SUBSPACE GAUSSIAN MIXTURE MODEL FOR COMPUTER-ASSISTED LANGUAGE LEARNING

Rong Tong, Boon Pang Lim, Nancy F. Chen, Bin Ma and Haizhou Li

Institute for Infocomm Research, Singapore

ABSTRACT

In computer-assisted language learning (CALL), speech data from non-native speakers are usually insufficient for acoustic modeling. Subspace Gaussian Mixture Models (SGMM) have been effective in training automatic speech recognition (ASR) systems with limited amounts of training data. Therefore, in this work, we propose to use SGMM to improve the fluency assessment performance. In particular, the contributions of this work are: (i) The proposed SGMM acoustic model trained with native data outperforms the MMI-GMM/HMM baseline by 25% relative, (ii) when incorporating a small amount of non-native training data, the SGMM acoustic model further improves the performance of fluency assessment by 47% relative.

Index Terms— Computer Assisted Language Learning (CALL), Subspace Gaussian Mixture Model (SGMM), Automatic Speech Recognition (ASR), Goodness Of Pronunciation (GOP), Fluency assessment

1. INTRODUCTION

Acquiring second or third languages is getting more popular with the trend of globalization. This demand is outpacing the availability of human language experts and teachers, thus driving an urgent need for computer-assisted language learning (CALL). CALL systems provide an easy interface for language learners. Unlike traditional language learning scenarios, CALL allows learners to practice on their own in private at their own pace. Even though automatic and self-learning options might not be effective as engaging learning with human teachers, CALL systems that emulate human assessment and scoring can save a lot of manual labor and time [1].

A CALL system usually provides feedback of the speech spoken of the second language (L2) learner. This feedback can be categorized into two levels: (1) Segmental level, which focuses on the pronunciation accuracy of the individual phonetic units [2, 3]. (2) Suprasegmental level, which focuses on the rhythm, stress, and intonation of the non-native speech [4, 5]. Unlike most studies in CALL, which only focus on either one of these two aspects, in this work, we attempt to model the fluency level of non-native speakers using both aspects even though it is a more challenging task.

The underlying mechanism for generating such language learning feedback is usually implemented as an automatic fluency assessment system. Automatic fluency assessment of non-native speech takes advantage of techniques from related fields such as language identification [6] and automatic speech recognition (ASR) [7]. For example, the goodness of pronunciation (GOP) [8] score derived from the posterior probability at the phonetic level is often used to quantify the segmental aspect of pronunciation quality. On the other hand, fluency assessment at the suprasegmental level could be modeled by prosodic features, such as the speed of articulation and frequency of pauses [4, 5, 9]. Therefore, the characterization power of the acoustic model plays an essential role in automatic fluency assessment.

Compared to native speech, higher variations are observed in non-native speech. These variations are influenced by various factors [10] like speaker's native language(s) and the amount of exposure to the target language. To build superior acoustic models that characterize those variations, some researchers employ discriminative training technique to enhance the acoustic models [3, 11], while other researchers train deep neural networks [12, 13, 14] to refine the acoustic models. Various machine learning methods are explored to model the variations in non-native speech [15, 16, 17]. When incorporating non-native data to train the acoustic model, one challenge is the amount of available non-native training data is often limited. To compensate for the lack of non-native data, speaker adaptation techniques are often applied [11, 18] by using a small amount of non-native data.

In this work, we utilize Subspace Gaussian Mixture Models (SGMM) [19] to characterize the acoustic properties of non-native speech. The speaker information derived from SGMM models have provided complementary information for automatic language recognition [20]. SGMMs have also shown great advantage in multi-lingual speech recognition [21] and non-native speech recognition [22]. The success of SGMM can be attributed to its compact parameter sharing mechanism, which is especially effective when training data is limited. Given the effectiveness of prior work adopting SGMM, we expect SGMM to be suitable in fluency assessment in CALL applications. In this paper we empirically show that SGMM outperforms a MMI-GMM/HMM baseline in assessing fluency scores of non-native speakers when only

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

using native Mandarin training data. In addition, further improvements are shown when adding a small amount of non-native data to the training set.

