns a comment indicating whether the prominent F0 peak appears on the expected syllable of the non-native realisation and if it is sufficiently marked. At the same time, visual displays are shown on the spectrogram of the learners' realizations: arrows indicate whether the pitch of the target syllables should be raised or lowered (the colour of each arrow provides an indication about the degree of difference between native and non-native realisations with respect to F0 height); a red curve represents the F0 contour; the syllable and vowel durations of the reference and those of the learner appear in the form of bars on the top of the learner's realization; the length of bars varying with the duration of these segments (see Fig. 3). The reduction phenomenon requires a specific treatment. If learners do not reduce syllables which have been strongly reduced by the English reference, they are invited to repeat their realization and to reduce the appropriate syllable.

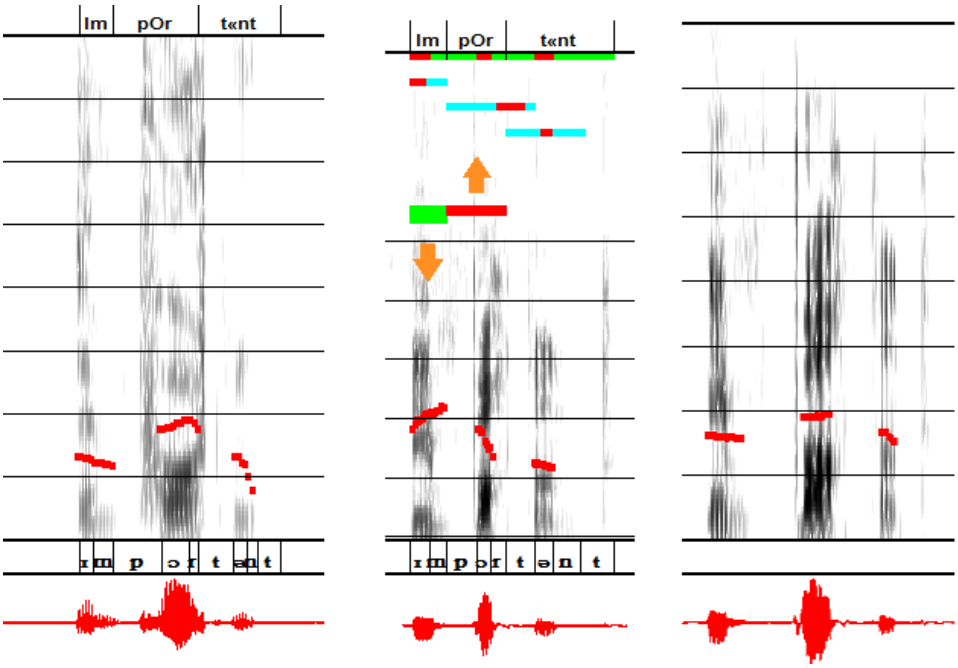

Fig. 3. Example of automatic diagnosis and speech modifications. Each panel shows the spectrogram, the melodic curve (in red) and the waveform (at the bottom, also in red) of a realization of the word "important". The first panel show the realization of this word by an English speaker; the second panel its realization by a French learner, with the diagnosis provided by the software (see text), and the third panel the modification of the learner's voice.

2.5 Modus operandi

Let us present the procedure proposed for providing automatic feedback to learners of a foreign language. Before clicking on the function in the menu, the subject should record a word or a small sentence, belonging to our database. He can, depending upon his wish or the task he is involved in, listen to the English reference before producing the item. When the subject selects the "correction" function, he is asked to choose a reference (one male and one female English speakers have recorded the corpus). Then the software realizes a text-to-speech automatic alignment of the native and non-native versions, and substitutes the prosodic cues of the native speakers to those of the non-native speakers.

Then the subject is provided with three versions of the utterance analysed, displayed in three windows, one per each version (see fig. 3). The first version consists in the native speaker's production, the second version is the learner's production, and the third version is the learner's modified production. Each window contains the representation of the melodic curve onto the spectrographic display, with the automatic alignment just under the spectrogram. The automatic diagnoses appear onto the spectrogram of the second window, that of the learner's original production. The subject can listen to the three versions and is invited to give a special attention to his own modified voice. The subject is free to produce the word again or select another one.

