International Journal of Future Generation Communication and Networking
Vol. 9, No. 12 (2016), pp. 65-74
http://dx.doi.org/10.14257/ijfgcn.2016.9.12.06

Acoustic Model Fusion Method of British and American Compatible Mode

Wei He*

College of English, Zhongzhou University, Zhengzhou City, 450000, Hebei Province, China
E-mail: *hewei1202@hotmail.com*
Telphone: +8613939061358

Abstract

The British and American English are the most common target accents. British and American English have different phonetic symbol and pronunciation system. These two accents have many learners. And some learners have mixed phenomenon of British and American accent. Based on the mixed phenomenon British and American English accent for English learners, this paper put forwarder a model of American and British accent fusion method, improve the quality of the pronunciation evaluation performance system, and realize the embedded compressed acoustic model. This method divide acoustic model into alternative model, fusion model and encourage model by replace probability. The alternative model could be removed, and isolated model could be reserved. The fusion model could be merged based on model interpolation and model clip. Pronunciation quality evaluation results showed that the correlation of speaker level increased by 14.1%, compared with single accent model and in fusion model,; fusion model was similar to the performance of the hybrid model, the figure of gaussian component compressed by 10.7%.

Keywords: *Computer assisted language learning, Embedded application, Pronunciation quality evaluation, Model integration*

1. Introduction

The pronunciation evaluation aims to automatically evaluate the pronunciation quality of target language input by learners through machine automatic evaluation, and has a vast potential for future development. In recent years, embedded pronunciation quality evaluation system attracts more and more research and application for its portability and real-time features. In the embedded platform, the algorithm complexity is limited by the processor ability, storage capacity, and system power consumption and so on. The pronunciation quality evaluation based on the widely adopted hidden Markov Model (HMM) requires improvement and clipping on algorithm optimization, then could run on the embedded platform [1].

Accent pronunciation problem are the focus of pronunciation quality evaluation research, it mainly includes two aspects: target language accent and learners' native language accent. The objective of this article was that there were two target accents in English, namely British English and American English. These two accents had many learners. And some learners had mixed phenomenon of British accent and American accent. So, English pronunciation evaluation system shall support this two accents at the same time.

The solution for account problem for people were pronunciation modeling and acoustic modeling [2-3]. Pronunciation modeling structure constructed multiply accent dictionary or words-and-accent conversion model through regulation and data statistics

ISSN: 2233-7857 IJFGCN

method. Firstly, acoustic model was mainly based on model integration of some distance measure; secondly, acoustic model itself adapted to and overcome the accent problem with insufficient data; thirdly, acoustic model combined with pronunciation modeling and formed the joint modeling and adapted to themself [3].

This paper proposed an acoustic model fusion method compatible with British and American English, which was suitable for embedded English pronunciation quality evaluation system. It was different from the traditional accented pronunciation recognition task, the features of this article task were below. Firstly, both accent have relatively sufficient training data; Secondly, two accents have two not exactly same pronunciation system; Thirdly, the two accents were equally important, one accent could not be depressed based on the increasing of the other accent evaluation performance; Fourthly, it was the compromise method to realize embedded performance and complexity. The literature [7] proposed State-Dependent Phoneme-Based Model Merging (SDPBMM) significantly increased the dialectal accent recognition rate without decrease of mandarin pronunciation recognition rate. In view of the above characteristics, based on the acoustic distance, this paper divided this two accents acoustic model into alternative model, fusion model and isolation model. Based on SDPBMM thoughts, the fusion model could be merged and clipped.

2. British and American English Phonetic System

British and American English have different phonetic symbol and pronunciation system. This study had took American English CMUdit [4], (The CMU Pronunciation Dictionary) and British English BEEP [5] (The British English Example Pronunciation Dictionary) as examples. CMUdict had 39 phonemes, while BEEP has 44 phonemes, the extra 5 phonemes were respectively ax, ea, ia, oh and ua. For the same word spelling, the BEEP and CMUdic had phonetic difference as well, such as word CONTORT was read /kantaot/ in BEEP, while read /kahntaort/ in CMUdict, which existed phonetics change of /ax/ and /ah/, and American English /r/ phenomenon.

