Towards Automatic Tone Correction in Non-native Mandarin

Mitchell Peabody and Stephanie Seneff
{mizhi,seneff}@csail.mit.edu *

Computer Science and Artificial Intelligence Laboratory,
MIT, Cambridge, MA 02139, USA

Abstract. Feedback is an important part of foreign language learning and *Computer Aided Language Learning* (CALL) systems. For pronunciation tutoring, one method to provide feedback is to provide examples of correct speech for the student to imitate. However, this may be frustrating if a student is unable to completely match the example speech. This research advances towards providing feedback using a student's own voice. Using the case of an American learning Mandarin Chinese, the differences between native and non-native pronunciations of Mandarin tone are highlighted, and a method for correcting tone errors is presented, which uses pitch transformation techniques to alter student tone productions while maintaining other voice characteristics.

1 Introduction

Feedback is essential in foreign language learning, and can take many forms, depending on the particular aspect of speech production being taught. A simple conversation involving teacher feedback is illustrated here:

1 Student "ni3 hao3! wo3 **jiu3** mi2 zhi4."
2 *Teacher* "*bu2 shi4* ***jiu3****. shi4* ***jiao4****.*"
3 Student "ni3 hao3! wo3 **jiao3** mi2 zhi4."
4 *Teacher* "*wo3* ***jiao4*** *mi2 zhi4.*"
5 Student "ni3 hao3! wo3 **jiao1** mi2 zhi4."
6 *Teacher* "***jiao4****. si4 sheng1.*"
7 Student "**jiao4**."
8 *Teacher* "*hao3.*"
9 Student "ni3 hao3! wo3 **jiao1** mi2 zhi4."

In this simple conversation, the student, whose primary language is American English, is attempting to say, "Hello! My name is Mitch," using Mandarin Chinese. The word pronounced incorrectly is presented in bold. In this case, there are two aspects of the pronunciation that are incorrect: the phonetic aspect and the tone aspect.

In order to correct the student's pronunciation, the teacher provides a correct template, in the form of their voice, for the student to imitate. The first correction

* This research was funded by the Industrial Technology Research Laboratory and by the Cambridge MIT Institute.

is to change "jiu" which has the wrong segmental form to "jiao." The second correction is to change the tone of the word from tone 3 to tone 4. The student is immediately able to correct the segmental part of his speech (changing "jiu" to "jiao"), however has trouble correcting the tonal aspect of his speech. In this example, the student unsuccessfully tries to correct his tone within a sentence. He then produces the tone correctly in isolation, but immediately fails to incorporate this change in a sentential context.

One problem is that the student does not know how his voice should sound and only has one reference to base his pronunciation on. The student's anxiety about correctly producing the language may be increased if he is unable to imitate the teacher's voice or identify what the teacher feels is lacking in his pronunciation [1, 2]. Because Mandarin is a tonal language, and the student's primary language, English, is not, the student may have trouble perceiving the distinctions between the tones [3, 4].

A major problem not illustrated by the example is that a teacher is not always available to give the student practice and guidance. However, a *Computer Aided Language Learning* (CALL) system can support practice at any time, in a non-threatening environment, and can provide feedback when a teacher is unavailable. CALL systems, which are designed to facilitate learning a foreign language using a computer, have three essential elements: speaking practice, hearing and understanding practice, and feedback.

A number of methods can be employed to provide the student with correct examples of speech. One method, largely employed by pronunciation dictionaries, is to pre-record native versions of speech being corrected. While this provides very high quality samples of speech for the student to emulate, it suffers from two major flaws. First, it is not scalable in that it is impossible to predict the full range of sentences that could be corrected. Second, the same problem that exists with the teacher is present: a student may still be unable to perceive and correct problems with their own speech. An alternative is to provide samples of speech using a speech synthesizer. This eliminates the scalability problem, but retains the problem of the student's inadequate perception of the error source. It also introduces the additional challenge of providing very high quality speech synthesis, which is very difficult.

