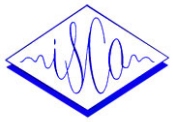

International Symposium on Chinese Spoken Language Processing (ISCSLP 2006)
Kent Ridge, Singapore
December 13-16, 2006

Automatic Scoring of Flat Tongue and Raised Tongue in Computer-assisted Mandarin Learning *

Bin Dong, Qingwei Zhao, and Yonghong Yan

ThinkIT Speech Lab
Institute of Acoustics, Chinese Academy of Sciences, Beijing
{bdong, qzhao, yyan}@hccl.ioa.ac.cn

Abstract. In this paper, a paradigm for the automatic scoring of flat tongue and raised tongue in computer-assisted Mandarin learning is presented. In this paradigm, distinctive features of Mandarin is used to distinguish the two classes consonants and score them. The confusion of them mainly results from different articulation places. The method based on the distributing of concentrated frequency area is proposed for scoring these consonants. In this method, recursive calculation of searching maximum energy value with sliding window is used to remove the high energy area of low frequency and the ratio of average energy is used as feature to evaluate flat-tongues and raised-tongues. With the method, the correct rate of scoring for the two classes consonants can reach 98.35%.

1 Introduction

With the rapid progress in communication and information technology and the development of market globalization during the past decades, the demand for foreign language teaching has increased steadily. At the same time, the teaching paradigm within language teaching has been shifted from traditional grammar teaching towards a teaching model focusing on the ability to communicate orally and pronunciation skill. The demand of each language learner can not be satisfied in the traditional classroom environment.

Computer-assisted language learning(CALL) can help language learners to communicate between computer and human being. In the past, computer-assisted language instruction could only support reading, listening and simple text input. But now users can get feedback from computer. The feedback is very important to language learners because they can get to know the quality of their pronunciations and through which they can better learn the language. This is especially important for beginners.

Nowadays, Chinese has become one of the most spokened language in the world and Chinese is one of the six official languages in United Nations. With the wave

* This work is supported by Chinese 973 program (2004CB318106), National Natural Science Foundation of China (10574140, 60535030).

of globalization, more and more countries pay attention to Chinese and more and more people begin to learn it.
Currently, the method of scoring pronunciation quality based on Hidden Markov Model(HMM) is widely used [1–3]. In this method, a set of context-independent models along with the HMM phone is used to compute an average posterior probability for each phone. The posterior probability is selected as the evaluation feature for scoring and the context-independent model is regarded as the correct pronunciation of each phone. The method with HMM can achieve a good performance generally[4–6], but still fail to distinguish some confusable phones accurately.
In Mandarin, flat tongue and raised tongue are one of the most confusable consonants. The confusion of them mainly results from different articulation places. In this paper, the method based on distinctive features is proposed to score the two classes consonants. On the confusion of them, the method based on the distributing of concentrated frequency area is proposed. In this method, recursive calculation of searching maximum energy value with sliding window is used to remove the high energy area of low frequency and the ratio of average energy is used as feature to evaluate flat-tongues and raised-tongues. With the method, the correct rate of scoring for the two classes consonants can reach 98.35%.
The rest of paper is organized as following. Section 2 describes the flat tongue, raised tongue and their confusions. Section 3 proposes the method for scoring the flat/raised tongue. Section 4 analyzes the experiment results, and conclusions and future directions are given in Section 5.

2 Confusion of Flat Tongue and Raised Tongue

For beginners, it's more difficult to learn the pronunciation of consonants because of their short pronunciation duration comparing with that of vowels and strict articulation manner and articulation places.
Generally, beginners used to imitate the pronunciation of second language with that of their mother language. When some pronunciations are not included in their mother language, beginners can not master them very well in second language. And if some pronunciations are not distinguished in their mother language, these pronunciations will be confusable in second language for them.
Flat tongue and raised tongue are two classes phones whose pronunciations can not be easily mastered in Mandarin. There are six phones which belong to flat tongue and raised tongue. They are [s], [sh], [c], [ch], [z], [zh]. In them, [s] and [sh] are fricative, the others are affricative.
In this paper, the pronunciations of 200 students from Guangzhou of China who took part in the PUTONGHUA SHUIPING CESHI(PSC examination) are analyzed in detail. In the wrong pronunciations, confusions occur between flat tongue and raised tongue very often. The details are shown in Table 1 and Table 2.