3. Experiment

We conducted a pilot experiment to analyse the immediate impact of diagnosis and advanced perceptual feedback on French subjects speaking English as a second language. In France, teaching of English prosody for nonspecialists focuses on the main intonation patterns as well as the place and strength of English lexical stress accent. We have chosen to test the production of English lexical accent in isolated words by French speakers. French and English are very different from a prosodic point of view since French is considered as a syllable-timed language and English as stress-timed language (Dauer, 1983). The place of the English lexical accent is free, whereas the French one is fixed, and the English accent is very well marked on an acoustical point of view whereas the French one is relatively weakly marked. Indeed, in English, the stressed syllable of isolated words is more intense, higher in pitch and longer than the unstressed ones. Furthermore, post-stressed syllables are often reduced: their vocalic nucleus tends to be very short, weak, and its vocalic timbre can become similar to that of a schwa (the schwa is, as an example, the first and last vowels in the word "America"). The French accent is essentially characterized by a lengthening of the last syllable of a word or a group of words; non accentuated vowels are always produced with their "full" timbre (there is no reduction).

It is commonly said that learners of a foreign language are "deaf" to its prosodic system. This explains why learners tend to use the prosodic features of their mother language instead of the prosodic features of the target language. For example, when speaking English, French learners generally lengthen the last syllable of a word, even when they know that this syllable is unstressed, and produce post-stressed vowels with their full timbre (without any reduction).

3.1 Experimental protocol

3.1.1 Corpus and subjects

The corpus, made up of 40 English transparent words, has been recorded by two English speakers, the "reference" speakers, one male and one female, born and educated in England,

and currently living in France. Transparent words have similar spelling in French and English but do not have the same pronunciation (e.g. "important", "favourite"). They constitute a good way of making learners aware of the differences between accentuation in both languages. The corpus was made up of two and three-syllable words of the type: `S1S2 (the accent is on the first syllable), S1`S2S3 (accent on the second syllable) and `S1(S2)S3. Words with this last syllabic structure (e.g. "governor", "family"), are generally pronounced by English speakers with a very strong degree of reduction on the second vowel whereas French speakers pronounce this vowel with its full timbre. We kept ten words for a "familiarization" phase, which precedes all the recording sessions; these words were pronounced by speakers but not analysed. Thus thirty words –ten per each type- remained for the analysis. Note that, in this corpus, the lexical accent never appears on the last syllable of a word.

Ten French subjects, five male and five female speakers from 15 to 50 years old, participated in the experiment. Taking into account learners' pronunciation, we chose four (relatively) advanced speakers and six low (production) level speakers. The advanced speakers have lived for a short period in USA or made frequent trips in this country, whereas the six speakers with a low oral production level (let us call them beginners, for short) have studied English at school for at least five years but do not master English pronunciation. Since ten subjects produced thirty words in each of the two recording sessions (presented below), the corpus contains 600 words uttered by non-native speakers.

English reference speakers and French subjects recorded the corpus in a quiet office with a Sennheiser headset microphone connected to a laptop. The mono signal was digitized at 16 bits with a sampling frequency at 22500 Hz. The software Audacity was used for all the recording sessions except for the one requiring the use of the platform (second session of the test condition).

3.1.2 Experimental conditions

There were two experimental conditions: the test and control ones, and two sessions of recording for each condition. We selected five speakers for each condition, two advanced speakers and three "beginners", trying to balance both groups. Note that we do not let learners chose between the test or the control conditions.

The first session was the same for both conditions, subjects from both groups (the test and control groups) recorded the corpus without listening previously to English references. This gives us an indication of the learners' mastering of English lexical accent at the beginning of the experiment. Each word was written on a sheet of paper in orthographic transcription. To avoid the production of a list intonation, a short pause was enforced between each word.

In the second session, subjects participating in the test condition were submitted to the platform for prosody learning. The procedure was the following: subjects were invited to select a word, written in orthographic transcription, from a list corresponding to the corpus. For the purpose of this experiment we asked subjects to follow the list order, and select a male or a female reference according to their gender. Then subjects uttered the selected word without listening previously to its English reference; the system analysed the realisation, and visual, perceptual and textual (diagnosis) feedback was displayed on three windows as explained in section 2.5. After subjects took knowledge of the various kinds of feedback, they recorded the word one more time, and selected the following word to repeat the procedure until the end of the list. Only word repetitions (the pronunciation after feedback) have been analysed and compared to words pronounced during the first session.