If a step back is taken and the target language is considered, another method presents itself. Mandarin is a tonal language which means that tone quality and phonetic quality can be considered independently. Focusing on only the tonal aspect of Mandarin, we propose a method that modifies the *tonally* incorrect portions of student speech to sound correct. We wish to do this in a contextually sensitive manner for entire sentences by predicting an overall pitch contour based on native models. The advantages of this method are that a large database of pre-recorded speech is not required, a speech synthesizer is not utilized, the number of phrases that can be corrected is virtually unlimited, and the feedback is in the student's voice. Furthermore, by listening to two minimally different versions of their spoken utterance, students can tune in to the perceived differences, which pertain directly to tonal aspects.

In this research, a number of questions need to be addressed. What are some characteristics of natively produced tones? How are these different from those produced by non-native speakers? What, if any, variations occur with respect to sentence position? How can the corrections be realized? How can the tone of the speech be modified such that the result has few artifacts? How can the quality of the tones that are produced be tested?

Section 2 provides a brief overview of modern CALL systems that utilize dialogue interaction to allow for student practice. Section 3 gives a brief overview of native Mandarin tones and discusses differences between native and non-native productions of tones. Section 4 discusses our approach to correcting non-native tone errors. Section 5 presents some results and Section 6 summarizes the main points of this paper and provides directions for future work.

2 Background

A common approach to learning in a foreign language classroom is the task-based method. Task-based language learning is a communicative approach in which the student participates in a dialogue with another student, teacher, or native speaker on a particular topic with a specific end goal in mind [5, 6].

Feedback pertaining to various aspects of language learning is given either during or after the conversation. The idea is that, by encouraging the student to come up with sentences and phrases on their own, even if they are imperfect, learning will take place. Feedback may be given to correct major problems, but other problems are allowed to slide.

In recent years, CALL researchers have attempted to enable this form of foreign language learning using dialogue systems. In contrast to tapes or CDs, a computer system has the ability to dynamically create dialogues based on a given scenario for a student. By highly restricting the domain of a lesson to those one might find in a language book, it is feasible to construct dialogues that are dynamic in content and flow.

The *Tactical Language Tutoring System* (TLTS) [7, 8] immerses a student in a 3D world using the *Unreal Tournament 2003* [9] game engine, where he is instructed to accomplish missions by interacting with characters in the environment using speech and non-verbal communication. Speech recognition is done on highly restricted sets of sentences using the Cambridge *Hidden Markov Model Toolkit* (HTK) [10] augmented with noisy-channel models to capture mispronunciations associated with English speakers learning Arabic [11].

Raux and Eskenazi [12] adapted a spoken dialogue system [13] to handle non-native speech through adaptation techniques [14] using a generic task-based dialogue manager [15]. Another key feature of the system is the use of clarification statements to provide implicit feedback through emphasis of certain parts of a student's utterance [16] allowing feedback to be given as part of the dialogue.

Our general approach to CALL [17] is also modeled on the task-based approach. For all of our experiments, we have focused on the Mandarin/English language pair. A prototype system created by Lau [18] was able to carry on

short conversations with a student about simple topics such as family relationships. LanguageLand [19] was developed as a multi-modal system intended to help students learn to give directions in a foreign language.

It is within the task-based learning pedagogical framework that we wish to provide feedback. An overview of general pronunciation feedback in CALL can be found in [20]. Our focus here is on feedback as it pertains to pitch, for which a number of strategies have been previously attempted. An oscilloscope system [21] from the early 1960s provided direct visual feedback to the student through a real-time pitch display. A more recent example comes from [22], where a student received feedback in the form of a video game. A simple car driving game indicated to the student the quality of their feedback by how well the car remained in the center of a twisting and curving road.

Instead of explicitly indicating problems with pitch, which only hint at ways to correct the errors, some methods of feedback involve presenting the student with a corrected version of their own voice. For example, the *WinPitchLTL* [23] program provides visual feedback to the student in the form of pitch contours that can be compared against teacher provided models. The program has the additional capability of transforming the pitch of the student's speech to train on aspects such as intonation, stress, or tone. This functionality is obtained through a manual editing process. An automatic method was introduced in [24] where the prosody of isolated words was repaired using *Pitch Synchronous OverLap and Add* (PSOLA) [25–27]. Reference pronunciations were provided by recorded teacher utterances or by KTH's text-to-speech system [28]. Experiments in [29] generated a pitch contour for phrases using linear regression and ToBI [30] transcriptions. The generated contour was compared against a reference contour to show improvement. A similar technique was used in [31] where the authors attempted to repair intonation structure with a focus on stress patterns.

