EuroSpeech'99, Vol. 1, pp.187-190, Budapest, Hungary, Sept. 1999

AN EFFECTIVE SCORING METHOD FOR SPEAKING SKILL EVALUATION SYSTEM

Zhanjiang Song, Fang Zheng, Mingxing Xu, Wenhu Wu

Speech Lab., Dept. of Computer Science & Technology,
Tsinghua University, Beijing 100084, P.R.China

szj@sp.cs.tsinghua.edu.cn

ABSTRACT

The Speaking Skill Evaluation (SSE) technologies are derived from speech recognition technologies and are used for language learning and instructing. In this paper, an effective automatic pronunciation scoring method for SSE systems is proposed. The Center-Distance Continuous Probability Model (CDCPM) is incorporated to model the speech. The Merging-Based Syllable Detection Automaton (MBSDA) and the Non-Linear Partition (NLP) method are used to perform the time alignment. And the Critical Area Percentage (CAP) based scoring method is used to score the learner's pronunciations or reject invalid utterances. Subjective assessments show that this method is concise, fast, and effective. The SSE system based on it has achieved a satisfying performance.

Keywords: Speaking Skill Evaluation, CDCPM, CAP, Automatic Pronunciation Scoring

1. INTRODUCTION

Intelligent Speaking Skill Evaluation (SSE) tech-nology is a brand-new research field derived from the conventional Continuous Speech Recognition (CSR), it is also an integration of computer technology and speech signal digital processing technology when they are applied to language learning.

SSE includes many key technologies, such as time alignment strategy and automatic scoring method for Speech Recognition Units (SRU), utterance verifi-cation, error locating and detecting, auditory based feedback, etc. And the progress made in the research of SSE will also be a great help to CSR.

In recent years, SSE has been studied and practiced. Some early SSE systems are text-dependent, while later ones are becoming text-independent [3], with utterance error locating ability [4][5] and auditory feedback [4]. Some practical automatic pronunciation scoring methods [2][3] have been put forward too.

In general, an intelligent SSE system should (1) provide standard utterances of instructive sentences and already-built basic SRU models; (2) evaluate the learner's pronunciations and map the scores to corresponding levels consistent with human subjective senses; (3) locate possible errors precisely in the learner's utterances, and (4) feed back the evaluation results and utterance-rectifying suggestions for the learner to correct his pronunciations.

According to above descriptions, we come to the conclusion that SSE technology is not just CSR, since there are some additional particular requirements. Conventional CSR focuses on the accurate mapping from utterances to proper SRUs, which is equivalent to a sequence of binary decisions whether a piece of utterance belongs to a certain SRU or not. However, the SSE systems additionally need to find out how the utterances are similar to or different from the SRUs they belong to, and then map the similarities to the levels consistent with human senses.

For the purpose of some multimedia applications aiming at Chinese language learning, we developed a Chinese based SSE system named CSSE [1], whose structure is illustrated in Figure 1.

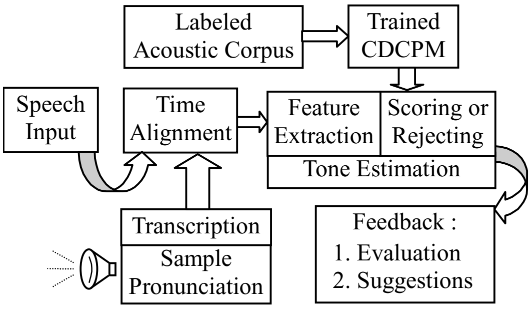

Figure 1: System structure of CSSE

In Section 2 the acoustic model and time alignment strategy are described. In Section 3 the novel scoring method based on CAP principle is described. In Section 4, some experimental results are given and some conclusions are drawn.

2. MODELS AND TIME ALIGNMENT

2.1 The CDN Distribution

Denote the probability density function (PDF) of a random variable ξ with a normal distribution by $N(x;\mu_x,\sigma_x)$, where μ_x is the mean value and σ_x is the standard deviation. Define a new random variable $\eta = |\xi - \mu_x|$, the PDF of which is

$$p(y;\sigma_x) = \frac{2}{\sqrt{2\pi}\sigma_x}\exp(-y^2/2\sigma_x^2), \quad (y \geq 0)$$

where the mean value μ_y of η can be calculated to be $\mu_y = 2\sigma_x/\sqrt{2\pi}$. In fact, η is the distance between the normal distributed variable ξ and its mean value μ_x, thus the defined distribution is referred to as a Center-Distance Normal (CDN) distribution. Define the (weighted Euclidean) distance between two scalars or two D-dimensional vectors as

