Learning Chinese Tones

Valery A. Petrushin

Accenture Technology Labs, Accenture
161 N. Clark St., Chicago, IL 60601
petr@techlabs.accenture.com

Abstract

This paper is devoted to developing techniques for improving learning of foreign spoken languages. It presents a general framework for evaluating student's spoken response, which is based on collecting experimental data about experts' and novices' performance and applying machine learning and knowledge management techniques for deriving evaluation rules. The related speech analysis, visualization, and student response evaluation techniques are described. An experimental course for learning tones of Standard Chinese (Mandarin) 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, such as the IBM's ViaVoice, 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 low-level 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. 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 hint 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 it.

The rest of the paper has the following structure. First, I shall describe a general framework for applying speech analysis, machine learning and visualization techniques for spoken language learning. Then I shall demonstrate how the methodology works in an experimental course for learning Standard Chinese (Mandarin) tones.

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. Use images, animation or video clips to illustrate the gestures.
- Select acoustic features and create a quantitative model of the task [1]. For example, in an experiment for learning English 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 created 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's 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.

3. 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 nine tones in Cantonese [5].

Learning tones is very hard problem for the most of learners. Figure 1 shows profiles for four tones of Standard

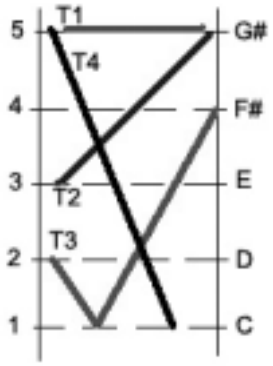

Figure 1. Mandarin Tones

Chinese (Mandarin) language [6]. A native Mandarin speaker distinguishes among five pitch levels, which correspond to the following music notes: C, D, E, F#, and G#. These values 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.

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 fell to 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

Category	Tone 1	Tone 2	Tone 3	Tone 4
Tone 1	**87.5**	9.5	0.5	2.5
Tone 2	2.5	**74.5**	22.5	0.5
Tone 3	0.5	19.5	**77.5**	2.5
Tone 4	2.5	2.5	0.5	**94.5**

Table 2: Accuracy of tone recognition for non-native speakers

Category	Tone 1	Tone 2	Tone 3	Tone 4
Tone 1	**70.8**	13.2	4.4	11.6
Tone 2	9.0	**60.4**	18.2	12.4
Tone 3	10.0	10.0	**59.6**	20.4
Tone 4	21.6	8.6	4.8	**65.0**

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 3 and 4 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 3. Tone recognition accuracy of for learner A.

Category	Tone 1	Tone 2	Tone 3	Tone 4
Tone 1	**72.0**	2.0	10.0	16.0
Tone 2	0.0	**48.0**	20.0	32.0
Tone 3	0.0	4.0	**66.0**	30.0
Tone 4	2.0	4.0	14.0	**80.0**

Table 4. Tone recognition accuracy of for learner B.

Category	Tone 1	Tone 2	Tone 3	Tone 4
Tone 1	**60.0**	22.0	10.0	8.0
Tone 2	28.0	**30.0**	16.0	26.0
Tone 3	18.0	4.0	**56.0**	22.0
Tone 4	22.0	16.0	8.0	**54.0**

3.1. 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 1 have been verified using the above-mentioned dataset.

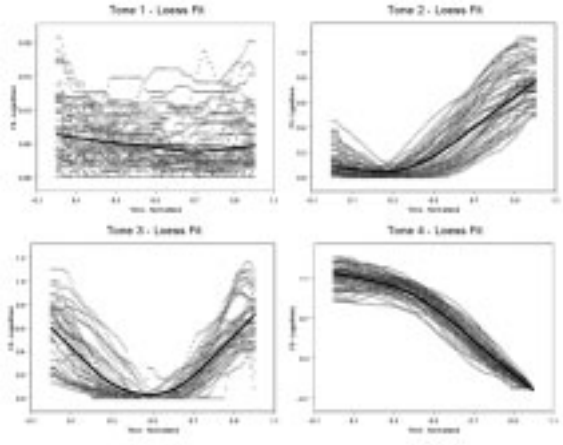

Figure 2. Experimental tone models.

The experimental profiles and their characteristics are depicted on Figure 2, which 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.

Table 5 shows statistics for syllable duration in milliseconds for different tones. You can see that the average duration for tone 3 is about three times longer than the duration for tone 4. Syllables of tone 1 and tone 2 have approximately the same duration, which is about 2/3 of average duration for tone 3. Figure 3 shows tone duration histograms.

Table 5. Syllable duration statistics for different tones.

