Student Response Evaluation for Spoken Language Learning: A Case Study of Learning Chinese Tones

Valery A. Petrushin
Accenture Technology Labs, Accenture
161 N. Clark St., Chicago, IL 60601, USA
E-mail: petr@techlabs.accenture.com

Abstract

This paper presents speech analysis, visualization, and student response evaluation techniques that have been used for computer-aided 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. A case study of learning tones of Standard Chinese (Mandarin) is discussed in details.

1. Introduction

Learning a foreign language is a difficult and timeconsuming 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, such as the IBM’s ViaVoice™, Dragon’s Naturally Speaking™, or Lernout & Hauspie’s VoiceXpress™, which were created for building voice-enabled applications for native speakers rather than for teaching foreigners. They have means to adapt themselves to a particular speaker but cannot teach a learner the normative speech. On the other hand, there are commercial products that can perform sophisticated lowlevel speech analysis. For example, the Kay Elemetrics’ Speech Lab suite of speech analysis products can perform various types of speech analyses that have been traditionally used for pathological voice evaluation. As I show below, many of these analyses can be successfully applied to foreign language learning. The main drawback of the current computerized language learning products is a 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 or she gets several “wrong” or “bad” grades without any hints how to improve his/her performance. The learner needs a system that helps to visualize his/her performance, compare it to the teacher’s performance, and provide feedback on how to improve the performance. The rest of the paper has the following structure. First, I shall introduce characteristics (features) of speech signal and various types of speech analysis. Then I shall describe a general framework for applying speech analysis, machine learning and visualization techniques for spoken language learning. After this I shall demonstrate how the methodology works using a case study of learning Standard Chinese (Mandarin) tones.

2. Speech signal features

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. For example, speech dictation systems require a signal that has frequencies up to 8-10 kHz; hence the signal is sampled at 16 kHz or 22.05 kHz. Figure 1 presents some speech features for an utterance “Accenture” pronounced by an adult male. The graphical representation of speech signal as a function of time is called a waveform (see Figure 1a). 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. Energy characterizes the volume of speech at a particular moment of time. It is calculated as a mean of squared samples for each time interval. Figure 1b depicts the energy of the signal for each 10 ms. The graphical representation of energy can be useful for a learner. Pitch is represented by the fundamental frequency (F0) that is the dominating frequency of the sound produced by the vocal chords.
Pitch exists only for voiced fragments of speech, i.e. for vowels and voiced consonants. Figure 1c presents pitch as a function of time for the above utterance.

Formants are the resonances of the vocal tract and their values depend on the positions of speech organs (tongue, velum, lips, teeth, and jaw). Formants also make sense to consider only for the voiced part of the utterance. Figure 1d depicts two first formants as dark dashed lines superimposed on a spectrogram of the above utterance. Speaking rate characterizes the speed of speech. Roughly it is proportional to the number of syllables a speaker pronounces per second. It is calculated as the inverse of the average length of the voiced part of utterance. The speaking rate for the above utterance is 10.4. The spectrum gives a picture of the distribution of frequency amplitude at a particular moment in time. Figures 2a and 2b present spectrum for vowel /e/ at the moment 0.45 sec and for the consonant /s/ at the moment 0.35 sec correspondingly. The horizontal axis represents frequency, and the vertical axis amplitude. The bumps on Figure 2a correspond to formants of vowel /e/. The spectrum of consonant /s/ (Figure 2b) looks quite different; it has no formants and contains much more energy in high frequency band. The spectrogram presents the spectrum of the speech signal over time. The horizontal axis represents time, and the vertical axis represents frequency. The third dimension – frequency amplitude – is represented by shades of darkness. Figure 1d shows a spectrogram of the above utterance. Cepstrum can be considered as a spectrum of spectrum at a moment in time. Two kinds of cepstrum are in use: linear prediction cepstrum and mel-frequency cepstrum [1]. The cepstrum coefficients are the standard input for speech and speaker recognition systems. The cepstragram presents the cepstrum of the speech signal over time. Transcription is a sequence of phonemes or words that correspond to the utterance, for example, transcription for the utterance depicted on Figure 1 is /ak-'sench&r/. An utterance also has such characteristics as duration, number of voiced, unvoiced and silent parts, their duration, duration of phonemes, etc. Beside this, a number of statistics, such as mean, standard deviation, maximum, minimum, range, etc., can be calculated for the abovementioned characteristics.