As mentioned above, the pronunciation evaluation system should support the two mainstream English accents of American English and British English. First of all, the pronunciation model should be compatible with British and American pronunciation model. Secondly, the acoustic model should be compatible with the American English and British English, considering the constrains of embedded system calculation and storage capacity, and fused the American acoustic model and British acoustic model.

3. The American and British Acoustic Model of Phonetics

The model fusion method was widely applied on pronunciation recognition task of various accents and non-native accent. Different from pronunciation recognition, pronunce evaluation could not tolerate various of non-target accent phonetics change, but to evaluate the difference of learner’s pronunciation accent and target accent (American English and British English).

The study proposed that acoustic model of the other target accent (English accent) could be divided into alternative model, fusion model and encourage model, based on one target accent (American accent). The acoustic distance was very little for alternative model and basic model. Almostly, there was no difference on recognition performance, if alternative model was replaced by basic model. Therefore, alternative model could be integrated with basic model. The acoustic distance of isolation model and basic model was bigger, and the isolated model recognition performance significantly lower after the replacement of basic model. So it should be remained in the study.

In order to achieve the above model integration ideas, the following two key issues shall be solved. (1) the definition and implementation of acoustic model distance, which was to distinguish the three kinds of model; (2) how to integrate basic model and

integration model. The following two sections would respectively discuss the two problems.

3.1. Acoustic Model Classification

In this paper, the method of process in acoustic model classification was shown in Figure 1.

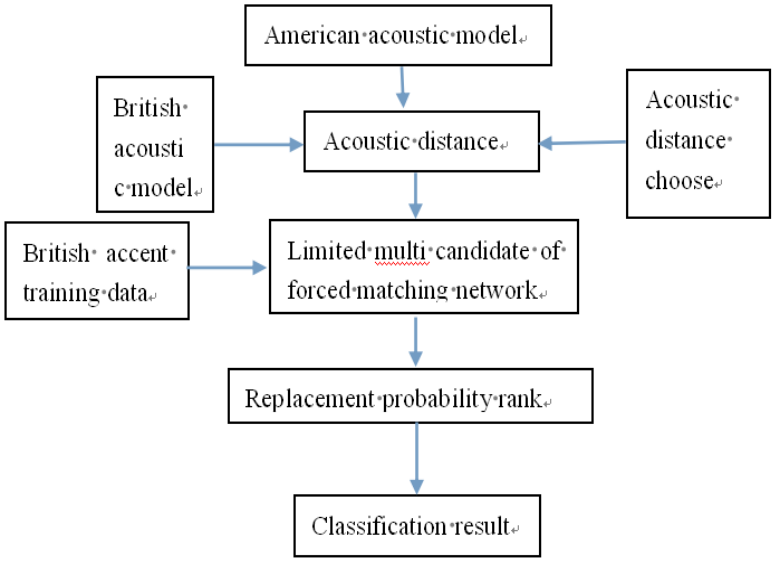

Figure 1. Implementation Method of the Acoustic Model

This paper generally set the basic model on the basis of the American accent model, and classified the British sound model. In order to overcome the shortcoming of high error recognition rate and poor matching effect of unlimited phoneme recognition network, this paper utilized the restricted multi-candidated mandatory matching method [3,6]. First, N piece of American accent models with nearest acoustic distance were selected accordingly to each British accent model, and named N candidate. Secondly, data was labeled under British Englishing training. The mandatory matching method constructed by N candidates compulsively match English account training set. Secondly, the substitution probability of N candidates American accent was statistically calculated for each British accent model, and output the highest probability by ranking. Finally, the highest probability was compared with two pre-setted threshold values. Isolated model was the British model below the lower threshold value. Alternative model was judged above upper threshold. Fusion model was the British model between the two thresholds.