$d(x_1,x_2) = |x_1 - x_2|$, and

$$d(\vec{x}_1,\vec{x}_2) = \sqrt{\sum_{d=1}^{D} w_d (x_{1d} - x_{2d})^2}$$

So, the pseudo-PDF of scalar-CDN can be

$$N_{CD}(x;\mu_x,\mu_y) = \frac{2}{\pi\mu_y}\exp(-d^2(x,\mu_x)/\pi\mu_y^2)$$

The case of D-dimensional vectors is similar to that of scalars. Denote the (weighted Euclidean) distance between a D-dimensional normal distributed vector $\vec{\xi}$ and its mean value vector $\vec{\mu}_x$ by another random variable η. Assume η is a CDN variable, then similarly, the pseudo-PDF of vector-CDN is

$$N_{CD}(\vec{x};\vec{\mu}_x,\mu_y) = \frac{2}{\pi\mu_y}\exp(-d^2(\vec{x},\vec{\mu}_x)/\pi\mu_y^2)$$

As a matter of fact, $N_{CD}(\vec{x};\vec{\mu}_x,\mu_y)$ is not the PDF of $\vec{\xi}$ but that of $d(\vec{\xi},\vec{\mu}_x)$, the distance between a normal distributed vector and its mean vector.

2.2 CDCPM, The Acoustic Model

Researches and experiments on the distance measure between models shown that the transition probability matrix plays a far less important role in HMM than the observation probability matrix does [6][7][8]. In view of the characteristics of Chinese language, a new acoustic model named Center-Distance Continuous Probability Model (CDCPM) [9][10][11] is derived from conventional HMM to model Chinese speech.

CDCPM is similar in topology to a left-to-right HMM without state skipping. The most prominent feature of CDCPM is that it discards the transition probability matrix of HMM and adopts the mixed CDN distribution instead of the normal distribution. The mixed density CDCPM we developed can be described by the following parameters: (1) N, the number of states per model. (2) M, the number of mixed densities per state. (3) D, the number of dimensions per feature vector. (4) $\vec{\mu}_{xnm} = (\mu_{xd}^{(nm)})$, the mean vector of the *m*'th density component in *n*'th state. (5) μ_{ynm}, the mean center-distance of the *m*'th density component in *n*'th state. (6) g_{nm}, the gain of *m*'th mixture density component. Here $1 \leq n \leq N$, $1 \leq m \leq M$, $1 \leq d \leq D$, and the observation PDF of state n has the similar form, which is called a mixed CDN density

$$b_n(\vec{c}) = \sum_{m=1}^{M} g_{nm} N_{CD}(\vec{c};\vec{\mu}_{xnm},\mu_{ynm})$$

where the subscript n means the *n*'th CDCPM state, and $\vec{c}$ denotes the speech frame feature vectors.

2.3 Training Corpus and Acoustic Features

The training data for CDCPM are taken from a giant Chinese speech database, uttered naturally by 76 people aged from 16 to 25 from all over the country, more than 500 sentences were read by each person. Those speakers consist of 38 males and 38 females. The speech is digitized for 16 bits per sample at 16 kHz sampling rate. The boundaries of all Chinese syllables in each sentence are pre-labeled manually.

We use LPCC and (weighted) auto-regression LPCC [12][13] as the acoustic features of mixed density CDCPM in CSSE, and take Chinese toneless syllables as SRUs to be modeled. Since the structure of CDCPM is very simple, it is easy to estimate all the model parameters of CDCPMs.

2.4 Time alignment of SRU states

In CDCPM, because the transition probability matrix has been discarded, the state transition strategy is quite different from that of HMM. To obtain time alignment of SRU states, a two-stage procedure is performed. First, a Merging-Based Syllable Detection Automaton (MBSDA) is used to determine syllable boundaries of the learner's utterances accurately. Second, a Non-Linear Partition (NLP) criterion is used to determine the state boundaries in each learner-uttered syllable.

MBSDA makes full use of speech parameters such as (differential) momentary frame energy, zero-crossing rate, pitches, and statistical knowledge of Chinese syllables (or initials and finals) and noises, etc. It agglomerates those neighboring frames having similar features to form Merged Similar Segments (MSS).

The MSSs will then be sent to a Syllable Detection Automaton (SDA) that contains some nodes of stable attributes. Figure 2 illustrates the state transitions in it.

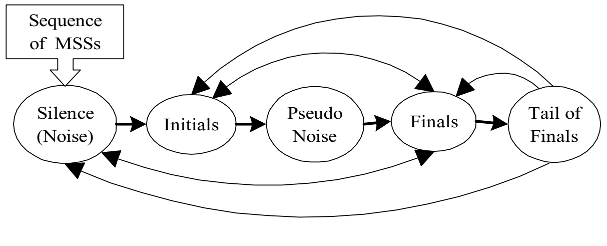

Figure 2: State Transitions in SDA

One of the most distinct differences between SSE and CSR is that the transcription to be uttered should be definite. That is to say, for a cooperative learner, he should imitate the instruction pronunciations of the SSE system, thus the content and the number of SRUs in his readings are fixed. With the aid of that extra information, the MBSDA can obtain the proper boundaries of each syllable without refined acoustic searching procedure.