In order to avoid false corrections due to erroneous segment boundary detection, the system stops just after the automatic alignment of the word under analysis to ask the experimenter whether he/she desires to continue the process or modify speech boundaries before. This is a necessary step if we want to test the impact of feedback without incurring the risk of disconcerting the learner by inappropriate corrections. Let us recall that we are currently working on the elaboration of confidence levels to limit the risk of erroneous corrections.
In the second session, subjects participating in the control condition read each word on a sheet of paper, listened to its reference, and uttered it.
Both conditions enable us to compare the effect of simple auditory feedback, received by learners in control condition, with that of advanced feedback, received by learners in test condition.

3.2 Acoustic cues

We estimated differences in height (F0) and duration ratios for `S1S2 and S1`S2S3 words with the lexical accent falling on the penultimate vowel, and we analysed the presence or absence of reduction in the second syllable of `S1S2S3 words.
To analyse all productions, including those not analysed by the platform during the test condition (second session), we segmented words into speech sounds, and estimated the values of the acoustic cues taken into account, i.e. fundamental frequency and segment duration. F0 was evaluated in semi-tones.
For words with `S1S2 and S1`S2S3 syllabic structures, we estimated the following criteria.

1. The differences between F0 values of the (to be) stressed vowel (VS) and the following unstressed vowel (VUF, for unstressed final vowel) :
   F0(VS) - F0(VUF) .
   F0 was averaged across all the frames of each vowel except the frames close to the vowel boundaries.
2. The number of times the F0 maximum fell on the right syllable (the theoretically stressed one), given in percentage.
3. The ratio of the duration of the stressed vowel to that of the following unstressed vowel:
   D(VS)/D(VUF).
   Note that VUF is the last vowel of the word, and could be lengthened by French subjects. Taking into consideration ratio and not difference in the above formula allows us to remove the effect of temporal variations due to speech tempo.

For `S1(S2)S3 words, we noted whether French learners pronounced the second vowel (V2) with its full timbre. To estimate the presence/absence of reduction, we calculated the duration of V2, when it was audible, and we compared it to the averaged duration of V1 and V3. If the duration of V2 was inferior to 40 ms and at least two times shorter than the averaged duration of other vowels in the word, we considered that the vowel had been reduced. With these criteria, all the second vowels in `S1(S2)S3 words uttered by the English speakers used as references in this study are considered as reduced.
We had also compared F0 height and durations between the stressed and the initial unstressed vowels (for S1`S2S3 words) but these parameters did not contribute a lot to global results and we do not give them here in order to simplify the discussion.

We used Student's t-tests, paired samples, associating the data of each speaker and each word to compare the pronunciation of the words in the first session (without feedback) to their pronunciation in the second session (with feedback). Each condition (auditory feedback and advanced feedback) and each group of speakers (i.e. advanced speakers and beginners) were tested separately. We also used Student's t-tests to compare results obtained in both conditions (only words uttered during the second session were taken into account). Once more advanced speakers and beginners were considered separately. We accepted a level of 0.05 for significant effect and considered that results were highly significant when this level was inferior to 0.001. We submit the following parameters to statistical tests: the first parameter (F0 height), the third parameter (relative duration ratio) as well as, for vowel reduction in 'S1S2S3 words, the ratio of the duration of the second vowel to that of other vowels in the words.

4. Results

Table 1 provides, for each session, the results obtained for advanced learners and beginners as well as the results for both reference speakers. Parameter values are averaged across all words and all speakers of a given level (advanced or beginner). For estimating the averaged differences in height (in semi-tones) between stressed and unstressed vowels (column 2), we only took into account the cases where the F0 maximum is on the correct syllable. The third column displays the number of times the F0 peak is located on the right syllable, and the fifth column the number of times the second vowel in `S1S2S3 words is reduced. The duration ratios are given in the fourth column.
We also calculated the averaged values obtained for each speaker in each parameter, and we discuss them below, when interesting.

Speakers	F0 (ST)	Max F0 place (%)	Duration ratio	Reduction (%)
REF (M)	11	100	1.7	100
REF (F)	9	100	1.7	100
Adv.	6.9	90	1.05	7.5
Adv.	7.4	100	1.2	100
Adv.	7.5	100	1.25	100
Beg.	2.5	33	0.85	2
Beg.	2.9	43	0.83	37
Beg.	3.9	75	1.1	63

Table 1. Results for the first session, without feedback (blue line) and for control and test conditions with feedback (white and green lines, respectively), for advanced learners (Adv), and beginners (Beg). Values for reference speakers, the male (M) and female (F) are given on the top of the table. See text for more explanations.

4.1 First session

In this session, all learners (from test and control groups) uttered the words of the corpus without listening previously to their English references.