Category	Mean	Median	s.d.	Max	Min
Tone 1	454.0	430	141.2	750	250
Tone 2	391.0	375	95.9	660	240
Tone 3	638.4	650	158.9	930	280
Tone 4	231.5	225	44.1	370	170

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.

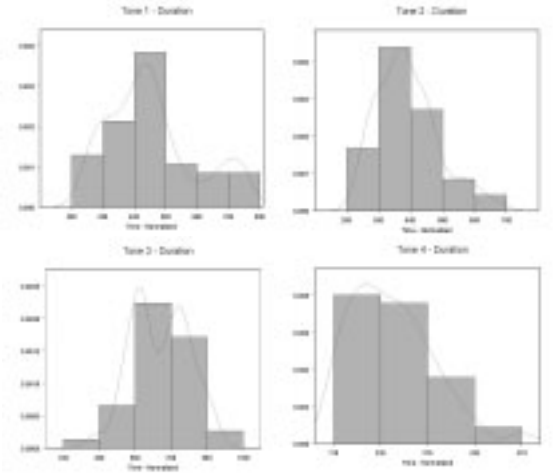

Figure 3. Tone duration histograms.

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.

3.2. 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 listen 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. The 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 or model speech 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 prototypes of the 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:

if (f0beg1 < level5_low_limit)
and (f0beg2 < level5_low_limit)
and (f0beg3 < level5_low_limit)
then trigger_error ("low pitch for tone 1")

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

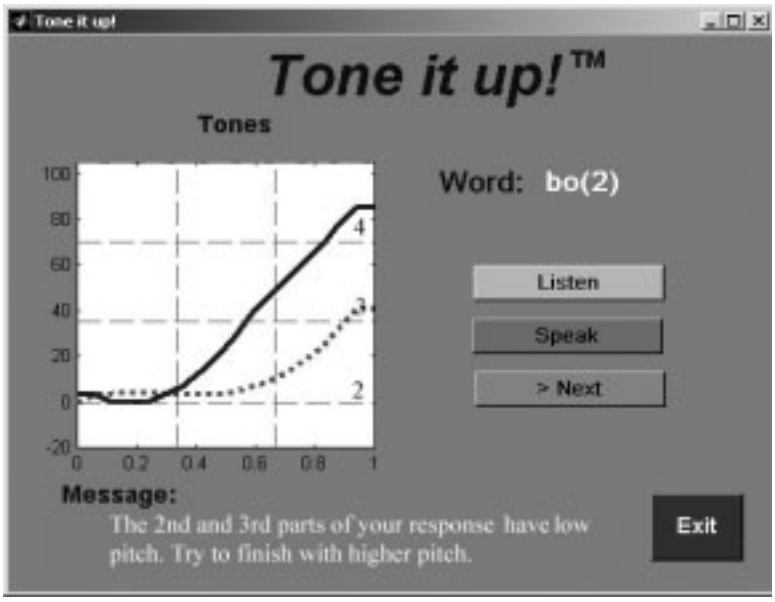

Figure 4. Performance visualization.

because your pitch is low. Try to start with higher pitch and maintain it evenly."

3.3. 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 4 shows an example of learner performance visualization. The learner tried to pronounce Mandarin word /bo/ with tone 2 ("thin"). 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 ending pitch for the utterance is not high enough.

4. Summary

The paper presents the methodology that allows creating diagnostic systems for student 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. Using this approach allows creating system with more intelligent feedback that can positively influence students' motivation and increase their learning pace.

Currently the methodology has been applied for two language learning tasks – learning English 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.

5. References

[1] L. Rabiner, and B.-H. Juang Fundamentals of speech recognition, Prentice Hall, Englewood Hills, NJ, 1993.

[2] J.M. Pickett, The acoustics of speech communication, Allyn and Bacon, Needham Heights, MA, 1999.

[3] TIMIT, http://www.ldc.upenn.edu/Catalog/LDC93S1.html

[4] Petrushin V., Using Speech Analysis Techniques for Language Learning, In T. Okamoto, R. Hartley, Kinshuk, J.P. Klus (eds), IEEE Int. Conf. On Advanced Learning Technologies 2001, Madison, WI, August 6-8, pp. 129-130.

[5] Lee, T. , Kochanski, G., Shih, Ch., and Li, Yu. Modeling Tones in Continuous Cantonese Speech. In. Proc. 7th International Conference on Spoken Language Processing (ICSLP 2002), pp. 2401-2404.

[6] Zongji Wu, "From traditional Chinese phonology to modern speech processing. Realization of tone and intonation in standard Chinese.", In Proc. 6th International Conference on Spoken Language Processing (ICSLP 2000), vol. 1, pp. B1-B12.