Then, the NLP criterion is chosen to determine the separating points between inner states of each isolated syllable. Denote $\left\{\vec{O}_1,\vec{O}_2,\cdots,\vec{O}_T\right\}$ as the sequence of speech feature vectors of a certain SRU, where T is the length in frames. Define the difference between two neighboring vectors as $\Delta_t = \left\|\vec{O}_{t+1} - \vec{O}_t\right\|$, where $1 \le t \le T$. Assume the number of states per SRU is N, the average of total feature differences per state will be $\Delta_{state} = \frac{1}{N}\sum_{t=1}^{T}\Delta_t$. For $1 \le n \le N-1$, let L_n be the total number of feature vectors belonging to first n states. If there exists k such that

$$\sum_{t=1}^{k}\Delta_t < n \cdot \Delta_{state} \le \sum_{t=1}^{k+1}\Delta_t$$

the value of L_n is k. That is, L_n is the separating point between state n and state $n+1$. Apparently, L_N is equal to T. Thus, the procedure to find out the state boundaries can be convert to calculating the total number of vectors belonging to the first n states, the time alignment of each inner state of syllables is determined at the same time too.

In SSE applications, at least the number of syllables that a cooperative learner uttered should be equal to that of the transcription being followed, otherwise, the SSE system will reject his utterance and ask him to read the sentence again. For the former case, each syllable uttered by the learner is matched with the appropriate acoustic model according to its position in current transcription, evaluating it with the following scoring method, or rejecting it due to possible too large utterance distortions.

3. AUTOMATIC SCORING

Introduced in this section will be the novel automatic scoring method employed in our CSSE system.

Considering the normal distribution of $x \in (-\infty,\infty)$, $N(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-(x-\mu)^2/2\sigma^2)$, there exists approximately 95% of all samples falling into the area $[\mu-2\sigma,\mu+2\sigma]$, this kind of areas is defined as Critical Area (CA). Similarly, for each CDN distribution with the parameters of $\vec{\mu}_x$ and μ_y, there exists $\sigma = \sqrt{2\pi}\mu_y/2 \approx 1.25\mu_y$. Thus for a CDN distribution, the CA containing 95% of all samples is about $[0,2.5\mu_y]$. This kind of relation-ships between the speech feature vectors and the CAs of mixed CDN distributions in CDCPM are applied to evaluate the learner's utterances automatically. That is to say, this scoring method is primarily based on the Critical Area Percentage (CAP).

A threshold TH is chosen as the *Percentage* of CAP. According to this threshold, we can measure the proportion of the samples falling into the CA of $[0,TH\cdot\mu_y]$. Denote CDCPM parameters as

$$\Lambda = \left\{\vec{\mu}_{xnm},\mu_{ynm} \mid 1 \le n \le N, 1 \le m \le M\right\}$$

where n is the *n*'th state of the model, m is the *m*'th mixed CDN component in that state, N and M are defined as in Section 2.2.

Define the sequence of speech feature vectors within one learner-uttered syllable determined by the above time alignment strategy as $\mathbf{O} = \left\{\vec{O}_1,\vec{O}_2,\cdots,\vec{O}_T\right\}$. The CAP based score of the uttered syllable is defined as

$$\mathbf{S}(\mathbf{O} \mid \Lambda) = \sum_{t=1}^{T}\sum_{m=1}^{M} S(\vec{O}_t \mid n(t),m,\Lambda) \Big/ (T\cdot M)$$

where $n(t)$ is the state number that the feature vector at frame t belongs to within the current syllable. $n(t)$ is obtained directly through MBSDA and NLP. $S(\vec{O}_t \mid n,m,\Lambda)$ is the *partial* frame score of the *m*'th mixed CDN component within state n, defined as

$$S(\vec{O}_t \mid n,m,\Lambda) = \begin{cases} 1, & d(\vec{O}_t,\vec{\mu}_{xnm}) \in [0,TH\cdot\mu_{ynm}] \\ 0, & otherwse \end{cases}$$

Chinese is a toned language. The existing 4 tones are playing a very important role in the understanding of

Chinese sentences. However, one of the most important problems is that the Chinese sentences read by some foreigners or some dialectal people of China are in distorted tones often. So, the tones must be taken into consideration when developing a SSE system aiming at Chinese learning.

In our CSSE system, two most possible tones for each learner-uttered syllable are estimated by means of calculating the segmented slope of its F0 contour. If the two candidate tones of the syllable cannot cover the desired tone, a certain score will be deducted from the original CAP score as a penalty.