The experiment results of the highest probability distribution interval of British model (monophone) were given in Table 1. Apart from “dh, b, d, l, th, t”, most of the replacement probability were higher than 60%, with higher fungibility. Apart from “oy, iy”, most vowels’ substitution probability were less than 60%, with lower fungibility. When the numbers of candidates was more than 7, the substitution probability reduced significantly. The number N was set as 7 in the later experiment.

Table 1. Distribution Interval of the Highest Substitution Probability for British Accent Model

Probability interval (%)	British accent model	Probability interval (%)	British accent model
0-25	-	25-30	ow
30-35	-	35-40	ia, oh, uh, ae
40-45	eh, dh, aa	45-50	aw, er, ua, b

50-55	ih, ax, d, l, th	55-60	ay, uw, ao, ey, t, ea
60-65	v, p	65-70	jh, g, n, ah, z, r
70-75	k, m, zh, w	75-80	ng, iy, ch, f
80-85	oy, hh	85-95	sh, y, s

3.2. The Choice of Acoustic Distance

From the acoustic model classification above, the choice of a good acoustic distance was crucial for the subsequent study. The common acoustic distance were Euclidean distance, Markov distance, Bhattachyaryya distance and divergence distance [3, 6]. Which acoustic distance could more precisely describe the acoustic differnce?

The acoustic distance between the two HMM with same topological structure could be defined as below formular 1:

$$dis\left(H_{1j}, H_{2j}\right)=\frac{1}{K}\sum_{i=1}^{K} dis\left(S_{1ji}, S_{2ji}\right) \tag{1}$$

In which, H_{1j} and H_{2j} respectively represented the j th phoneme model of two model set; S_{1ji} and S_{2ji} respectively represented the i th condition of j th phoneme model from two different model set; K represented the condition quantity of HMM.

In this article, the British model and American model mostly had two different phoneme systems, which could not directly utilize acoustic distance measure standard under the hypothesis of same phoneme system, for example, the proposed evaluation strategy based on phoneme candidates rank. Therefore, a new acoustic distance judgment formula for phoneme system inconsistencies was shown below formular 2:

$$\text{Score}=C_{eq}+\sum_{j=1}^{J}\frac{dis\left(H_{1j}, \breve{H}_{2j}\right)-dis\left(H_{1j}, H_{2j}\right)}{dis\left(H_{1j}, \breve{H}_{2j}\right)} \tag{2}$$

In which, H_{2j} represented phoneme model of the j th phoneme model of dictionary phonetic symbols H_{1j} in American accent model set; $\breve{H}_{2j}$ represented the phoneme model with nearest acoustic distance with H_{1j} in British acoustic model, apart from H_{2j}, namely the $\breve{H}_{2j}=\arg\min_{l\neq j} dis\left(H_{1j}, H_{2l}\right)$; C_{eq} represented the quantity of phoneme model with same phonetic symbol and smallest phoneme acoustic distance in two kinds of model set; J was the quantity of the British phoneme acoustic models. Obviously, the previous part of C_{eq} of formular 2 was integer, in larter part, the results of the fraction in summation symbol was less than 1, and was the increasing function of $dis\left(H_{1j}, \breve{H}_{2j}\right)$. The result of formular 2 was mainly decided by C_{eq}. if the C_{eq} of two acoustic distance was same, the bigger the $dis\left(H_{1j}, \breve{H}_{2j}\right)$, the higher score it was. In theory, each phoneme model was nearest to itself; a better distance measurement shall not only guarantee the above condition, but also should make a larger distance between each phoneme model and other phoneme models.

Table 2. The Score of Different Distance Measurement

Distance metric	Score	Distance metric	Score
Euclidean distance	47.67	Divergence distance of Euclidean	53.41
Markov distance	50.98	Divergence distance of Markov	54.10
Bhattachyaryya distance	54.49	Divergence distance of Bhattachyaryya	56.65

The scores of different distance metric was given in table 2. The combination of divergence distance of Bhattachyaryya and Dhattachyaryya had highest accuracy, which was used as acoustic distance metric. The results was same as the results of standard mandarin and Minnan accent mandarin model in literature [3]. Visibly, the relatively accurate degree of this distance metric described acoustic difference had certain data and to be popularized.