Table 1. Confusion of Raised Tongue

Confusion Rate	Flat Tongue	Affricative	Others
Raised Tongue(Fricative)	27.8%	63.3%	8.9%

Confusion Rate	Flat Tongue	Fricative	Others
Raised Tongue(Affricative)	18.9%	70.9%	10.2%

Table 2. Confusion of Flat Tongue

Confusion Rate	Raised Tongue	Affricative	Others
Flat Tongue(Fricative)	75%	12.5%	12.5%

Confusion Rate	Raised Tongue	Fricative	Others
Flat Tongue(Affricative)	69.7%	18.4%	11.9%

3 Scoring for Pronunciations of Flat Tongue and Raised Tongue

There are three groups of flat tongues and raised tongues. They are [s] and [sh], [c] and [ch], [z] and [zh]. As shown in Table 1 and Table 2, Confusions occur between flat tongues and raised tongues very often. In this section, the method of scoring them will be proposed. The figure. 1 is the block diagram of the scoring method.

3.1 Feature Selection

The articulation place is defined as the location of the constriction by the tongue in the oral tract. The location of the constriction by the tongue at the back, center, or front of the oral tract, as well as the teeth or lips, influences which fricative sound is produced. The constriction separates the oral tract into front and back cavities with the sound radiated from the front cavity. Although the front cavity dominates the spectral shaping of the sound, the back cavity introduces anti-resonances in the transfer function, absorbing energy at approximately its won resonances. Because the front cavity is shorter than the full oral cavity and because anti-resonances of the back cavity tend to be lower in frequency than the resonances of the front cavity, the resulting transfer function consists primarily of high-frequency resonances which change with the location of the constriction.[7] So different articulation place results in different resonant cavity and different resonant cavity results in different energy distributing. In the six flat/raised tongues, the articulation place of flat tongue is at the gingiva and the articulation place of raised tongue is at the front of hard palate. The different articulation

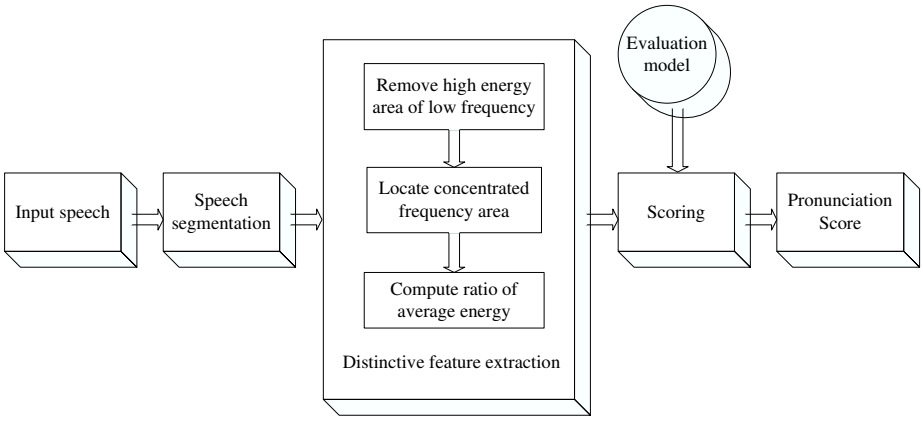

Fig. 1. Block diagram of scoring for pronunciation with different articulation places

place results in the different energy distributing. The energy curve of frequency domain is shown in Figure. 2 and Figure. 3. In the two figures, [s] is selected as example of flat tongue and [sh] is selected as example of raised tongue.

As shown in Figure. 2 and Figure. 3, concentrated frequency area is the obvious difference of the two classes consonants. The concentrated frequency area of flat tongue is at the high frequency section and that of raised tongue is at the mid-frequency area. So, the feature is selected as the distinctive feature to score the two kinds of consonants.

3.2 Feature Extraction

Concentrated frequency area is selected as the main feature to score flat/raised tongues with different articulation places. In this section, the feature extraction method which can describe the concentrated frequency area will be discussed in detail.

Removing high energy area of low frequency As shown in Figure. 2 and Figure. 3, concentrated frequency area is obvious to distinguish flat/raised tongues with different articulation places. But the curves in the two figure are not the whole curves. The high energy area of low frequency has been removed manually. The Figure. 4 shows the full curve corresponding to that in the Figure. 2. The high energy area of low frequency is generally introduced by recording environment and the effect of alternating current. From Figure. 4, we can see that the high energy area of low frequency influences the difference with concentrated frequency area.