2. SUBSPACE GAUSSIAN MIXTURE MODEL FOR AUTOMATIC FLUENCY ASSESSMENT

2.1. Subspace Gaussian Mixture Model

The Subspace Gaussian Mixture Models can be formulated as follows:

$$p(\mathbf{x}|j) = \sum_{i=1}^{I} \omega_{ji} N(\mathbf{x}; \mu_{ji}, \Sigma_i) \quad (1)$$

$$\mu_{ji} = \boldsymbol{M}_i \nu_j \quad (2)$$

$$\omega_{ji} = exp(\omega_i^T \nu_j)/(\sum_{k=1}^{I} exp(\omega_k^T \nu_j)) \quad (3)$$

where $p(\mathbf{x}|j)$ is the probability model for state j, I is the number of shared mixtures, and ν_j is the state-specific vector. Σ_i is a full covariance matrix for i-th Gaussian; it is shared globally among all the states for this Gaussian, μ_{ji} and ω_{ji} are mean and mixture weights for states j respectively. Unlike the conventional Gaussian mixture model training, the state mean and mixture weights are not trained directly, they are derived from globally shared mean matrix M_i and weight vector ω_i using a low dimensional vector ν_j.

As the full covariance matrix is shared across states, an SGMM model has fewer parameters than the conventional GMM model. This means we can use less training data to achieve comparable models without losing accuracy. It is very useful in language learning task, as we are always facing the problem of insufficient foreign accented data.

2.2. Automatic fluency assessment

2.2.1. Fluency assessment features

Even though fluency is more often associated with the suprasegmental aspect of speech, it is usually an implicit prerequisite that a fluent speaker is able to achieve segmental competency (i.e., to pronounce words in an articulate manner). Therefore, in this work fluency is modeled from both segmental and suprasegmental aspects (see Table 1).

The goodness of pronunciation (GOP) [8] and its statistics are used to characterize the segmental level of non-native speech. GOP is a phone level confidence measure to gauge how differently a particular phone is pronounced compared with a native model. Given phone p, the GOP score is:

$$GOP(p) = \frac{1}{n} \frac{P(O|p)P(p)}{\max_{q \in Q} P(O|q)p(q)} \quad (4)$$

where O is the acoustic observation, Q is the set of all phones, n is the number of frames. $P(O|p)$ stands for the likelihood of the observation on model p, it can be obtained by performing forced alignment with the canonical transcription. $\max_{q \in Q} P(O|q)$ is the maximum likelihood of all the phones in the phonetic inventory, often derived from a phone-loop decoding process.

Prosodic features are important in modeling the suprasegmental level of fluency [4]. For example, a fluent L2 speaker is more likely to have a higher speaking rate and less hesitation. In this work, we characterize prosody using features derived from ASR results. Specifically, they are rate of speech, phonation ratio, articulation rate and mean pause length (see Table 1 for more details.)

2.2.2. Performance measurement

Correlation score is a widely adopted performance measurement for fluency assessment. Let x presents a set of fluency labels given by human raters and y denotes system derived fluency labels, their correlation score is estimated as:

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (5)$$

where $\bar{x}$ and $\bar{y}$ are the means of x and y, n is the number of test. A high correlation score indicates the system has good performance in fluency prediction.

3. CORPUS AND EXPERIMENTAL SETUP

3.1. Native Mandarin corpus

The Chinese Mandarin Speech Recognition Database (King-ASR-214) [1] is used for the acoustic model training. The corpus contains speech collected over 4 different mobile operating systems: iOS, Android, Windows Mobile and Symbian. The speech data are recorded in 16 KHz. There are 1000 gender balanced speakers with 444.6 hours of speech in total.