There are two signals in a language-learning task – a sample utterance from the expert (teacher) and a student’s response. The problems are how to choose the proper characteristics, how to compare the utterances, and how to visualize the results of comparison. In the next section I shall describe the general framework for learner’s performance evaluation.

3. 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 [2], and a two-dimensional Gaussian model was built for each vowel based on TIMIT database [3];
- **Collect experimental data from native speakers (teachers) 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.

The typical language-learning tasks and some solutions using the above approach are discussed in [4]. In the next section I shall describe in details how to use the above approach to learn syllabic intonation or tones in polytonal languages.

### 4. A case study: learning Chinese tones

In polytonal languages, such as Chinese, Tai, and Vietnamese, the meaning of a word depends on syllabic intonation or tones. There are four tones in Standard Chinese (Mandarin), five tones in Tai, and eight tones in Cantonese.

Learning tones is very hard problem for the most of learners. Figure 3 shows profiles for four tones of Standard Chinese (Mandarin) language [5]. A native Mandarin speaker distinguishes among five pitch levels, which correspond to the following music notes: C, D, E, F#, and G#. These levels are coded by numbers from 1 to 5. The absolute pitch value does not matter but relative intervals do. Tone 1 (High or Plain) starts and maintains at the level 5. Tone 2 (Rising) starts at the level 3 and goes up to the level 5. Tone 3 (Low or Checking) starts at the level 2, goes down to the level 1, and, then goes up to the level 4. Sometimes a pause may occur in between falling and rising parts of the tone. Tone 4 (Falling) starts at the level 5 and goes rapidly to the level 1.

![Mandarin Tones](image)

Learning tones includes developing two skills: tone recognition and tone portraying. An experiment in tone recognition has been conducted using a dataset of 200 utterances (4 tones by 10 one-syllable words by 5 speakers). The average accuracy of recognition for two native speakers is 93.5%, for four speakers who were exposed to a polytonal language in childhood it dropped down to 83.5%, but for non-native learners it is 63.95%.

Tables 1 and 2 present the average confusion matrices for native (exposed) and non-native speakers correspondingly. The rows and the columns represent true and evaluated categories respectively, for example, second row of Table 1 says that 2.5% of utterances that represent tone 2 were evaluated as tone 1, 74.5% as true tone 2, 22.5% as tone 3, and 0.5% as tone 4.

#### Table 1. Accuracy of tone recognition for native speakers.

<table>
<thead>
<tr>
<th>Category</th>
<th>Tone 1</th>
<th>Tone 2</th>
<th>Tone 3</th>
<th>Tone 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tone 1</td>
<td>87.5</td>
<td>9.5</td>
<td>0.5</td>
<td>2.5</td>
</tr>
<tr>
<td>Tone 2</td>
<td>2.5</td>
<td>74.5</td>
<td>22.5</td>
<td>0.5</td>
</tr>
<tr>
<td>Tone 3</td>
<td>0.5</td>
<td>19.5</td>
<td>77.5</td>
<td>2.5</td>
</tr>
<tr>
<td>Tone 4</td>
<td>2.5</td>
<td>2.5</td>
<td>0.5</td>
<td>94.5</td>
</tr>
</tbody>
</table>

#### Table 2. Accuracy of tone recognition for non-native speakers.

<table>
<thead>
<tr>
<th>Category</th>
<th>Tone 1</th>
<th>Tone 2</th>
<th>Tone 3</th>
<th>Tone 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tone 1</td>
<td>70.8</td>
<td>13.2</td>
<td>4.4</td>
<td>11.6</td>
</tr>
<tr>
<td>Tone 2</td>
<td>9.0</td>
<td><strong>60.4</strong></td>
<td>18.2</td>
<td>12.4</td>
</tr>
<tr>
<td>Tone 3</td>
<td>10.0</td>
<td>10.0</td>
<td><strong>59.6</strong></td>
<td>20.4</td>
</tr>
<tr>
<td>Tone 4</td>
<td>21.6</td>
<td>8.6</td>
<td>4.8</td>
<td><strong>65.0</strong></td>
</tr>
</tbody>
</table>