4.1.1 Reduction phenomon in `S1S2S3 words

Subjects utter these words the French way: they realize post-stressed vowels with their full timbre. We note only few exceptions: one advanced speaker reduces the vowels in approximately a third of the cases (probably when he knows well the word pronunciation) and one beginner realizes one reduced vowel. For all other realizations the duration of the second vowel varies between 45 and 130 ms and its averaged duration with respect to other vowels of the word is approximately 0.8, for both groups. The averaged percentages of reduction were 7.5% and 2% for advanced learners and beginners, respectively.

4.1.2 Vowel duration ratio in `S1S2 S1`S2S3 words

Most speakers do not master L2 vowel durations (the averaged ratios are 1.05 and 0.85 for advanced speakers and beginners). Indeed, for all learners but one, the duration ratio (d(VS)/d(VUF)) is in between 0.7 and 1.1 (vs. 1.7 for English speakers). This confirms the tendency, well known for French learners, to lengthen the last syllable of the word and make relatively weak differences between syllable durations. The duration ratio is relatively high (1.5) and close to that of English speakers for only one speaker (an advanced one).

4.1.3 F0 pattern in `S1S2 and S1`S2S3 words

The realisation of this pattern seems to provide a good indication of the degree of proficiency of the subjects involved in this experiment. Let us first present the case of speakers with a low production level. We observe two kinds of pattern for these learners. Most speakers exhibit relatively flat patterns, with sometimes a tendency to rise F0 at the end of word, a behaviour typical of beginners and people apprehensive about speaking in a foreign language. One speaker exhibits systematically falling patterns. Since in the corpus, the accent never falls on the last syllable of the word, the percentage indicating the number of times the F0 maximum is on the right syllable is relatively low (33%). The difference in height, only taken into account when the F0 maximum is on the correct syllable, is weak (2.5 semi-tones). The advanced learners tend to exhibit correct F0 patterns, with the location of the F0 peak most of the times (90%) on the expected syllable and a substantial difference between the height of the stressed vowel and that of the post-stressed vowel (6.9 semi-tones).

4.1.4 Summary

To summarize the results of this first session, we remark strong differences in the mastering of F0 pattern and duration cues. Concerning duration, all speakers but one apply L1 duration rules when speaking English. Most of them are not aware of the reduction of post-stressed vowels, and others seem to encounter problems in predicting its occurence. On the contrary, some speakers, the more advanced ones, realize correct F0 patterns.

4.2 Second session. Effect of feedback for advanced learners

In this subsection, we examine the effet of feedback for advanced learners: auditory feedback (control condition) and diagnosis, F0 displays, as well as speech modification (test

condition). There is no statistical difference between results for advanced speakers in test and control conditions so results for both groups are discussed at the same time.

4.2.1 Reduction phenomon in `S1S2S3 words

Feedback has a very high impact on advanced speakers, both in amplitude and degree of significance, whatever the condition. Indeed, in both condition, all speakers change their pronunciation to realize strongly reduced vowels for all words. We thus obtain a score of 100% for all speakers. It then appears here that auditory feedback was sufficient to enable advanced learners to improve their realization.

4.2.2 Vowel duration ratio in `S1S2 and S1`S2S3 words

All speakers who exhibit relatively low averaged ratios –i.e. all but one- improve themselves, whatever the condition, showing once more the high impact of auditory feedback for advanced learners. The averaged ratios are 1.2 and 1.25 in control and test conditions, respectively. This increase is mainly due to the lengthening of the stressed vowel, rather than a shortening of the last one.

4.2.3 F0 pattern in `S1S2 and S1`S2S3 words

We observed slight but significant increases (of about 0.6 semi-tones) in F0 relative height in both conditions. This might not be important on a perceptual point of view, but it seems that learners are aware of the important difference in height exhibited by the reference speaker between the stressed and the following unstressed vowel and try to imitate it. There was no more error in the location of F0 maximum.

4.2.4 Summary

Due to effect of feedback in both conditions, advanced speakers change the relative duration of the vowels, and reduce vowels that have been strongly reduced by reference speakers. For learners involved in this study, auditory feedback appears to be as efficient as more complex feedback such as the one proposed in this study. This result appears to be in agreement with that obtained by Herry and Hirst (2010) who tested advanced learners.