3.2. Non-native corpus: iCALL

The non-native speech corpus used in this study is iCALL corpus [23]. In this corpus, 300 beginning learners of Mandarin Chinese were asked to read 300 Pinyin prompts, including 200 words (each consists of 2-4 characters) and 100 sentences. Each speaker received a different set of utterances. The speech was sampled at 16 kHz and recorded in quiet office rooms. The speech data are transcribed in Pinyin through perceptual listening tasks. These transcriptions represent the surface pronunciation – mistakes made by the speaker were transcribed as is – while the Pinyin prompts are served as the underlying canonical pronunciation.

3.2.1. Perceptual fluency scoring guidelines

For each speaker, one third of the utterances are examined by three human judges, who rated the fluency level on the scale of 1 to 4 (where 4 indicates the most fluent).

[1] http://www.speechocean.com/en-ASR-Corpora/839.html

Feature type	Specific feature	Description
segmental	GOP	GOP score for each phone in phone inventory
	AVE INIT GOP	Average GOP score for initial phones
	AVE FINAL GOP	Average GOP score for final phones
suprasegmental	ROS	Rate of speech, number of phones per frame
	PTR	Phonation/time ratio, total duration of speech w/o pause/total duration of speech
	ART	Articulation rate, number of phones per seconds excluding pause
	MLP	Mean length of pauses

Table 1: Features for fluency assessment

A first set of scoring guidelines were established through literature survey and consulting experienced human judges. The scoring guidelines were further refined through pilot tests. Discrepancies among scores rated by the human judges in the pilot test were resolved through discussion, and used to refine the scoring guidelines. The number of pronunciation errors in each utterance was the main criteria for scoring. Phonetic errors were penalized more heavily than tonal ones, as phonetic errors affect intelligibility more.

3.2.2. Consistency tests for perceptual fluency scores

After establishing the scoring guidelines, another consistency test was conducted to ensure the scores from the human judges were properly calibrated. A set of 30 unique utterances were randomly selected for performing a consistency test. These 30 utterances were duplicated (i.e. total 60 utterances) and presented randomly to each human rater. The correlation coefficients for intra-rater scores and those for inter-rater scores are shown in Table 1(a). The Cohen's Kappa coefficients are reported in Table 1(b) to quantify the inter-rater agreement. Consistency tests are conducted periodically to ensure the scores from the human judges maintain such desired quality.

	R1	R2	R3
R1	1.00	0.89	0.90
R2	0.89	0.95	0.93
R3	0.90	0.93	0.95

(a) Correlation

	Kappa
R1 R2	0.641
R2 R3	0.709
R1 R3	0.698

(b) Kappa

Table 2: Inter-rater and intra-rater correlation and Kappa

Germanic	Train	Test	Roman	Train	Test	Slavic	Train	Test	Others	Train	Test
UK	24	6	Italy	13	4	Russia	19	6	Unknown	27	9
USA	37	12	France	22	7	Ukraine	2	0	Hungary	2	0
Canada	13	3	Mexico	3	1	Croatia	2	0	Greece	3	1
Australia	3	1	Spain	8	2	Bulgaria	1	0	Georgia	3	1
Germany	8	2	Argenti	1	0	Poland	3	1	Finland	1	0
Sweden	3	1	Brazil	6	1	Belarus	2	0			
Switzerlan	1	0									
Ireland	2	0									
Norway	3	1									
	94	26		53	15		29	7		36	11

Fig. 1: Number of speakers in iCALL train and test sets

As inter-rater consistency is high, we proceeded to assign each rater a distinct set of utterances to score in order to save time. This human-rated data set is split into a training subset and a test subset of distinct speakers. Gender and utterance length are balanced across the subsets. Figure 1 shows the number of speakers in the iCALL training and test set, the speakers are grouped into 4 broad categories according to their country of origin [23]. There are a total of 14800 and 5810 utterances in the training and test sets, respectively.