3.3. Acoustic Mode Fusion

First of all, referring to the model interpolation method in SDPBMM, many HMM with same topological structure was performed model merging. The merged model, each state contained the Gaussian component of all corresponding model condition which involved in merging. For example, model H_{2j} was the fusion objects, and were merged with M piece of models, the output probability density function of the i th condition of the merged model $H_{2j}^{'}$ shall meet:

$$p\left(o\left|S_{2ji}^{'}\right.\right)=\lambda p\left(o\left|S_{2ji}\right.\right)+\sum_{m=1}^{M}(1-\lambda)d\left(S_{1mi}\right)p\left(o\left|S_{1mi}\right.\right) \tag{3}$$

In this formula (3), o represented pronunciation feature vector, S_{1mi} represented the i th condition of m th model; $d\left(S_{1mi}\right)$ was merging regulation related undetermined parameter, met $\sum_{m=1}^{M}d\left(S_{1mi}\right)=1$; λ was the interpolation factor, which decided by experiment. This article also used the pronunciation changes of phoneme model to decide the $d\left(S_{1mi}\right)$ value in formula (3). Namely, for all condition of m th model, $d\left(S_{1mi}\right)$ was a fixed value $d\left(S_{1m}\right)$. The calculation of $d\left(S_{1m}\right)$ was the below formula (4).

$$d\left(S_{1m}\right)=\frac{P\left(H_{1m}\right)}{\sum_{m=1}^{M}P\left(H_{1m}\right)} \tag{4}$$

Among them, $P\left(H_{1m}\right)$ represented the substitution probability of model H_{2j} for model H_{1m} model involved in merging. Thus, the more model involved in merging, the higher of Gaussian component of merged model condition.

Furthermore, in order to control Gaussian component number expansion and reach the purpose of compression model scale, this study proposed a simple and effective model clipping method. The minimum confidence $\left(Miniconf\right)$ and maximum support number $\left(Maxnum\right)$ were imported to control the number M involved in merging number. British English phonemes H_{2j} were associated with N candidates of American English phonemes, the substitution probability was ranked from high to low, expressed as $H_{1j}^{1},H_{1j}^{2}\cdots H_{1j}^{N}$. When the substitution probability met $P\left(H_{1j}^{n-1}\right)-P\left(H_{1j}^{n}\right)>Miniconf$, the then $H_{1j}^{1},\cdots,H_{1j}^{n-1}$ were remained to involve merging. If the remained candidates number larger than $Maxnum$, then the $Maxnum$ pieces of candidates were remained. This candidates was adopted to reconstruct mandatory matching network to achieve new substitution probability. Based on the merged Gaussian components mode condition weight coefficient of merged model, the minimum confidence and maximum support number were adopted to restrict Gaussian components number. The fused acoustic model

was achieved.

4. Pronunciation Quality Evaluation Algorithm based on Prosterior Probability

This study adopted the prosterior probability pronunciation quality evaluation [1], as shown in formula (5).

$$p(H_t|o_t)=\frac{p(o_t|H_t)p(H_t)}{\sum_{j=1}^{J}p(o_t|H_j)p(H_j)} \tag{5}$$

In this formula, o_t represented the t th frame characteristic vector; the numerator at the right side of equation performed mandatory matching based on object text. H_t represented the t th frame acoustic model in mandatory matching pathway; based on the phoneme segmentation results of mandatory matching output, all phoneme models performed mandatory matching in a certain phoneme segment, the output was the denominator in the right side of equation. This background model calculation method had considered its lower complexity was suitable for embedded implementation.

5. Experiment Result and Analysis

5.1. Database

American model training adopted WSJ1 [7] database and CMU dict dictionary. British model training adopted WSJCAM0 [8] database and BEEP dictionary. Considering the embedded platform porting and previous experimental analysis, the single monophone of British Acoustic model and American Acoustic model were utilized, and each Gaussian component was 8 for each condition [1]. The development set of cross-identification of British English and American English were respectively clipped 1,000 phrases (the isolated words recognition task) from WSJ1 and WSJCAM0 test set [6]. Development set was used to determine the undetermined parameters in the algorithm.