We propose repairing non-native tonal errors in a sentence by producing a model pitch contour based on native data. We examine Chinese tones to determine properties that can be incorporated into this model contour. We also examine differences in tone production between native and non-native speakers.

3 Tone Studies

In this section, we investigate speech data from three corpora: the Yinhe [32] corpus, the Lexical Tone (LT) [33] corpus, and the *Defense Language Institute* (DLI) corpus. The Yinhe data consists of 5,211 Mandarin utterances spoken by native speakers interacting with a dialogue system that provides information about flights and weather. The LT data consists of 497 Mandarin utterances, also in the weather domain, spoken by Americans in their first or second year of studying Mandarin at a college level. The DLI data consists of 5,213 utterances taken from oral proficiency interviews at DLI.

A tonal language uses pitch, the perception of fundamental frequency (f_0), to lexically distinguish tones. Mandarin Chinese is a tonal language in which every syllable is marked with a tone. Syllables in Chinese are composed of two parts: an initial and a final. The initial phone is either a consonant or the null initial

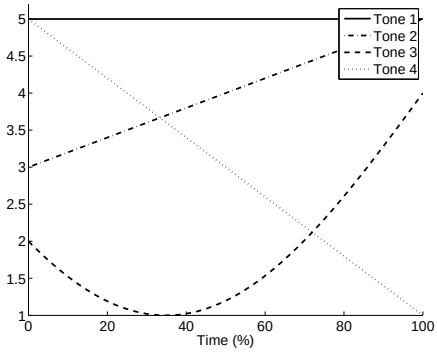
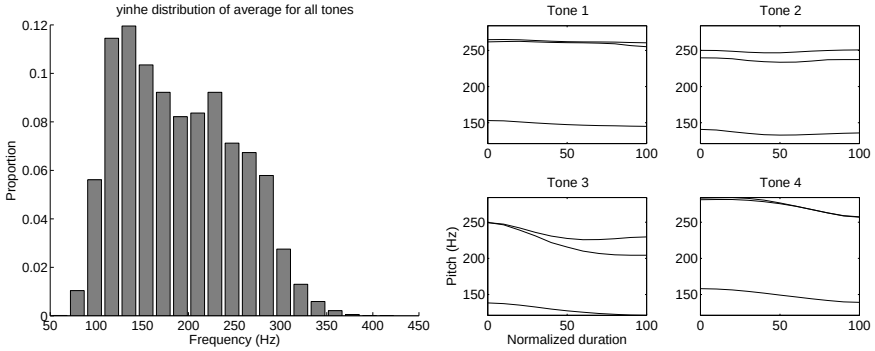


Fig. 1. Canonical forms of f_0 for tones produced in isolated syllables.



(a) Global f_0 distribution. (b) Raw contours.

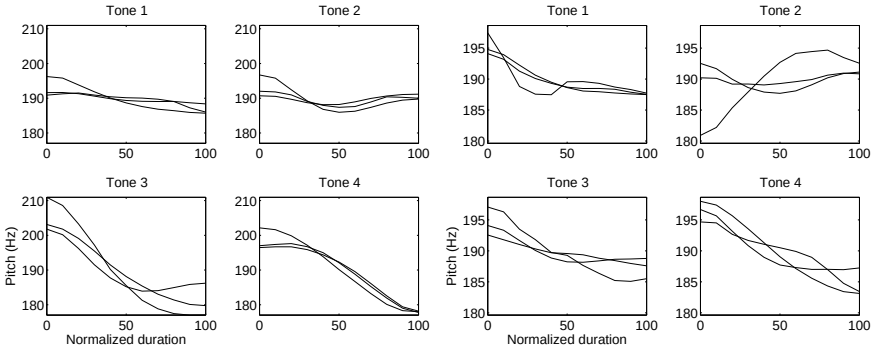
Fig. 2. Illustrating the need for normalization of f_0 contours.