4. AUTOMATIC FLUENCY ASSESSMENT

4.1. Experimental Setup

4.1.1. Baseline MMI-GMM/HMM ASR Systems

A baseline ASR system was trained from native Mandarin Chinese (as described in Section 3.1). The corpus was split into training and test sets. The training set has 326000 utterances from 975 speakers, while the test set (King-test) consists of speech data from 25 speakers, with 1980 utterances in total. The acoustic feature consists of 13 dimensional MFCC feature, 1 dimensional tone feature, and their derived deltas, acceleration and third-order deltas, resulting in 56 dimension. The acoustic model consists of 175 phones and 8534 tied states.

The baseline acoustic model is a GMM-HMM model discriminatively trained with Maximum Mutual Information (MMI) criterion, this model is denoted as *MMI native*. Another MMI model (*MMI native+iCALL*) is trained from both King-ASR-214 and iCALL training set using the same parameter set up as *MMI native*.

4.1.2. SGMM ASR systems

An SGMM model (*SGMM native*) was trained on top of the MMI native model with the same set of native training data. It is configured with 800 shared mixture components and 6081 states. To learn the pronunciation variability of foreign speaker's, another SGMM model is built on top of the MMI native model by using the iCALL training set and the native training set. It uses the same configuration as the *SGMM native* model, we refer this model as *SGMM native+iCALL*. Note that the surface pronunciation (derived by raters) of the iCALL training set is used in the above mentioned acoustic model training process.

Table 3 reports the ASR performance of the acoustic models on the native test set (King) and non-native test set (iCALL). The King test set is decoded with a 5-gram language model while the iCALL test set is decoded with a Pinyin loop. The results show that the SGMM models outperform their MMI counterparts on both native and non-native test set. The accuracy of the non-native set is improved by incorporating non-native training data while the accuracy of

Test set	MMI native	MMI native+iCALL	SGMM native	SGMM native+iCALL
King	30.8	38.5	29.0	48.0
iCALL	66.6	64.3	65.6	51.3

Table 3: Performance of acoustic models, King test set in character error rate and iCALL test set in phone error rate

the native test set is degraded. Intuitively, a larger pronunciation variance can be observed among non-native speakers due to different nationality and language group. The acoustic models trained from those non-native speech are adapted to capture those pronunciation characteristics which are not observed in native speech.

4.2. Fluency assessment results

A gender-dependent support vector machine (SVM) classifier is built for each fluency level following the one-vs-rest criterion. Hence there are 2 models for each of the 4 fluency levels (one female and one male). Each test utterance is evaluated on a set of gender-matched SVM models.

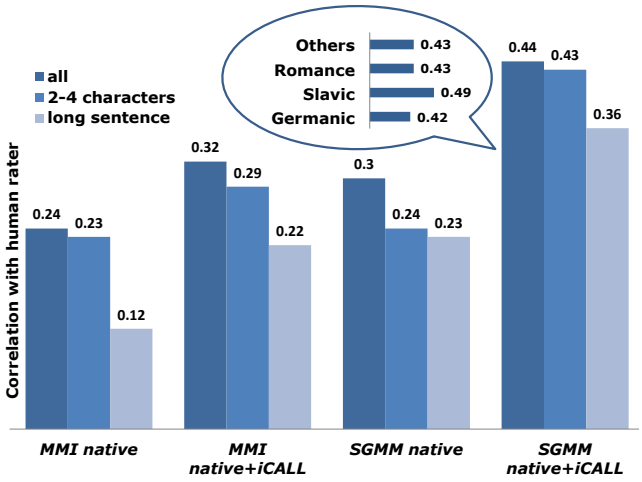

Fig. 2: Fluency assessment results of different acoustic models

Figure 2 illustrates the correlation scores of the four acoustic models described in section 4.1. We observe the following trends:

1) SGMM outperforms MMI-GMM/HMM baselines in fluency prediction. The results show that the two SGMM models consistently achieve higher correlation with the human raters than their corresponding baseline MMI models. We believe it is attributed to the compactness of the SGMM architecture which makes it more sensitive to the variability of the non-native pronunciation and speaking style. The correlation increases by 25% relative (from 0.24 to 0.30) by using *SGMM native* model instead of the *MMI native* model.