Pronunciation quality evaluation test set was from the site acquired 401 people's voice library from the simulation spoken language test organized by College English Test 4/6 (CET-4/6) committee. It can be abbreviated as CET401. Each people read 10 sentence English without referring to text. Each sentence pronunciation was subjectively evaluated by committee authorized English professor: 1 for good, 0.5 for ok, 0 for bad. The summary of 10 sentence scores was the speaker's score.

5.2. Preferences

The method and main parameters of British English and American English mentioned in this article were below. First, the substitution probability threshold value for model classification, which were abbreviated as Th_{low} and Th_{high}; secondly, interpolation factor λ was used to merge models; thirdly, it were used for the minimum confidence and maximum support number for clipping models.

For simplicity, $Th_{low}=Th_{high}=Th$ was proposed firstly, there was no existence of fusion models, the effect of substitution probability threshold value Th was discussed on development set error rate and model size. The experiment results were shown in table 3.

Table 3. Threshold Value Th Effect on Development set Error Rate and Model Size

Th	Model size	WER (%)		
		Dev_AE	Dev_BE	Avg.
20%	43	12.7	16	14.6
30%	44	12.6	15.6	14.3
40%	49	12.7	14	13.5
50%	60	12.3	13.7	13.1
60%	67	13.3	10.6	11.8
70%	77	13.8	8.7	11.0
75%	77	13.8	7.6	10.3
89%	83	13.7	7.2	10.1
90%	83	13.4	7.1	9.9
95%	87	13.4	6.7	9.7

Note: model size was phoneme model number, WER: Word Error Rate, Dev_AE represented American Development set.

Dev_BR represented British English deveopment set, Avg. Represented average word error rate.

From table 3, with the increase of Th, model size increased gradually, the word error rate of American development set slightly increased, the word error rate of British development set decreased, and the average word error rate decreased as well. When Th was 20%, all British Acoustic model were substitution models, when Th was 95%, all British Acoustic models were isolated models. When Th changed between 20% and 95%, the word error rate floated a lot in two range, which were 70% ~ 75% and 40% ~ 50%. Therefore, this study choose Th_{low} was 40%, Th_{high} was 75%. The fusion model was the substitution between 40% ~ 75%.

The model fusion method in this study aimed to compress model size, so maximum support number of preset substitution probability was 2, the maximum support number of Gaussian component was 16; in order to confirm the coincidence of HMM condition topological structure, experiment did not take minimum confidence parameters. The model size of that time were below, phoneme size number was 50, Gaussian component number was 1864. The interpolation factor effect on word error rate of British model and American model were shown in Figure 2. When λ was 0, all fusion models were British Acoustic model, when λ was 1, all fusion models were American models. When λ increased to 1 from 0, the word error rate of Dev_AE decreased, the word error rate of Dev_BE increased; when λ was 0 and λ was 1, three word error rate floated a lot; when λ was between 0.3 and 0.8, three word error rate nearly remained the same. Therefore, when one accent model was totally substituted by another accent model, the word error rate could apparently increase. In addition, word error rate was not sensitive to the change of λ. Finally, this study choose λ equal to 0.5.

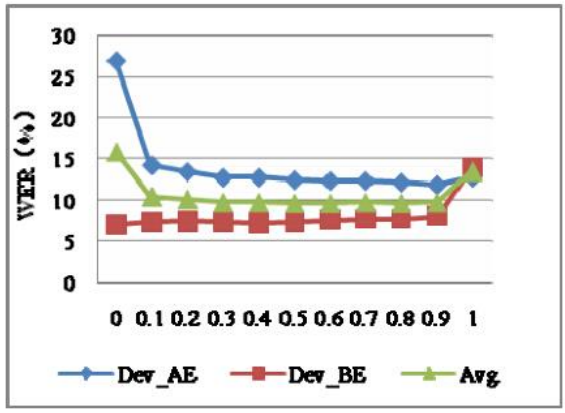

Figure 2. Topological Structure Effect on Word Error Rate of British English Development set and American English Development Set

5.3. Pronunciation Quality Evaluation Experiment

Based on CET401 database, the performance of British Acoustic model, American Acoustic models, mixed model of British English and American English, and fusion model of British English and American English was illustrated in table 4.