4.3 Second session. Effect of feedback for low production level speakers

4.3.1 Reduction phenomon in `S1S2S3 words

Auditory feedback does not have a high impact on low production level speakers. According to the comments they made after the experiment, they seem to have been disturbed by the way English speakers utter seemingly familiar words (transparent words) and slightly (but significantly) improved their realizations. On the whole, all speakers diminish the averaged duration of the second vowel but the averaged percentage of reduction is relatively low (37%).

Of course the effect of advanced feedback is far better. Indeed, the system detects the number of syllables and when this number is different from that of the reference, subjects were informed of this deviation and asked to change their pronunciation. Then, aware of what is expected, subjects make strong efforts to reduce vowels. The results varies with speakers and words: only one speaker reduces all vowels, but all speakers drastically reduce

V2 averaged duration (this reduction was highly significant). The average percentage of reduction was 63%. The difference between both conditions is highly significant.

4.3.2 Vowel duration ratio in `S1S2 and S1`S2S3 words

In control condition, there is no significant improvement with respect to vowel duration. The averaged duration ratio (d(VS)/D(UF)) estimated for all speakers stay in the same range as that observed in the first session.
In test condition, subjects significantly improved their realization, making generally longer stressed vowels (the averaged ratio rises from 0.84 up to 1.1). The difference between test and control condition is highly significant.

4.3.3 F0 pattern in `S1S2 and S1`S2S3 words

In control condition, the observed improvement is very small in amplitude (0.4 semi-tones) and significance and varies with speakers and words. Two speakers exhibit clear tendencies to raise the pitch of the stressed syllables with respect to the unstressed one, but this rise not systematic, i.e. not observed for all words. The overall percentage of words with the maximum of F0 on the right syllable increases from about 10%.
In test condition, there is a very significant improvement. All speakers were clearly informed of what was expected from them (thanks to small texts and arrows) and improved their realization most of the times (the percentage concerning the location of the F0 maximum raises from 30% up to 75%). The amplitude of the modification (1.4 semi-tones in the average) varies with speakers and words. Once more, the difference between test and control conditions is highly significant.

4.3.4 Summary

Due to effect of "advanced" feedback, subjects with a low oral production level significantly improve their realizations. Advanced feedback appears to more interesting than simple auditory feedback for these subjects.

5. Conclusion

Feedback on L2 prosody based upon visual displays, speech modifications and automatic diagnosis has been elaborated and a pilot experiment undertaken to test its immediate impact on listeners. Results show that the various kinds of feedback provided by the system enable French learners with a low production level to improve their realisations of English lexical accents more than (simple) auditory feedback. These results shoud be reinforced with a large number of speakers but based upon the important differences between results obtained for speakers in test and control conditions, we are confident in the interest of the system presented here. In particular, the system analyses learners' realisations and provide indications on what they should correct, a guidance which is considered as necessary by specialists in the oral aspects of language learning, such as Chun (1998) or Germain-Rutherford (Germain-Rutherford & Martin, 2000).
The perspectives of this work are twofold, technological and experimental. On a technological point of view, results from the experiment encourage us to pursue the work

undertaken with the speech team ("Parole" team) at LORIA on speech alignment for non-native speakers. This work concerns both the detection of precise segment boundaries when possible and the elaboration of confidence levels on boundary detections. The design of such confidence levels would allow the system to propose feedback only when it can rely upon detections with high confidence levels and thus avoid erroneous corrections. We also plan to refine our method for modifying learners' voice. We use (an improved version) of TD-PSOLA method to modify segment durations and melodic curves. This algorithm does not allow us to modifiy vowel quality. Yet automatic modifications of vowel quality would be interesting to take into account vowel (timbre) reduction in post-stressed vowels. This modification could be obtained by using techniques such as voice morphing or techniques working in the frequency space. Since these techniques could generate degradations in speech quality, compromises should be found.

On an experimental point of view, it would be interesting to separate the impact of each kind of feedback (visual displays, diagnosis and perceptual feedback based upon speech modification). We also plan to investigate the long-term effect of automatic feedback should in collaboration with teachers in foreign language at the University of Lorraine.

6. Acknowledgment

This work has been partially funded by the INTERREG (european project) ALLEGRO. We would like to thank Odile Mella for her help concerning the manuscript, the reference speakers, and the subjects participating in the experiment, as well as Yves Laprie, Guillaume Henry and Christian Gillot for their contribution to the software.

7. References

Bouselmi, G.; Fohr, D.; Illina, I. & Haton J.-P. (2005). Fully Automated Non-Native Speech Recognition Using Confusion-Based Acoustic Model Integration, *Proceedings of Interspeech* Lisbon, Portugal.