2) Non-native training data improves fluency prediction. Higher fluency correlations are obtained by incorporating non-native speech in both MMI and SGMM model training. Adding non-native training data improves the *MMI native* model by 33.3% (0.24 to 0.32) relative and improves the correlation of *SGMM native* model by 46.7% relative (0.30 to 0.44). In addition, fluency assessment of the *MMI native+iCALL* model gives slightly better performance (6.67% improvement) than the *SGMM native* model. This observation implies that SGMM's acoustic characterization ability can partially make up for the lack of non-native training data. Thus SGMM is a good alternative to MMI-GMM/HMM models in scenarios where non-native data is unavailable.

3) Homogeneity of first language background affects fluency scores. The correlation score of *SGMM native+iCALL* system on the four non-native speaker groups are also shown in Figure 2. The fluency prediction of the Slavic speakers outperforms the other three groups of speakers. One possible reason is that there is less first language variation in the Slavic group. As shown in Figure 1, 6 out of 7 test speakers in the Slavic group are from Russia.

4) Utterance length affects fluency scores. To analyze how utterance length might influence fluency scores, the correlation coefficients are broken down into those from short utterances (2-4 characters) and long utterances in Figure 2. We see that automatic scores of short utterances consistently show higher correlation with human scores for all four models.

		Rated Fluency Score			
		1	2	3	4
Actual Fluency Score	1	59.0	23.8	17.0	0.0
	2	19.4	25.7	29.9	24.8
	3	22.0	13.4	32.2	32.2
	4	0.2	1.3	33.7	64.7

Fig. 3: Confusion matrix of *SGMM native+iCALL*

Figure 3 shows the confusion matrix (in percentage) of the automatic fluency prediction from the *SGMM native+iCALL* model. The automatic fluency scores predict level 4 and level 1 the best. The fluency level 3 and 2 are less accurately predicted. This trend actually follows that of human evaluation as this discrepancy is also observed in human ratings, where extreme scores are easier to reach rater-consensus and there is more subjectivity when giving intermediate scores.

5. CONCLUSION AND FUTURE WORK

In this paper, we proposed using SGMM to characterize the acoustic properties of non-native speech. The SGMM model outperforms the MMI-GMM/HMM model in automatic fluency assessment. An SGMM model trained from a small amount of non-native data further improves the fluency correlation than SGMM model trained with only native data.

A speaker's L2 pronunciation is often influenced by his native language. In future work, we plan to use SGMM to further refine non-native speakers into subgroups based on their first language background. For example, the Romance language speakers are more likely to de-aspirate their stop initials [23], which is potentially due to phonemic characteristics of their first language.

6. REFERENCES

[1] M. Peabody, *Methods for pronunciation assessment in computer aided language learning*, Ph.D. thesis, Massachusetts Institute of Technology, 2011.

[2] A. Lee and J. Glass, "A comparison-based approach to mispronunciation detection," in *Spoken Language Technology Workshop (SLT)*, 2012.

[3] K. Yan and S. Gong, "Pronunciation proficiency evaluation based on discriminatively refined acoustic models," *International Journal of Information Technology and Computer Science*, pp. 17–23, 2011.

[4] C. Cucchiarini, H. Strik, and L. Boves, "Quantitative assessment of second language learners fluency by means of automatic speech recognition technology," *Journal of the Acoustical Society of America*, vol. 107, no. 2, pp. 1989–1999, 2000.

[5] C. Cucchiarini, H. Strik, D. Binnenpoorte, and L. Boves, "Towards an automatic oral proficiency test for dutch as a second language: Automatic pronunciation assessment in read and spontaneous speech," in *Proceedings of Instil*, 2000.