(silence). The final portion of the syllable is composed of vowels and possibly a post-vocalic nasal. The final also functions as the tone bearing unit of a Chinese syllable [34]: the portion of the syllable where pitch differentiates tones.

Mandarin has 5 official tones, of which the first four are the most important for understanding. The fifth is often referred to as the neutral tone. Tonal languages lexically distinguish tones using pitch, or f_0 perception, in two main ways: by shape or by absolute height (register). Mandarin tones are mainly distinguished by shape, though there are other perceptual cues [35].

When pronounced in isolation, tones 1 through 4 have shapes that ideally look like those seen in Fig. 1 (tone 5 has no canonical shape, and is not shown). When pronounced as part of a word, phrase, or sentence, the pitch of the tones is altered in complex ways that depend on such factors as left and right contexts, anticipation [36], pitch declination [37], or tone sandhi rules [38].

In general, speakers of a non-tonal language who are learning Mandarin as a foreign language have difficulty both perceiving and producing tone (see, for example [4]). The major preceptual cue for distinguishing Mandarin tones is



(a) Native. (b) Non-native.

Fig. 3. Comparison of native vs non-native normalized f_0 contours.

pitch shape, which makes it a natural starting point for comparison between native and non-native speakers. In order to make meaningful comparisons of shape, the f_0 of the data must be normalized.

Fig. 2a is a histogram of the average f_0 for all voiced portions of speech over the entire Yinhe corpora. The bimodal distribution is due to gender differences in average f_0. Fig. 2b shows the f_0 contours for three randomly selected speakers from the Yinhe corpus: two female and one male. The obvious difference in the average f_0 of the male and females is one of the main reasons for normalization.

The normalization process has three main steps. First, an overall f_0 value is obtained for the entire Yinhe corpus. For each utterance in both the Yinhe and LT corpora, the f_0 values of each syllable are adjusted to be close to the utterance mean f_0. This step effectively removes tilt due to f_0 declination. Finally, each f_0 in the utterance is scaled by a constant factor to make the utterance mean f_0 equal to the corpus mean f_0 (189.75 Hz). This moves the mean utterance f_0 for each utterance to the same f_0 location as the overall mean for the Yinhe data.

After normalization, native and non-native pitch contours can be compared directly. Fig. 3a shows the f_0 contours for the same three speakers from the Yinhe corpus after normalization. The most important aspect to note is the consistency of the shapes for the native speakers, when contrasted with contours from three random non-native speakers from the LT corpus shown in Fig. 3b. It is evident that these non-native speakers have difficulty producing the contours correctly. For instance, there is very little contrast between the shapes of tone 1 and tone 3 for the non-native speakers.

It is also apparent from analysis of the Yinhe data that Mandarin tones differ in their mean f_0 values, although sentence declination effects must be accounted for to effectively exploit this feature. Averaged over all speakers, tones 1, 2, 3 and 4 have mean f_0 values of 203.73 Hz, 179.37 Hz, 178.05 Hz, and 196.3 Hz respectively. Thus, tones 1 and 4 have relatively greater mean values, and tones 2 and 3 have relatively smaller mean values.

Intonation, which encodes phrase level and sentential structure, interacts with tone production in complex ways that are not fully understood. It is known, however, that Mandarin pitch has a generally negative downward slope throughout phrases. This effect, known as pitch declination, must be accounted for explicitly in order to effectively exploit the intrinsic mean f_0 property of tones.

Quantitatively, we can define relative f_0 to be the ratio of f_0, averaged over the duration of the syllable final, over the mean f_0 of the sentence. Fig. 4a shows two histograms illustrating the distribution of relative f_0 for tones 3 and 4, for data over all syllable positions in the sentence. Fig. 4b shows the same plots, but restricted to syllable position 5. It is evident that the two distributions are much better separated when the data are restricted to a single syllable position.

If the relative f_0 of each tone is plotted as a function of syllable position, an interesting picture emerges. Fig. 5a plots the relative f_0 of native Mandarin speakers vs syllable position, and clearly shows that the separation between tones by relative f_0 persists throughout the duration of an utterance. This means that a pitch generating algorithm needs to adjust f_0 for declination based on both syllable position and tone assignment. Intonation generally plays a large role in the quality of pronunciation [39]. As with tone shape, non-native speakers have poor control over the interplay of tone relative f_0 and intonation, as illustrated by Fig. 5b.