Table 1 shows that the tones 2 and 3 cause a lot of confusion even for the speakers who were exposed to polytonal languages in their childhood. For non-native
speakers the pattern is quite different – the most confusion is caused by recognizing tone 3 as tone 4 (20.4%) and tone 4 as tone 1 (21.6%). But individual patterns can be very different. Tables 3a and 3b show confusion matrices for two learners. The learner A recognized many utterances as the tone 4, but the learner B put many utterances in the tone 1 category and recognized tone 2 very poorly.

Table 3a. Accuracy of tone recognition for learner A.

<table>
<thead>
<tr>
<th>Category</th>
<th>Tone 1</th>
<th>Tone 2</th>
<th>Tone 3</th>
<th>Tone 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tone 1</td>
<td>72.0</td>
<td>2.0</td>
<td>10.0</td>
<td>16.0</td>
</tr>
<tr>
<td>Tone 2</td>
<td>0.0</td>
<td>48.0</td>
<td>20.0</td>
<td>32.0</td>
</tr>
<tr>
<td>Tone 3</td>
<td>0.0</td>
<td>4.0</td>
<td>66.0</td>
<td>30.0</td>
</tr>
<tr>
<td>Tone 4</td>
<td>2.0</td>
<td>4.0</td>
<td>14.0</td>
<td>80.0</td>
</tr>
</tbody>
</table>

Table 3b. Accuracy of tone recognition for learner B.

<table>
<thead>
<tr>
<th>Category</th>
<th>Tone 1</th>
<th>Tone 2</th>
<th>Tone 3</th>
<th>Tone 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tone 1</td>
<td>60.0</td>
<td>22.0</td>
<td>10.0</td>
<td>8.0</td>
</tr>
<tr>
<td>Tone 2</td>
<td>28.0</td>
<td>30.0</td>
<td>16.0</td>
<td>26.0</td>
</tr>
<tr>
<td>Tone 3</td>
<td>18.0</td>
<td>4.0</td>
<td>56.0</td>
<td>22.0</td>
</tr>
<tr>
<td>Tone 4</td>
<td>22.0</td>
<td>16.0</td>
<td>8.0</td>
<td>54.0</td>
</tr>
</tbody>
</table>

4.1 Descriptive and quantitative models

The descriptive models were created for each tone. A learner should control only two acoustic variables (pitch and duration) to achieve correct performance. Examples of correct pronunciation have been provided for each tone. The theoretical descriptive models depicted on Figure 3 have been verified using the above-mentioned dataset. The experimental profiles and their characteristics are depicted on Figure 4, that shows experimental tone models for normalized time axis and logarithmic (musical) pitch axis. Moreover, all individual pitch ranges were aligned by subtracting minimal pitch values. As you can see, experimental profiles for the tone 2 and tone 3 are different from theoretical – for the tone 2 the first 1/3 of the profile is rather flat or even falling, and for tone 3 the falling and raising part have approximately the same range.

Utterance duration and the following pitch (fundamental frequency) statistics were used as features for the quantitative models: starting pitch, ending pitch, difference between starting and ending pitch, maximal pitch, minimal pitch, pitch range, mean of the pitch, median of the pitch, and pitch standard deviation. Beside these integral features, each pitch contour is represented as a normalized time series of 101 points of pitch values using logarithmic (musical) scale. Using the above integral features and native speakers’ (expert) data I have created 15 neural network classifiers. Most of them showed 100% accuracy for the test native speaker data.

Having applied them to learners’ data that has not been used for training I achieved 95.6% accuracy.