[6] H. Li, K. A. Lee, and B. Ma, "Spoken language recognition: From fundamentals to practice," *Proceedings of the IEEE*, vol. 101, pp. 1136 – 1159, May 2013.

[7] X. Huang and L. Deng, "An overview of modern speech recognition," in *Handbook of Natural Language Processing, Second Edition*, N. Indurkhya and F. J. Damerau, Eds. CRC Press, Taylor and Francis Group, Boca Raton, FL, 2010, ISBN 978-1420085921.

[8] S. Witt, *Use of Speech Recognition in Computer-assisted Language Learning*, Ph.D. thesis, Cambridge University, 1999.

[9] F. de Wet, C. van der Walt, and T. Niesler, "Automatic large-scale oral language proficiency assessment," in *Interspeech*, 2007, pp. 218–221.

[10] J. E. Flege, "Factors affecting degree of perceived foreign accent in English sentences," *Journal of the Acoustical Society of America*, vol. 84, no. 1, pp. 70–79, 1988.

[11] X. Qian, H. M. Meng, and F. K. Soong, "Discriminatively trained acoustic model for improving mispronunciation detection and diagnosis in computer-aided pronunciation training (CAPT)," in *Interspeech*, 2010.

[12] A. Lee, Y. Zhang, and J. Glass, "Mispronunciation detection via dynamic time wrapping on deep belief network-based posteriorgrams," in *ICASSP*, 2013.

[13] X. Qian, H. M. Meng, and F. K. Soong, "The use of DBN-HMMs for mispronunciation detection and diagnosis in L2 English to support computer-aided pronunciation training," in *Interspeech*, 2012.

[14] W. Hu, Y. Qian, and F. K. Soong, "A new DNN-based high quality pronunciation evaluation for computer-aided language learning (CALL)," in *Interspeech*, 2013.

[15] S. Wei, G. Hu, Y. Hu, and R. Wang, "A new method for mispronunciation detection using support vector machine based on pronunciation space models," *Speech Communication*, pp. 896–905, 2009.

[16] K. Truong, A. Neri, F. D. Wet, C. Cucchiarini, and H. Strik, "Automatic detection of frequent pronunciation errors made by l2-learners," in *Interspeech*, 2005, pp. 1345–1348.

[17] M. Goudbeek, A. Cutler, and R. Smits, "Supervised and unsupervised learning of multidimensionally varying non-native speech categories," *Speech Communication*, vol. 50, pp. 109–125, 2008.

[18] F. Ge, L. Lu, C. Liu, F. Pan, B. Dong, and Y. Yan, "An Mandarin pronunciation quality assessment system using two kinds of acoustic models," in *Research Challenges in Computer Science, ICRCCS*, 2009, pp. 68–72.

[19] D. Povey, "A tutorial-style introduction to Subspace Gaussian Mixture Models for speech recognition," Tech. Rep., Microsoft Research, 2009.

[20] O. Plchot, M. Karafit, N. Brummer, O. Glembek, P. Matjka, E. V. de, and J. ernock, "A two-stage speaker adaptation approach for Subspace Gaussian Mixture Model based nonnative speech recognition," in *Proceedings of Odyssey 2012, The Speaker and Language Recognition Workshop*, 2012.

[21] L. Burget, P. Schwarz, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. Goel, M. Karafiat, D. Povey, A. Rastrow, R. C. Rose, and S. Thomas, "Multilingual acoustic modeling for speech recognition based on subspace gaussian mixture models," in *ICASSP*, 2010.

[22] B. Li and K. C. Sim, "A two-stage speaker adaptation approach for Subspace Gaussian Mixture Model based nonnative speech recognition," in *Interspeech*, 2012.

[23] N. F. Chen, V. Shivakumar, M. Harikumar, B. Ma, and H. Li, "Large-scale characterization of mandarin pronunciation errors made by native speakers of European languages," in *Interspeech*. IEEE, 2013, vol. II, pp. 803–806.