4 Approach

Our general approach to providing tonal corrections in sentences is to modify a waveform of the student's speech. This is done in a two stage process. The first stage generates a pitch contour from native tone models. The second stage alters the pitch in the student waveform to match the generated contour.

In the first stage, we assume that an aligned transcription of the correct initials and finals for each syllable in a waveform of student speech is available. The pitch contour is extracted from the original speech using a dynamic programming algorithm described in [40].

For each syllable, the tone assignment for the final portion is determined from the transcription. A series of f_0 values in the shape of the tone is generated over the duration of the final. The f_0 values are adjusted to be appropriate for the current syllable position according to a declination model. For those time segments in which there is no final (and hence, no tone), the f_0 values are linearly interpolated to make the contour continuous.

Tone shapes are represented by four coefficients from the discrete Legendre transform as described in [41]. The model f_0 contour can be reconstructed from the first four coefficients. Parametrically characterizing the pitch contour of the tones has two benefits: pitch contours for different syllable durations can be easily generated, and less training data is required for each tone model.

The intonation declination models are linear equations derived from a regression on the first 10 syllables of the relative f_0 for each tone. For each tone, this gives a parametric model that can be used to adjust the f_0 values for a given tone at a given syllable position.

An example of a corrected utterance can be seen in Fig. 6. Normalization for speaker f_0 range and for sentence declination have been incorporated into the connected contour plot, and thus it has a very flat declination and correctly shaped tones.

In this example, there are very evident changes in the shapes of the tone contours. For example, in syllable position one, the syllable "luo4" is seen. This tone should have a falling pitch, but the speaker produced it with a rising pitch. The generation algorithm has produced a contour that is qualitatively closer to the correct native contour.

In the second stage, the pitch contour of the original speech and the generated pitch contour are both available. The speech in the student waveform is processed so that the pitch at each time interval is adjusted according to the generated version. The adjustments are done using a phase vocoder as described in [42, 43], which allows pitch to be adjusted up or down depending on a real-valued factor.

The advantage of using a phase vocoder is two-fold: it can produce very high quality pitch transformations and it can do these transformations by manipulating the waveform itself; no pitch extraction and resynthesis is required. In previous usage, the phase vocoder had adjusted the pitch of speech by a constant factor, but for this application, the pitch needs to be adjusted by a different factor for each time frame.

The end result of this algorithm is to produce a version of the student waveform that has been corrected for tone. The voice can still be identified as originally belonging to the speaker, but the tones will sound closer to native quality.

Dataset	# Utts	Original	Predicted
LT	497	41.3%	92.2%
DLI	5213	29.0%	81.3%

Table 1. Classification accuracy for original and generated contours.

5 Results

To evaluate the quality of the generated tone contours, we used a tone classifier trained on native data to classify tones from both original pitch contours and from the corresponding generated contours. The reasoning is that, if the generated pitch contour is closer to native quality, then the classification accuracy for a given utterance should be much better than for the original pitch contours. The choice of using a classifier to evaluate quality was motivated by the expectation that, in the future, this research will be incorporated into a larger CALL system that has automatic tone evaluation. We wished to establish that any corrective guidance regarding tone would be detectable by such a method.

The native training data used was the Yinhe data with normalized f_0. This normalization corrected for both syllable position and speaker pitch range. The feature vector used to train the classifier models was composed of the four Legendre coefficients that parameterize the tone shapes. Each tone model in the classifier was composed of 16 Gaussian mixture models.

Pitch contours were generated, as described in Section 4, for each utterance in the LT and DLI corpora, which contain non-native data. Table 1 shows the accuracy of the classifier on the original pitch contours and on the generated pitch contours. For both the LT and DLI corpora, there is a large increase in classification accuracy.