The high energy area must be removed in order to achieve good performance of distinctive feature. In this paper, sliding window is used to search the energy maximum point in the high energy area and recursive calculate average energy

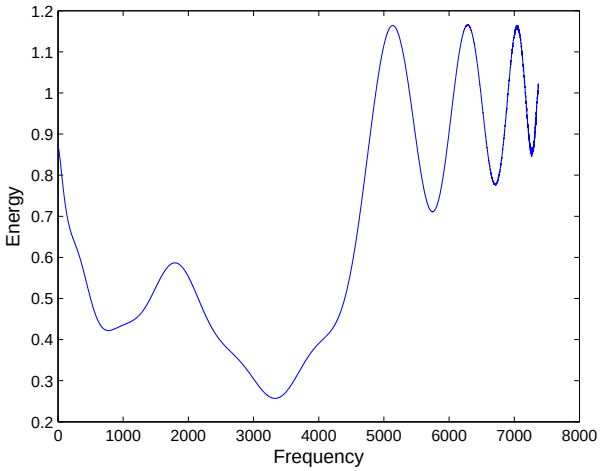

Fig. 2. Energy curve of flat tongue [s] in frequency domain

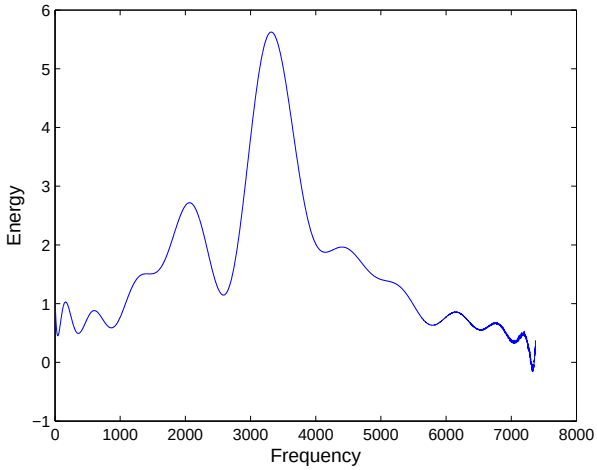

Fig. 3. Energy curve of raised tongue [sh] in frequency domain

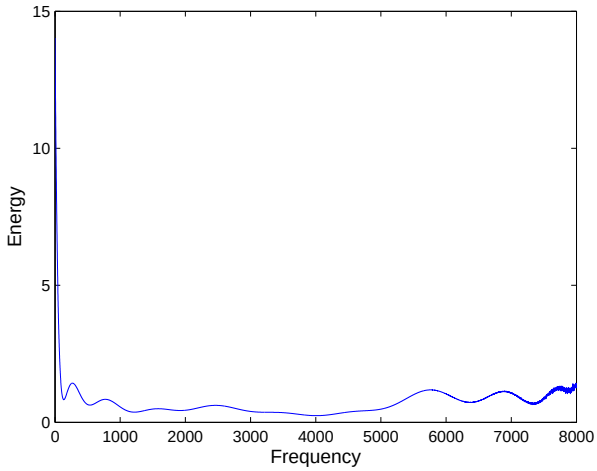

Fig. 4. Whole energy curve of flat tongue [s] in frequency domain

for searching the upper limit of the area.
Firstly, input speech is framed using a 25 ms window shifted every 10 ms, and processed by the hamming window. Amplitude spectrum is derived for fast Fourier transformation(FFT):

$$E(k) = \sum_{i=1}^{M} s(f_{ki}) \tag{1}$$

Where, M is the frame number of input speech, $s(f_{ki})$ is the spectrum energy on the each frequency point f_k of each frame.
Secondly, the maximum energy point can be found and sliding window is used in the process, as shown in Figure. 5.
Given, L is the sliding window size, W is the sliding step, the function of sliding windows is

$$w(k) = 1, 0 < k \leq L \tag{2}$$

The window is sliding in the pre-set range and is used to search the maximum energy value and corresponding frequency point k.

