Using Speech Analysis Techniques For Language Learning

Valery A. Petrushin
Center for Strategic Technology Research, Accenture
3773 Willow Rd., Northbrook, IL 60062
petr@cstar.accenture.com

Abstract
This paper presents an overview of speech analysis, visualization, and student response evaluation techniques that can be used for learning a foreign language by an adult learner. The general framework for student response evaluation is described. It is based on collecting experimental data about experts’ and novices’ performance and applying machine learning and knowledge management techniques for deriving evaluation rules. Application of the proposed approach to language learning tasks is discussed.

1. Introduction
Learning a foreign language is a difficult and time-consuming task. The best results are achieved in one-to-one interactions with a teacher who is a native speaker. Unfortunately, this approach is not affordable for most learners. Advances in speech technology resulted in proliferation of speech-enabled commercial language learning products that suppose to improve the quality and speed of language learning for a reasonable price. These products use commercially available speech toolkits, which were created for building voice-enabled applications for native speakers rather than for teaching foreigners. On the other hand, there are commercial products that can perform sophisticated low-level speech analyses that have been traditionally used for pathological voice evaluation. Many of these analyses can be successfully applied to foreign language learning. The main drawback of the current computerized language learning products is very weak feedback. Some systems can tell the user if his or her response is correct or wrong, or represent the quality of response on a scale “bad-satisfactory-good”. The learner’s frustration skyrockets when he gets several “wrong” or “bad” grades without any hints how to improve his performance. The learner needs a system that helps to visualize his performance, compare it to the teacher’s performance, and provide feedback on how to improve it.

A digitized speech signal is represented as a sequence of 8-bit or 16-bit integer numbers sampled at a frequency that is twice larger than the maximal frequency of the signal. The graphical representation of speech signal as a function of time is called a waveform. This representation might be very confusing for learner, and I don’t recommend showing it. The following more informative characteristics can be extracted or associated with speech signal: energy, pitch, formants, speaking rate, spectrum, cepstrum and transcription [1].

2. General framework
Learning to speak a foreign language involves the development of new motor skills, i.e. new movements of one’s speech organs. To expedite the process, precise diagnostics of wrong movements and detailed description on how to fix them are necessary. The general framework for evaluating learner’s performance for a particular task includes the following steps:

- Create a descriptive model for the task. The model describes gestures of the tongue, lips and jaw that are necessary to perform the task correctly.
- Select acoustic features and create a quantitative model of the task. For example, for learning vowels two formants F1 and F2 were selected as the features, and a two-dimensional Gaussian model was built for each vowel based on TIMIT database.
- Collect experimental data from native speakers and learners. For the vowel learning task performance data were collected and manually classified as correct or wrong. The recommendations on how to improve performance were created for each case of wrong performance.
- Use machine learning and knowledge management techniques for creating a diagnostic system. The diagnostic system contains a set of rules that tells how to compare the teacher and learner’s data and gives recommendations how to fix learner’s wrong performance. For the vowel learning tasks I used a decision three classifier that was based on the experimental data.
- Use visualization techniques to present data to the learner. In case of vowel learning, a F1-F2 chart was used for displaying teacher and learner’s data.
3. Language learning tasks

3.1. Learning sounds

Sounds or phonemes are the basic elements of speech. There are two broad categories of sounds: vowels and consonants. Vowels are further divided into pure vowels and combined vowels – diphthongs and triphthongs. It is well known that vowels can be distinguished by their first (F1) and second formants (F2). Hence the best visualization for practicing a pure vowel is to calculate F1 and F2 for sequential speech frames of length 10 or 20 ms for both the teacher’s and the student’s utterances and display them on a F1-F2 chart. To estimate the student’s performance the probability that the student’s data belong to the model is calculated and compared to a threshold. If the probability is less than the threshold than the system uses rules to do diagnosis and give recommendations. Some rules can be derived from historical data using data mining techniques such as decision trees; the others are created by an expert. Knowledge of the student’s native language can help significantly in creating precise diagnosis and useful recommendation because the typical mistakes are: 1) substituting the vowel by the closest vowel of student’s native language, and 2) substituting a non-stressed vowel by the neutral vowel (schwa).

A F1-F2 chart for a diphthong contains initial vowel’s data, transitional data and final vowel’s data. Besides the F1-F2 chart, it is very important to track the dynamics of a combined vowel. To achieve this, I suggest comparing teacher’s and student’s F1 and F2 profiles. The diagnostic rules take into account the probabilities of the student’s data to belong to initial and final vowels, and closeness of formant profiles.

Consonants are distinguished by their spectral and voiced/unvoiced features. The voiced/unvoiced indicator, spectrum and spectrogram can be used to compare teacher and student’s sounds.

3.2. Syllabic intonation and stress

In polytonal languages, such as Chinese, Tai, and Vietnamese, the meaning of a word depends on syllabic stress or tones. There are four tones in Mandarin Chinese, five tones in Tai, and eight tones in Cantonese. Learning tones is very hard problem for most of learners. But it could be much more easier if the system visualizes student’s pitch profile and compares it to the teacher’s profile. The figure above presents a student’s try to pronounce Mandarin word /wa/ with third (falling-rising) tone. Here the solid line represents a teacher’s and dashed line represents a student’s profiles. Both profiles are aligned, scaled, and normalized. The distance between profiles is calculated and compared to the threshold. The diagnostic rules take into account the duration and closeness of profiles. This approach can be extended to give an opportunity to the student to learn relationship among tones in multi-syllable words [2]. Most languages use energy stress, but some of them, such as Japanese, use tonal stress, i.e. a speaker raises his or her pitch to stress a word. Learning tonal stress versus energy stress and vice versa is a hard problem. Visualizing and evaluating the teacher’s and learner’s pitch and energy profiles can greatly improve the student performance. Another neat visualization of stress based on pitch, energy, and spectrum is proposed in [3].

3.3 Learning words, phrases and sentences

Learning to pronounce multi-syllable words, phrases and sentences is another hard problem. This problem has two sides: 1) how well recognizable is a learner’s pronunciation, and 2) how “natural” is the learner’s speech. The first side deals with quality of sounds. But the second side deals with prosodic features (intonation, stress, and pauses). Some language educators claim that the prosodic part of speech is more important for oral communication. In spite of some native speakers have problems in pronouncing some sounds they are perceived as native speakers based on their prosody and vocabulary. To address prosodic problems, visualization of teacher and learner’s pitch and energy profiles is recommended. Diagnostic rules can be created for particular words, phrases and sentences. However, creating a comprehensive diagnostic system is a hard problem, which waits to be solved. But giving a learner a visual feedback itself can improve his or her performance. Using a speech recognition system allows estimating how recognizable is the learner’s speech. Some speech recognition systems can output several possible transcriptions (N-best matches) ranked in accordance with their likelihood. The diagnostic rules can use this output in combination with prosody features for tailoring more precise recommendations.

References