6 Summary and Future Work

This paper proposes a novel method for pronunciation feedback for learners of Mandarin by providing students with a corrected version of their own speech. An examination of native and non-native productions of tone revealed that non-native speakers have difficulty producing Mandarin tones. Based on native Mandarin speech, models for tones and phrase declination were built that were used to generate a pitch contour for a given utterance spoken by a non-native speaker. The results indicate that this generated pitch contour produces tones that are much closer to native quality than the original non-native speech.

In order to correct the student's speech we need to reintroduce the sentence declination into the generated pitch contour, and then apply the phase vocoder technique to instantiate it. Implementation of this portion of the algorithm is planned for the immediate future. While the generated contours were found to be much closer to native quality than the original contours; there is not yet any indication that there is correlation with human perception. The utterances produced by the phase vocoder need to be evaluated for improved tone quality through listening tests conducted by native speakers of Mandarin.

The tone models represented the lexical tones averaged over all left and right contexts; however contextual variations should be accounted for explicitly in the models. Modeling these contextual variations into account will also help capture prosodic phenomena such as tone sandhi rules. To do this will require more non-native data, as context specific models will experience data-sparseness issues.

This research dealt explicitly with feedback with the assumption that all tones were produced incorrectly by the non-native speakers. Most likely, though, only some of the tones will be produced incorrectly. In the near future, data will be marked by a fluent Mandarin speaker for tone quality. Based on feature comparisons between native and non-native speakers, methods for detecting which tones are produced incorrectly will be explored. This will allow for more selective feedback to be given.

7 Acknowledgement

We would like to thank Chao Wang for providing us with the pitch detector as well as significant technical assistance for this research.

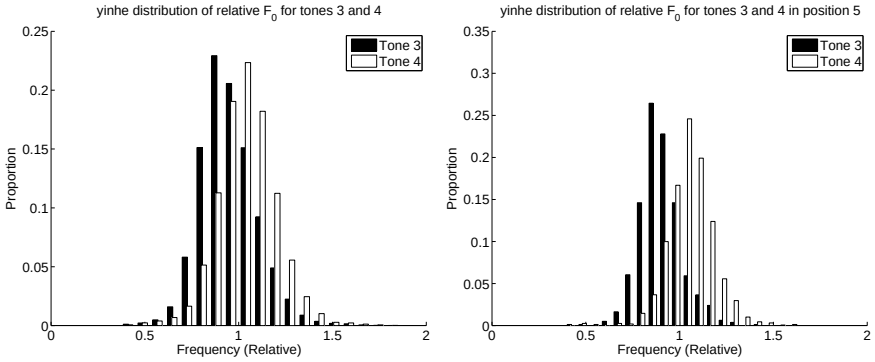
References

1. Horwitz, E.K., Horwitz, M.B., Cope, J.: Foreign language classroom anxiety. The Modern Language Journal **70**(2) (1986) 152–132