Chun, D. (1998). Signal analysis software for teaching discourse intonation. *Language learning and technology*. Vol. 2, No. 1, pp. 61-77.

Colotte, V.; Laprie, Y. & Bonneau, A. (2001). Percpetual experiments on enhanced and slowed down speech sentences for second language acquistion. *Proceedings on the international conference on speech communication and technology (ICSLP),* Aalborg, Denmark.

Colotte, V. & Laprie, Y. (2002). Higher pitch marking precision for TD-PSOLA, *Proceedings of European Signal Processing Conference (EUSIPCO),* Toulouse.

Dauer, R. (1983). Stress-timing and syllable-timing reanalysed. *Journal of Phonetics,* Vol. 11, pp. 51–62.

De Bot, K. (1983). Visual feedback on intonation I: Effectiveness and induced practice behaviour. *Language and Speech* , Vol.6 , No.4, pp. 331–350

Hazan, V. & Simpson, A. (1998). The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise," *Speech Communication*, vol. 24, pp. 211–226, 1998.

Fohr, D.; Mari, J. F. & Haton, J. P. (1996). Utilisation de modèles de Markov pour l'étiquetage automatique et la reconnaissance de BREF80, *Journées d'Etude de la Parole,* Avignon, France.

Germain-Rutherford, A. & Martin, P. (2000). Utilisation d'un logiciel de visualisation pour l'apprentissage de l'oral en langue seconde, *Apprentissage des Langues et Systèmes d'Information et de Communication (ALSIC),* Vol. 3, No.1, pp. 71-86.

Henry, G. ; Bonneau, A. & Colotte, V. (2007). Tools devoted to the acquisition of the prosody of a foreign language, *Proceedings of the International Congress of Phonetic Sciences*, Saarbrücken, Germany.

Herry, N. & Hirst, D. (2010). Subjective and Objective Evaluation of the Prosody of English Spoken by French Speakers: the Contribution of Computer Assisted Learning, *Speech Prosody*, Aix en Provence, France.

Ingram, J.; Mixdorff, H. & Kwon, N. (2009). Voice morphing and the manipulation of intra-speaker and cross-speaker phonetic variation to create foreign accent continua: A perceptual study, *SLaTE 2009 ISCA International Workshop on Speech and Language Technology in Education*, Wroxall Abbey Estate, Warwickshire, England. 3-5 September 2009.

James, E. (1977). The Acquisition of a Second-Language intonation Using a Visualizer. *Canadian Modern Language Review*, Vol.33, No.4, pp. 503-506.

Kommissarchik, J. & Kommissarchik, E. (2004). BetterAccent Tutor- Analysis and Visualization of Speech Prosody. *Proceedings of InSTIL/ICALL,* Dundee, Scotland.

Laprie, Y. (1999). Snoori, a software for speech sciences. *Proceedings of MATISSE*, London, England.

Loizou, P. (1998). Mimicking the human ear. *IEEE Signal processing magazine*. September 1998 pp. 101-130.

Martin, P. (2004). WinPitch LTL II, a Multimodal Pronunciation Software. *Proceedings of InSTIL/ICALL,* Venice, Italy.

Moulines, E. & Charpentier, F. (1990). Pitch synchronous wave-form processing techniques for a text-to-speech synthesis using diphones. *Speech Communication,* Vol.9, No.5-6, pp. 453-467.

Ortega, M. and Hazan, V. (1999). Enhancing acoustic cues to aid L2 speech perception. *Proceedings of the International Congress of Phonetics Sciences,* San Fransico, California. pp. 117-120.

Tallal, P.; Miller, S.; Bedi, G.; Byma, G.; Wang, X.; Nagarajan, S.; Schreiner, C.; Jenkins, W. & Merzenich, M. (1996). Language comprehension in language-learning impaired children improved with acoustically modified speech, *Science*, vol. 271, pp. 81–84.

Vardanian, R. (1964).Teaching English through oscilloscope displays, *Language Learning*, Vol.14, No. 3-4, pp. 109-117.

Yanagisawa, K. & Huckvale, M. (2007). Accent morphing as a technique to improve the intelligibility of foreign-accented speech, *Proceedings of the International Congress of Phonetics Sciences,* Saarbrücken, Germany.

www.tellmemore.com

Auralog
www.langmaster.com
LANGMaster
www.loria.fr/~laprie/WinSnoori
Yves Laprie