Table 4. Fusion of British English and American English Effect on Model Size and Pronunciation Quality Evaluation Performance

Model feature	Model size		Pronunciation quality evaluation performance	
	Phoneme number	Gaussian component number	Sentence relevancy	Speaker relevancy
British acoustic model	46	1104	0.613	0.713
American acoustic model	41	984	0.612	0.725
Mixed British and American acoustic model	87	2088	0.682	0.821
Fusion of British and American acoustic model	50	1864	0.685	0.827

Note: sentence relevancy represented normalized cross coefficient between scores from machine evaluation and subjective evaluation in sentence level; speaker relevancy represented the normalized cross coefficient between scores from machine evaluation and subjective evaluation in speaker level.

From table 4, the mixed British English and British English could obviously increase the correlation between machine evaluation and subjective evaluation in sentence level and speaker level. Fusion model of British English and American English and Mixed model of British English and American English was quite pronunciation quality

evaluation performance. While, compared to mixed model, the phoneme number of fusion model compressed 42.5%, Gaussian component number compressed 10.7%. This was due to mixed model only considered pronunciation model (dictionary), but fusion model considered both adjustment of pronunciation model and acoustic model. Fusion model had lower model complexity, suitable for embedded implementation. The proposed British and American English fusion model was achieved on UniSpeech platform [1].

6. Conclusion

In this paper, the study was based on the American acoustic model. The substitution probability from mutiply candidate identification restrictions was divided into substitution model, fusion model and isolated model. Then the fusion model was merged through model interpolation method, and the merged model was revised further. Finally, the final fusion model of British acoustic model and American acoustic model. The fusion model greatly improved the performance of pronunciation quality evaluation algorithm. Compared with the single accent model, the correlation of speaker increased by 14.1%; similar like mixed model had smaller model size and more suitable for the embedded application, Gaussian component compressed by 10.7%.

The experience was performed on the monophone model, the further study could be the contextual related model function on pronunciation quality evaluation, and contextual model extension of American and British accent fusion method.

References

[1] W. Q. Liang, "Objective Assessment of Pronunciation Quality Based on Hidden Markov Models", Tsinghua University, Beijing, **(2006)**.
[2] M. Y. Tsai, F. C. Chou and L. S. Lee, "Pronunciation Modeling with Reduced Confusion for Mandarin Chinese Using a Three-Stage Framework", IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 2, **(2007)**, pp. 661-675.
[3] L. Q. Liu, "Research on A Small Data Set Based Acoustic Modeling for Dialectal Chinese Speech Recognition", Tsinghua University, Beijing, **(2007)**.
[4] K. Lenzo, "The CMU Pronouncing Dictionary", http://www.speech.cs.cmu.edu/cgi-bin/cmudict.
[5] Robinson Tony, BEEP dictionary, http://svr-www.eng.cam.ac.uk/comp.speech/Section1/Lexical/beep.html, **(1997)**.
[6] K. Zhao, "Study on Algorithms for Speech Recognition Engine of English Learning Machine", Tsinghua University, Beijing, **(2008)**.
[7] November 1993 ARPA Continuous Speech Recognition Hub and Spoke Benchmark Tests Corpora and Instructions, http://www.ldc.upenn.edu/Catalog/readme_files/csr2/csrnov93.html, **(1994)**.
[8] F. Jeroen, P. Dave, R. Tony, W. Phil and Y. Steve, WSJCAM0 Corpus and Recording Description", http://svr-www.eng.cam.ac.uk/~ajr/wsjcam0/wsjcam0.html, 2 September 1994.