$$k = \arg\max E(k), k = nW + floor(L/2) + 1 \tag{3}$$

Where, n is the sliding number of window. Maximum energy value need on the condition:

$$E_{max} = E(k), \tag{4}$$

$$E(k) > \frac{\sum_{f_k=nw}^{floor(L/2)+nw} E(f_k)}{floor(L/2)} \tag{5}$$

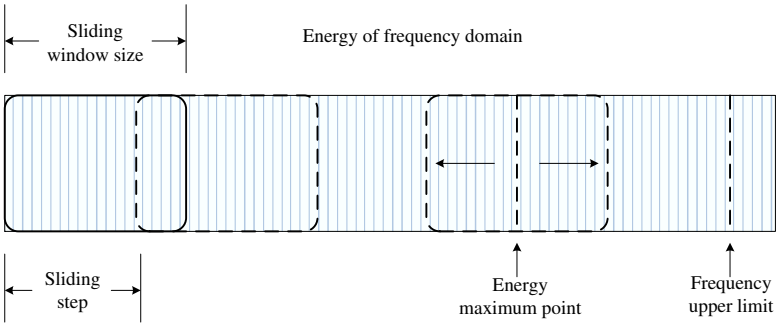

Fig. 5. Searching maximum energy point with sliding window

and

$$E(k) > \frac{\sum_{f_k=floor(L/2)+nw+1}^{L+nw} E(f_k)}{floor(L/2)} \tag{6}$$

In order to remove the effect of singularity value, maximum energy value $E(k)$ need satisfy further conditions:

$$E^\star = \frac{1}{2}(E(k-1) + E(k+1)) \tag{7}$$

$$E^\star > \frac{\sum_{f_k=nw}^{floor(L/2)+nw} E(f_k)}{floor(L/2)} \tag{8}$$

and

$$E^\star > \frac{\sum_{f_k=floor(L/2)+nw+1}^{L+nw} E(f_k)}{floor(L/2)} \tag{9}$$

Where, f_k is the frequency point.
Thirdly, alterable size window is used to search the upper limit of high energy area of low frequency, as shown in Figure. 6. Given E_0 is the initial threshold of energy, and

$$E_0 = \frac{1}{floor(L/2)} \sum_{m=k+1}^{k+floor(L/2)} E(m) \tag{10}$$

where, $k = \arg\max E(k)$.
With the largening of window size, the average energy is recursive computed as following:

$$E_i = \frac{E_{i-1} \cdot (floor(L/2) + (i-1)) + E_i}{floor(L/2) + i}, i = 0, 1, 2, \cdots \tag{11}$$

and

$$floor(L/2) + i < f_p \tag{12}$$

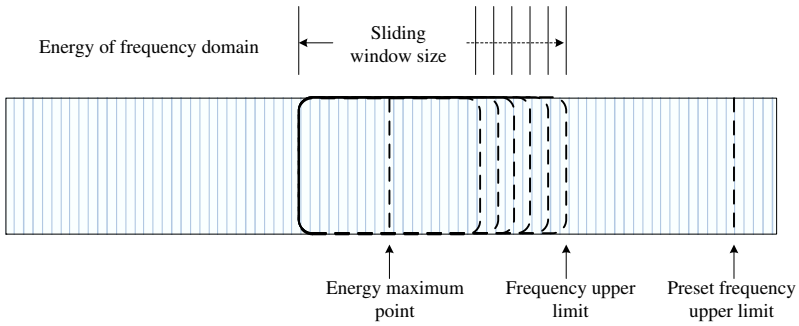

Fig. 6. Searching upper limit of high energy area with alterable size window

where, i is the frequency point, $E(i)$ is the spectrum energy of the ith frequency point, f_p is the pre-set upper limit of searching.
when

$$E_{i+1} > E_i \quad and \quad E_i > E_{i-1} \quad and \quad E_{i-1} > E_{i-2} \tag{13}$$

the frequency point i is regarded as the upper limit of high energy area of low frequency.

Feature extraction After removing the high energy area of low frequency, distinctive feature concentrated frequency area is displayed as shown in Figure. 2 and Figure. 3. So, extracting feature of concentrated frequency area is the crucial for scoring these consonants.
It can be found that concentrated frequency area is not located at a steady frequency band with stat. Locating concentrated frequency area is the next step for scoring.
a) Locating concentrated frequency area.
The searching range of concentrated frequency area can be pre-set according to the energy distributing of input speech because the transcript of input speech is known before hand. The principle of pre-setting searching range is enlarging. To flat tongue, the lower limit of frequency is set at 3500 Hz because its concentrated frequency area is at high frequency band. To raised tongue, the upper limit of frequency is set at 5500 Hz because its concentrated frequency area is at mid-frequency band.
Sliding window is used to search the first three maximum energy point k_1, k_2, k_3 in the pre-set range.