4.2 Developing diagnostic model

The tone-learning course consists of three stages. At the first stage a learner familiarizes herself with the concept of tone, listens to examples, and learns to recognize tones. During tone recognition exercises the learner listens to a randomly selected utterance and tries to determine its tone. The system informs the learner whether her guess is correct and plots the pitch contour for the utterance. The learner can listen to the utterance several times before picking up the next utterance. At the second stage the learner listens to an utterance pronounced by the expert and tries to replicate it. The system evaluates the learner’s response, plots the learner’s and expert’s pitch contours on the same graph, does diagnostics and displays recommendations. At the third stage the learner gets assignments to portray a word with a particular tone. The system evaluates the learner’s response. But instead of visualizing an expert’s pitch contour it displays the corresponding tone model profile.

The diagnostic model is based on a learner model that contains information about the learner’s normal pitch range. To get this information the learner is asked to sing a short musical fragment that covers her range. This data is used to calibrate the learner model.

The learner uses spoken responses at the second and third stages. At the second stage the learner’s response is compared to an expert’s performance but at the third stage it is compared to the corresponding tone model. In spite of this difference the processing procedure remains the same.

A learner’s spoken response is processed to find voiced fragments. The duration of learner’s speech is compared
to the duration of expert’s speech or to the model duration. If the duration of learner’s speech is shorter or longer than the duration of expert’s/model one by 30% then the corresponding error (“utterance is too short/long”) is triggered. The features are extracted from the signal and fed to the tone classifier. The result is compared to the required tone. If the learner’s utterance is classified as an incorrect one then the error “wrong tone” is triggered. Then, in spite of correctness of the learner’s utterance, the system does the detailed analysis. The utterance is divided into three equal parts (beginning, middle, and ending), and features are calculated for each part. Then a set of rules is applied to the features of each part to detect errors. The sets of rules represent expert knowledge in the problem domain and are mostly crafted and/or tuned manually. But to facilitate creating or tuning rules the following data mining techniques can be used:

- Clustering erroneous utterances for particular tone;
- Applying decision tree approach to derive rules;
- Creating recognizers for particular errors to expedite labeling typical cases and allowing experts to spend more time analyzing more complex cases.

Currently, the diagnostic model includes 41 rules that cover 23 typical errors. Some rules are simple. For example:

\[
\begin{align*}
\text{if } & (f0\text{beg}1 < \text{level5}\_\text{low}\_\text{limit}) \\
\text{and } & (f0\text{beg}2 < \text{level5}\_\text{low}\_\text{limit}) \\
\text{and } & (f0\text{beg}3 < \text{level5}\_\text{low}\_\text{limit}) \\
\text{then trigger_error ("low pitch for tone 1")}
\end{align*}
\]

This rule says that if the beginning pitch for each part is lower than the learner’s low boundary for pitch level 5 then the “low pitch for tone 1” error is triggered. When the detailed analysis is done, the system assembles the triggered errors and recommendations into a message. For example, “Your utterance cannot be recognized as tone 1 because your pitch is low. Try to start with higher pitch and maintain it evenly.”

### 4.3 Performance visualization

Besides providing a message to the learner the system visualizes the learner’s performance and allows comparing her performance to the expert’s one. The system plots the learner and expert pitch contours side by side. It also adds pitch levels to the graph. Figure 5 shows an example of learner performance visualization. The learner tried to pronounce Mandarin word /wa/ with tone 3. Here the solid line represents an expert’s and dashed line represents a student’s pitch contours. Both contours are aligned, scaled, and normalized. The learner can see easily her mistake – the starting pitch for the first part of utterance is too low.

### 5. Summary

The paper presents the methodology that allows creating diagnostic systems for students’ spoken responses in language learning environments. It uses signal processing and machine learning techniques to create quantitative and qualitative models for correct and typical wrong performances. This approach allows creating system with more intelligent feedback that can positively influence students’ motivation and their learning pace. Currently the methodology has been applied for two language learning tasks – learning sounds and learning Chinese tones. The paper describes in details the second task. A pilot experiment with five learners who recorded about 1,000 utterances portraying Mandarin tones (42% of utterances are erroneous) showed that the system diagnosed correctly 96% of utterances.

As some directions for future work I consider improving diagnostic models, creating tool for experts to expedite diagnostic model creation, and extending the current system for learning tones in multi-syllable words.

### 6. References


