Automatic prosody modification as a means for foreign language pronunciation training

Anna Sundström
Centre for Speech Technology, TMH/KTH, Stockholm
anna@speech.kth.se

Abstract
This paper proposes a method that allows students to hear the correct prosody of a foreign language as spoken with their own voice. This is made in two steps: First a phrase in the foreign language spoken by a student is compared to a reference pronunciation of the same phrase, which can be spoken by a teacher or provided by a CALL (Computer-Assisted Language Learning) system. Second, the student’s speech is resynthesized with duration and fundamental frequency parameters from the reference pronunciation. Our current results show that the method works well for single words and a given phonetic transcription.

1. Background
One prominent aspect of foreign accents is in many cases a deviant prosodic realisation. Therefore it is very important to train prosody in foreign language classes. Part of the difficulty in getting a correct prosody lies in perceiving the significant differences between one’s own pronunciation and the teachers. To apprehend the significant differences, it would be helpful for a student to hear his/her own speech with correct prosody.

In speech training for disabled speakers, the student’s current best production is often used as a model to aim towards [1]. In this project a somewhat similar method is used: The CALL system improves a foreign language student’s current best production by imposing prosody from the teacher’s speech. Thus the student can hear his/her own speech with correct prosody.

2. Overview
In this project a preliminary system for prosodic modification of student’s speech is developed, which in a later step can be incorporated in a CALL system.

A student records a phrase in the foreign language. The reference pronunciation is recorded by the teacher in the classroom or provided by the CALL system. Both student and reference speech are aligned with a phonetic transcription. For every phoneme in the transcription, the duration and fundamental frequency (F0) values for the two speakers are compared. Resynthesis of the phrase is made on basis of the student’s speech with the computed duration and F0 changes. An overview of the prosody modification system is shown in figure 1.

The resynthesis method which is used is Pitch Synchronous OverLap and Add (PSOLA) [2,3]. PSOLA was chosen because it preserves the naturalness of the speech well.

The student’s speech is phonetically labelled with the method described in [4]. All F0 extractions are made with the ESPSTM package.

3. Pronunciation models used as teacher
One of three different pronunciation models has been used as reference in the current study: a speech database, a phrase recorded by the teacher in the classroom, or text-to-speech synthesis.

3.1 Speech database as reference pronunciation
In the CALL system a student can choose among the phrases in a database, listen to a phrase and record his/her own attempt. If the database is not phonetically labelled, this is done in the same way as with the student’s speech. The F0 is also extracted.

3.2 Recorded teacher speech as reference pronunciation
The teacher in the classroom can record phrases, which can be used directly. This is useful if the phrases are not in the database or if the teacher wants to emphasize special prosodic features.

If the teacher gives the text for each phrase, a phonetic transcription is taken from a dictionary and the speech is labelled in the same way as the student’s speech. Otherwise phone recognition of the teacher speech is used to get the transcription. The recognizer is the general purpose Viterbi recognizer HVite from HTK [5] with a network and dictionary of 40 Swedish phonemes.

When the teacher’s speech is labelled or recognized, the F0 is extracted.

3.3 Text-to-speech synthesis as reference pronunciation
The pronunciation rules in the text-to-speech synthesis system at KTH [6] can be utilized as reference pronunciation. Currently, the synthesis is used in the same way as the speech database: the speech from the synthesis is aligned with the phonetic transcription and the F0 is extracted. However, the intention is to use the duration and F0 values given to the synthesizer directly without necessarily producing the synthesis.

4. Labelling and alignment
Before the student’s speech and reference pronunciation can be compared, the speech must be labelled and aligned. Instead of using words for the labelling, a phonetic transcription is given to the labelling program. The phonemes in the transcription are treated as single units. Pauses may be inserted between all phonemes.
This is especially useful if the student inserts pauses where there are no pauses in the reference pronunciation. So far, the pauses are ignored in the further treatment.

If the reference pronunciation is already labelled, the existing phoneme sequence is used for the labelling of the student’s speech. Otherwise the reference is labelled with a transcription from a dictionary, or phone recognition is used, which gives the transcription for the labelling of student’s speech.

As the labelling of the student and teacher speech use the same phonetic transcription, the utterances can be aligned with each other.

5. Resynthesizing the student’s speech

After labelling and alignment, the phoneme durations for the student and the reference pronunciation are compared. For each phoneme, the quotient between the reference’s duration and the student’s is used for the duration changes.

In the PSOLA synthesis method, duration and F0 changes can be made at the same time. To get F0 changes at the correct places in the student’s speech, one has to consider both F0 (for student and reference) and the duration changes that will be made. For every 10 milliseconds of the student’s speech, a new F0 value is extracted from the reference F0 curve. The point of time of the reference speech where the F0 value should be taken from is computed from the known duration differences between the student’s speech and reference speech. As the duration changes vary between phonemes, all F0 values of the reference are not used, while some of the values are used more than once.

It is important that the resynthesis has the same pitch level as the student's ordinary speech. Therefore the mean values of the student's F0 and the reference's are computed. The quotient of these two values is used to adapt the reference pitch level to the student's.