$$k = \arg\max E(k),\ k = k_0 + nW + floor(L/2) + 1 \tag{14}$$

where, k_0 is the initial point of searching range pre-set. And the first three maximum energy point need satisfy:

$$|k_1 - k_2| < f_{interval} \quad and \quad |k_2 - k_3| < f_{interval} \tag{15}$$

where, $f_{interval}$ is the interval between two maximum energy points.
As discussed above, the principle of pre-setting searching range is enlarging. The aim of enlarging is considering the instance shown in Figure. 7. It can be found by experiments that some wave crest of energy curve of flat tongue is near 3500 Hz which is at the mid-frequency area. If the initial lower limit of searching range is set too high, this kind of maximum energy point will be regarded as the part of non-concentrated frequency area. And it will influence the performance of the whole algorithm. So does raised tongue.
Considering the instance shown in Figure. 7, the pre-setting searching rang is enlarged. But to the normal instances of flat/raised tongue, the principle may introduce some pseudo maximum energy points. In order to resolve the problem, the initial window size is enlarged at the same time. With the larger window size, the satisfied maximum point may be not found. Alter window size is the method used in this paper.

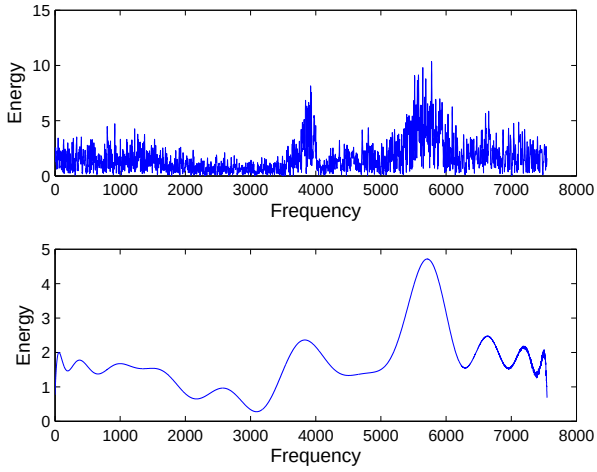

Fig. 7. Energy distributing of especial instance of flat tongue [s]

b) Computing the average energy of concentrated frequency area and it is defined as the average energy of sliding window which is at the maximum energy point found above.

$$E_H = \frac{1}{C(E_i)} \sum_{i=1}^{3} E_i \tag{16}$$

$$E_i = \frac{1}{L} \sum_{j=1}^{L} E(nW + k_0 + k_j) \tag{17}$$

$$C(E_i) = \begin{cases} 1\ E_i \neq 0,\ i = 1 \text{ or } i = 2 \text{ or } i = 3 \\ 2\ E_i = 0,\ i = 1 \text{ or } i = 2 \text{ or } i = 3 \\ 3\ E_i \neq 0,\ i = 1 \text{ or } i = 2 \text{ or } i = 3 \end{cases} \tag{18}$$

where, E_H is the average energy of concentrated frequency area. E_i is the average energy of sliding window which is the location of maximum energy point. $C(E_i)$ is a weight value which will alter according to E_i. L is the size of sliding window, n is the sliding number, W is sliding step of window, k_0 is the initial point of searching range pre-set.

c) computing average energy of non-concentrated frequency area

Given $f_{t_{ft}}$ is the lower limit of concentrated frequency area of flat tongue and $f_{p_{ft}}$ is the upper limit of non-concentrated frequency area of flat tongue.

$$f_{t_{ft}} = \arg\min E(k), \quad k = k_1, k_2, k_3 \tag{19}$$

$$f_{p_{ft}} = \begin{cases} f_{t_{ft}},\ f_{t_{ft}} - L < f_{k_{ft}} \\ f_{k_{ft}},\ f_{t_{ft}} - L > f_{k_{ft}} \end{cases} \tag{20}$$

where, $f_{k_{ft}}$ is the pre-set lower limit of concentrated frequency area of raised tongue. L is the sliding window size.

Given $f_{t_{rt}}$ is the upper limit of concentrated frequency area of flat tongue and $f_{p_{rt}}$ is the lower limit of non-concentrated frequency area of raised tongue.

$$f_{t_{rt}} = \arg\min E(k), \quad k = k_1, k_2, k_3 \tag{21}$$

$$f_{p_{rt}} = \begin{cases} f_{t_{rt}},\ f_{t_{rt}} - L > f_{k_{rt}} \\ f_{k_{rt}},\ f_{t_{rt}} - L < f_{k_{rt}} \end{cases} \tag{22}$$

where, $f_{k_{rt}}$ is the pre-set upper limit of concentrated frequency area of raised tongue. L is the sliding window size.