2. Onwuegbuzie, A.J., Bailey, P., Daley, C.E.: Factors associated with foreign language anxiety. Applied Psycholinguistics **20** (1999) 217–239
3. Kiriloff, C.: On the auditory discrimination of tones in mandarin. Phonetica **20** (1969) 63–67
4. Leather, J.: Perceptual and productive learning of chinese lexical tone by dutch and english speakers. In Leather, J., James, A., eds.: New Sounds 90, University of Amsterdam (1990) 72–97
5. Skehan, P.: Task-based instruction. Language Teaching **36**(01) (2003) 1–14
6. Ellis, R.: Task-based language learning and teaching. Oxford University Press, Oxford, UK (2003)
7. Johnson, L., Marsella, S., Mote, N., Viljhálmsson, H., Narayanan, S., Choi, S.: Tactical language training system: Supporting the rapid acquisition of foreign language and cultural skills. In: Proc. of InSTIL/ICALL Symposium: NLP and Speech Technologies in Advanced Language Learning Systems. (2004)
8. Johnson, L., Beal, C.R., Fowles-Winkler, A., Lauper, U., Marsella, S., Narayanan, S., Papachristou, D., Vilhjálmsson, H.: Tactical language training system: An interim report. In: Intelligent Tutoring Systems. (2004) 336–345
9. Epic Games, I.: Unreal tournament 2003. http://www.unrealtournament.com/ (2003)
10. Young, S., Odell, J., Ollason, D., Valtchev, V., Woodland, P.: The HTK Book. Cambridge University, Cambridge, UK (1997)
11. Mote, N., Johnson, L., Sethy, A., Silva, J., Narayanan, S.: Tactical language detection and modeling of learner speech errors: The case of arabic tactical language training for american english speakers. In: Proc. of InSTIL/ICALL Symposium: NLP and Speech Technologies in Advanced Language Learning Systems. (2004)
12. Raux, A., Eskenazi, M.: Using task-oriented spoken dialogue systems for language learning: Potential, practical applications and challenges. In: Proc. of InSTIL/ICALL Symposium: NLP and Speech Technologies in Advanced Language Learning Systems. (2004)
13. Raux, A., Langner, B., Eskenazi, M., Black, A.: Let's go: Improving spoken dialog systems for the elderly and non-natives. In: Eurospeech '03, Geneva, Switzerland (2003)
14. Raux, A., Eskenazi, M.: Non-native users in the let's go!! spoken dialogue system: Dealing with linguistic mismatch. In: HLT/NAACL 2004, Boston, MA (2004)
15. Bohus, D., Rudnicky, A.: Ravenclaw: Dialog management using hierarchical task decomposition and an expectation agenda. In: Eurospeech '03, Geneva, Switzerland (2003)
16. Raux, A., Black, A.: A unit selection approach to $f0$ modeling and its application to emphasis. In: ASRU 2003, St Thomas, US Virgin Islands (2003)
17. Seneff, S., Wang, C., Peabody, M., Zue, V.: Second language acquisition through human computer dialogue. In: Proceedings of ISCSLP. (2004)
18. Lau, T.L.J.: Slls: An online conversational spoken language learning system. Master's thesis, Massachusetts Institute of Technology (2003)
19. Lee, V.: Languageland: A multimodal conversational spoken language learning system. Master's thesis, Massachusetts Institute of Technology (2004) MEng.
20. Neri, A., Cucchiarini, C., Strik, H.: Feedback in computer assisted pronunciation training: technology push or demand pull? In: Proceedings of ICSLP, Denver, USA (2002) 1209–1212
21. Vardanian, R.M.: Teaching english through oscilloscope displays. Languate Learning **3**(4) (1964) 109–118