As mentioned in Section 1, it is also necessary to map the probabilistic scores to the levels consistent with human subjective senses. Furthermore, for some utterances having too large distortions compared to the anticipated pronunciations, the SSE system should refuse to evaluate their scores.

A simple linear function is taken here for the score mapping and utterance rejecting. Based on statistical results, an experimental threshold S_{REJ} is defined. If $\mathbf{S}(\mathbf{O} \mid \Lambda) \le S_{REJ}$, the uttered syllable is considered invalid and the system refuses to score it, otherwise, the final score $\mathbf{F}(\mathbf{O} \mid \Lambda)$ is calculated by

$$\mathbf{F}(\mathbf{O} \mid \Lambda) = \left(\mathbf{S}(\mathbf{O} \mid \Lambda) - S_{REJ}\right) / \left(1 - S_{REJ}\right)$$

It is obvious that

$$0 < \mathbf{F}(\mathbf{O} \mid \Lambda) \le 1 .$$

For some subtler SSE systems, more complicated mapping functions may be chosen to achieve better performance, such as the nonlinear mapping through a neural network [2].

4. CONCLUSION

We introduced a novel automatic scoring method for SSE systems. In general, it is hard to yield objective experimental results for such a kind of applications, whereas, subjective human assessments can be used to evaluate their performance. For our CSSE system, 315 Chinese sentences are chosen as instructive sentences, and 30 persons including some foreigners are invited to read the sentences and test it, then to assess its performance by comparing the output scores of each sentence with their own experiences and feelings.

According to their assessments, more than 76% of them thought the performance of CSSE is good enough for Chinese language learning and instructing, while the others suggested that there are still some important improvements need to be developed.

Since CSSE takes Chinese syllables as SRUs, by far the utterance distortions can be located at syllable level only. Modeling on initials and finals within syllables will help to give more exact feedback or suggestions on revising the learner’s pronunciations.

5. REFERENCES

[1] Wu, W.-H., Song, Z.-J., Wang, F. (1998), Research on the Strategies and Methodologies of Intelligent Speaking Skill Evaluation Systems, *Proceedings of Macau IT Congress*, pp. 174-177, Jan., 1998

[2] Franco, H., Neumeyer, L., Kim, Y., and Ronen, O. (1997), Automatic Pronunciation Scoring for Language Instruction, *Proceedings of ICASSP*, Vol. 2, pp.1471-1474, 1997

[3] Neumeyer, L., Franco, H., Weintraub, M., and Price, P. (1996), Automatic Text-Independent Pronunciation Scoring of Foreign Language Student Speech, *Proceedings of ICSLP*, Vol. 3, pp. 1457-1460, 1996

[4] Yoram, M., and Hirose, K. (1996), Language Training System Utilizing Speech Modification, *Proceedings of ICSLP*, Vol. 3, pp. 1449-1452, 1996

[5] Eskenazi, M. (1996), Detection of Foreign Speakers’ Pronunciation Errors for Second Language Training - Preliminary Results, *Proceedings of ICSLP*, Vol. 3, pp. 1465-1468, 1996

[6] Juang, B. H., and Rabiner, L. R. (1985), A Probabi-listic Distance Measure for Hidden Markov Models, *AT&T Technical Journal*, Vol. 64, No. 2, pp. 391-408, Feb., 1985

[7] Rabiner, L. R., and Juang B. H. (1986), An Introduction to Hidden Markov Models, *IEEE ASSP Magazine*, Vol. 3, No. 1, pp. 4-16, Jan., 1986

[8] Lee, K.-F. (1989), Automatic Speech Recognition – The Development of the SPHINX System, *Kluwer Academic Publishers*, Boston, 1989

[9] Zheng, F., Wu, W.-H., and Fang, D.-T. (1996), CDCPM and Its Applications in Speech Recog-nition, *Journal of Software*, Vol. 7, pp. 69-75, Oct., 1996

[10] Zheng, F., Chai, H.-X., Shi, Z.-J. et al. (1997), A Real-World Speech Recognition System Based on CDCPMs, *International Conference on Computer Processing of Oriental Languages, (ICCPOL’97)*, Vol. 1, pp. 204-207, Apr., 1997

[11] Zheng, F., Wu, W.-H., and Fang, D.-T. (1998), Center-Distance Continuous Probability Models and the Distance Measure, *Journal of Computer Science and Technology*, Vol. 13, No. 5, pp. 426-437, Sept., 1998

[12] Rabiner, L.R., and Schafer, R.W. (1978), Digital Processing of Speech Signals, *Prentice Hall Inc.*, 1978

[13] Furui, S. (1986), Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum, IEEE Trans. On ASSP, Vol. 34, No. 1, pp. 52-59, Feb., 1986