The average energy of non-concentrated frequency area of flat tongue is

$$E_{L_{ft}} = \frac{1}{f_{p_{ft}}} \sum_{k=1}^{f_{p_{ft}}} E(k) \tag{23}$$

The average energy of non-concentrated frequency area of raised tongue is

$$E_{L_{rt}} = \frac{1}{F - f_{p_{rt}} + 1} \sum_{k=f_{p_{rt}}}^{F} E(k) \tag{24}$$

where, $f_{p_{ft}}$ and $f_{p_{rt}}$ is the upper/lower limit of non-concentrated frequency area, F is the half of sampling frequency.

d) Computing the average energy ratio of concentrated frequency area and non-concentrated frequency area which can describe the distinctive feature energy distributing.

$$R = \frac{E_H}{E_L} \tag{25}$$

where, E_H and E_L is the average energy of concentrated frequency area and non-concentrated frequency area respectively.

3.3 Scoring method with Support Vector Machine

The average energy ratio is selected as the feature which can reflect concentrated frequency area of flat and raised tongue. In this paper, the pronunciation quality of each consonant is scored with feature in three levels: correct level, defective level and wrong level.
Support Vector Machine(SVM) is used to score each consonant. That is to say, the feature of consonants may be classified into three classes. For three-class, three classifiers can be constructed, one for each class. The nth classifier constructs a hyperplane between class n and the $n-1$ other classes. A majority vote across the classifiers can then be applied to classify a new point.

4 Experiments and results

In order to examine the proposed algorithm, a test database was prepared. The test database comprises of 300 utterances including all six flat tongues and raised tongues read by 50 speakers from Guangzhou of China. Five certified Chinese language testers rated the overall pronunciations on a scale of 0 to 2 corresponding to wrong level, defective level and correct level.
With the proposed algorithm described above, the pronunciation quality of test set is scored and the results are shown in Table 4. In the experiment, the performance of algorithm with removing high energy area of low frequency and without removing it is also compared and the results are shown in Table 3. At the same time, the method based on HMM is used to score the same test set and the results are also shown in Table 4.

Table 3. Correct rate of scoring on flat tongue and raised tongue

Consonant	Correct Rate of Scoring	
	no removing high energy area	removing high energy area
Flat Tongue	78.65%	98.52%
Raised Tongue	76.33%	98.05%
Average	77.49%	98.35%

It can be found that removing the high energy area of low frequency is necessary for the algorithm. And as shown in Table 4, the performance of algorithm proposed in this paper is better than that of method based on HMM on scoring flat tongue and raised tongue in Mandarin. The correct rate of classification with proposed method can reach 98.35%, while the correct rate of method based HMM is 82.25%.

Table 4. Correct rate of scoring on flat tongue and raised tongue

Consonant	Correct Rate of Scoring	
	HMM	Distinctive Feature
Flat Tongue	84.02%	98.52%
Raised Tongue	81.26%	98.05%
Average	82.25%	98.35%

5 Conclusion

An algorithm for automatic scoring consonants of flat tongue and raised tongue which have different articulation places and same articulation manner is presented. In this algorithm, the distinctive feature concentrated frequency area is selected as the feature and Support Vector Machine is used to score. The algorithm achieves a better result than HMM based method. However, different consonants may be scored with different distinctive features and no uniform frame has been proposed. This is our future work.

References

1. Neumeyer, L., Franco, H., Weintraub, M., Price, P.: Automatic text-independent pronunciation scoring of foreign language student speech. In: Proc of ICSLP. (1996)
2. Witt, S.M.: Use of Speech Recognition in Computer-assisted Language Learning. PhD thesis, University of Cambridge (1999)
3. Kim, Y., Franco, H., Neumeyer, L.: Automatic pronunciation scoring of specific phone segments for language instruction. In: Proc of EuroSpeech. (1997)
4. Franco, H., Neumeyer, L., Ramos, M., Bratt, H.: Automatic detection of phone-level mispronunciation for language learning. In: Proc of EuroSpeech. (1999)
5. Neumeyer, L., Franco, H., Digalakis, V., Weintraub, M.: Automatic scoring of pronunciation quality. In: Speech Communication. (2000)
6. Franco, H., Neumeyer, L., Digalakis, V., Ronen, O.: Combination of machine scores for automatic grading of pronunciation quality. In: Speech Communication. (2000)
7. Quatieri, T.F.: Discrete-Time Speech Signal Processing:Principles and Practice. Prentice Hall PTR (2002)