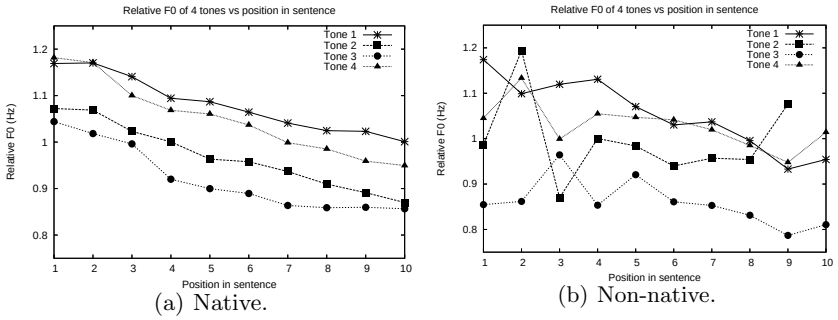
22. Álvarez, A., Martínez, R., Gómez, P., Domínguez, J.L.: A signal processing technique for speech visualization. In: STILL, ESCA, ESCA and Department of Speech, Music and Hearing KTH (1998)
23. Martin, P.: Winpitch ltl ii, a multimodel pronunciation software. In: Proc. of InSTIL/ICALL Symposium: NLP and Speech Technologies in Advanced Language Learning Systems. (2004)
24. Sundström, A.: Automatic prosody modification as a means for foreign language pronunciation training. In: STILL, ESCA, ESCA and Department of Speech, Music and Hearing KTH (1998) 49–52
25. Hamon, C., Moulines, E., Charpentier, F.: A diphone synthesis system based on time-domain prosodic modifications of speech. In: Proc. ICASSP '89, Glasgow, Scotland (1989) 238–241
26. Moulines, E., Charpentier, F.: Pitch synchronous waveform processing techniques for text-to-speech conversion using diphones. Speech Communication **9** (1990) 453–467
27. Moulines, E., Laroche, J.: Non-parametric techniques for pitch scaling and time-scale modification of speech. Speech Communication **16**(2) (1995) 175–207
28. Carlson, R., Granström, B., Hunnicutt, S.: Multilingual text-to-speech development and applications. In Ainsworth, A., ed.: Advances in speech, hearing and language processing. JAI Press, London (1990) 269–296
29. Black, A.W., Hunt, A.J.: Generating f0 contours from tobi labels using linear regression. In: Proceedings of the Fourth International Conference on Spoken Language Processing. Volume 3. (1996) 1385–1388
30. Silverman, K.E.A., Beckman, M., Pitrelli, J.F., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J.: Tobi: A standard for labeling english prosody. In: Proceedings of the 1992 International Conference on Spoken Language Processing. Volume 2., Banff, Canada (1992) 867–870
31. Jilka, M., Möhler, G.: Intonational foreign accent: Speech technology and foreign language testing. In: STILL, ESCA, ESCA and Department of Speech, Music and Hearing KTH (1998) 115–118
32. Wang, C., Glass, J.R., Meng, H., Polifroni, J., Seneff, S., Zue, V.: YINHE: A Mandarin Chinese version of the GALAXY system. In: Proc. EUROSPEECH'97, Rhodes, Greece (1997) 351–354
33. Peabody, M., Seneff, S., Wang, C.: Mandarin tone acquisition through typed interactions. In: Proc. of InSTIL/ICALL Symposium: NLP and Speech Technologies in Advanced Language Learning Systems. (2004)
34. Duanmu, S.: The Phonology of Standard Chinese. Oxford University Press (2002)
35. Whalen, D., Xu, Y.: Information for mandarin tones in the amplitude contour and in brief segments. Phonetica **49** (1992) 25–47
36. Xu, Y.: Contextual tonal variations in mandarin. Journal of Phonetics **25** (1997) 61–83
37. Shih, C.: Declination in mandarin. Prosody tutorial at 7th International Conference on Spoken Language Processiong (2002)
38. Chen, M.: Tone Sandhi: Patterns Across Chinese Dialects. Cambridge University Press, Cambridge, UK (2000)
39. Jilka, M.: The contribution of intonation to the perception of foreign accent. PhD thesis, University of Stuttgart (2000)
40. Wang, C., Seneff, S.: Robust pitch tracking for prosodic modeling in telephone speech. In: Proc. ICASSP, Istanbul, Turkey (2000) 887–890
41. Wang, C.: Prosodic Modeling for Improved Speech Recognition and Understanding. PhD thesis, Massachusetts Institute of Technology (2001)

42. Seneff, S.: System to independently modify excitation and/or spectrum of speech waveform without explicit pitch extraction. IEEE Trans. Acoustics, Speech and Signal Processing **ASSP-30**(4) (1982) 566
43. Tang, M., Wang, C., Seneff, S.: Voice transformations: From speech synthesis to mammalian vocalizations. In: Proc. Eurospeech 2001, Aalborg, Denmark (2001)



(a) Average over all positions. (b) At syllable position five.

Fig. 4. Separation of tones 3 and 4 by relative f_0.



(a) Native. (b) Non-native.

Fig. 5. Relative f_0 declination separated by tone.

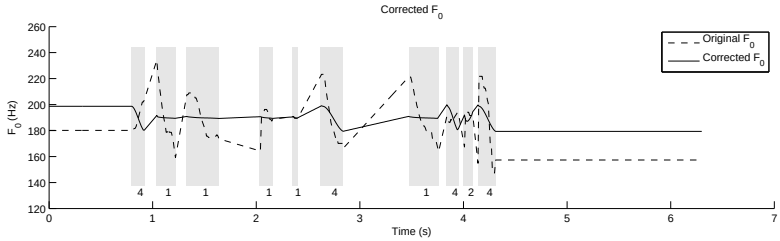


Fig. 6. Corrected contour for the sentence "luo4 shan1 ji1 xing1 qi1 si4 feng1 da4 bu2 da4" (*English: "Will it be windy in Los Angeles on Thursday?"*)