On the basis of the student’s speech, PSOLA synthesis is made with the computed duration and intonation changes. The tools for making PSOLA synthesis are developed by Gregor Möhler at Stuttgart University [7] as an extension to the ESPS™ package.

6. Pilot experiments

The automatic prosody modification was tested with recorded speech from four Bosnian students learning Swedish. The speech data was taken from the database described in [8]. The recordings were made on cassette tape and then digitized.

Four two-syllabled words and one sentence from every speaker were used, which made a total of 16 words and 4 sentences. Two of the words had the Swedish accent II, which is often difficult to pronounce for foreign students.

![Diagram](image.png)

**Figure 1.** Overview of the prosody modification system. The dashed rectangle shows the reference pronunciation, which in this case is recorded by the teacher in the classroom.
As reference pronunciation, the same words and sentence from a female Swedish speaker were used. The text-to-speech synthesis system at KTH was also used as reference pronunciation [6].

The evaluation was done by first carefully listening to the resynthesized speech, and for every word or sentence decide whether it sounded like natural speech or not. After that, the phrase was listened to again and compared to the original reference pronunciation, and it was decided whether the prosody sounded the same or not.

6.1 Prosody modification of single words
A number of experiments were done for the single words. The words which were used were the Swedish words "alla", "elev", "järnväg" and "sönder" (meaning "all", "pupil", "railway" and "broken"). First the modifications of the student’s speech were done with the human reference pronunciation and a given phonetic transcription. Second, the same reference was used without given transcription. Third, the modifications were done with the text-to-speech synthesis at KTH as reference pronunciation.

For comparison, the first experiment was redone with manually labelled speech data.

6.1.1 Human reference with given transcription
When the human reference pronunciation with given transcription was used, the modified student’s speech sounded like natural speech in 14 of the 16 words. For all these words, the intonation contour sounded like the reference. For the two words which did not sound like natural speech, the speech was distorted due to bad alignment between the reference and the student’s speech. Resynthesis of one word is shown in figure 2.

In this experiment, additional measurements on the resynthesis were made. The resynthesized speech was manually labelled and the phoneme durations compared to the phoneme durations of the reference speech. The form of the resynthesized F0 curves was also examined.

The results of the phoneme duration comparisons can be seen in figure 3. The four words consisted of 3, 4, 6 and 6 phonemes respectively. Four speakers who spoke four words each made a total of 76 phonemes in the resynthesized words. For 29% of these phonemes the new duration is between 80% and 120% of the duration in the original reference speech. An amount of 55% of the new durations lie between 50% and 150% of the original reference duration.

The F0 curves for the resynthesized speech were dislocated in correspondence to the displacement of the phoneme boundaries. However, the form of the F0 curve corresponded well to the reference for 10 of the 16 words. For four of the words, the correspondence was good except for the last phoneme.

It seems that the modification was more effective for the intonation than for the durations.
6.1.4 Manually labelled student and reference speech
The manual labelling of the speech data started from the automatic labelling which was then adjusted in a program where a spectrogram can be seen and segments listened to [9]. This was performed with both student and reference speech.

The prosody modification of manually labelled speech was more accurate than the modification of automatically labelled speech. All modifications of single words sounded like natural speech, and the prosody was closer to the reference prosody.

6.2 Prosody modification of whole sentence
With the whole sentence, the same experiments were done as with the words, but only for one of the speakers. The sentence was in Swedish "Alla barn hade ledigt från skolan i flera dagar" (meaning "All children were free from school for several days"). The automatic alignment between reference and student’s speech was incorrect, resulting in big distortions in the resynthesized speech.

With manually labelled speech the prosody modification was performed for all four students. For one of the speakers, the resynthesis sounded like natural speech and had a prosody similar to the reference. For the other three, most of the phrase sounded natural, but the end was distorted due to limitations in the resynthesis of big duration changes. In one of the cases the student had omitted one syllable in the last word, which made the alignment with the reference speech difficult.

7. Conclusions
The next step in the development should be to optimize the recognition, labelling and alignment better. Pauses which are detected in reference and student’s speech should be made use of instead of being treated as part of the surrounding phonemes. The new F0 curve needs to be smoothed in some way if a phoneme duration is extended several times.

The PSOLA synthesis can also be utilized in a better manner. Now the information about pitch periods is taken from the F0 file. Instead a label file with pitch synchronous information can be used, which gives better resynthesis quality.

In a full-scale working system for prosody modification of student’s speech, a module for assessing the student pronunciation should be incorporated. If the pronunciation is not good enough on the phonetic level, the prosody modification should not be done. This is both for technical and pedagogical reasons: with a pronunciation that differs too much from the teacher pronunciation the resynthesis will not sound natural, and with bad pronunciation of individual phonemes the resynthesis will not give a correct pronunciation and the student might maintain an incorrect pronunciation.

Acknowledgements
This work started as a M.Sc. thesis and has been developed at CTT, Centre for Speech Technology. The PSOLA tools were kindly supplied by Gregor Möhler at Stuttgart University.

References