Automatic Assessment of Non-native Accent Degrees using Phonetic Level Posterior and Duration Features from Multiple Languages*

Shushan Chen†, Yiming Zhou‡ and Ming Li†§
† SYSU-CMU Joint Institute of Engineering, Sun Yat-Sen University, Guangzhou, China
‡ Department of Physics, Zhejiang University, Hangzhou, China
§ SYSU-CMU Shunde International Joint Research Institute, Foshan, China
E-mail: liming46@mail.sysu.edu.cn

***Abstract*—This paper presents an automatic non-native accent assessment approach using phonetic level posterior and duration features. In this method, instead of using conventional MFCC trained Gaussian Mixture Models (GMM), we use phonetic phoneme states as tokens to calculate the posterior probability and zero-oder Baum-Welch statistics. Phoneme recognizers from five languages are employed to extract phonetic level features. It is shown that features based on these five languages' phoneme recognizers are complementary for capturing non-native information and phoneme duration based features are most effective in this task. The final proposed fusion system achieved 0.6089 Spearman's Correlation Coefficient on the test set, which outperformed the openSMILE baseline by 43.3%.**

I. Introduction

Applications of speech technology in support of second language (L2) learning proliferates in recent years, especially for English language learners [1]. With the introduction of computer-assisted pronunciation training (CAPT) system , L2 learners are able to practice their pronunciation without the presence of human teachers. To support CAPT system, many studies have used automatic speech recognition technology to evaluate the non-native accent. Reference [2] presented the Goodness of Pronunciation (GOP) scoring methods which calculate an individual score for each phoneme of an utterance for which the dictionary-based transcription is known. Given an acoustic segment of an utterance, it uses a set of Hidden Markov Models (HMM) to calculate the likelihood corresponding to a specific phoneme and defines the pronunciation quality score as the duration normalized log posterior probability. On the other hand, a few studies evaluated the non-native accent through prosodic parameters which are based on duration, energy, pitch and fundamental frequency information and achieved high correlation with human scores [3][4][5].

Recently, GMM based supervector approaches are widely used for supervised learning of utterance level labels, e.g., age, gender, emotion [6][7][8]. These supervector features originally were proposed for speaker verification tasks [9], but also performed well in the paralinguistic challenges. However, when the utterance duration is short, the first-order Baum-Welch statistics based supervector features perform poorly as feature frames are not enough to calculate the statistics, while the zero-order statistics based supervector achieves better performance [6]. Due to the short duration of speech utterance in our data set (average 5.23 seconds, include silence), we adopted zero-order Baum-Welch statistics based posterior probability supervector as features.

Furthermore, we extended the tokens from the acoustic MFCC trained GMM components to the phonetic phoneme states to calculate the posterior probability supervector feature. Different from previous studies which calculate a score for each phoneme segment, here the posterior probability feature is calculated for each frame and then zero-order Baum-Welch statistics is applied on these frame-based features to calculate the supervector feature for each utterance.

When determining the phoneme segments, most of previous researches only use the English phoneme recognizer. However, non-native pronunciation is a linguistic phenomenon that non-native speakers tend to carry the intonation, phonological processes and pronunciation rules from their mother language [10]. Only use the English phoneme recognizer may not be able to capture some special information from non-native speakers' mother language. Hence, we further applied phoneme recognizers from four additional languages to capture non-native information.

Due to different language proficient levels between native and non-native speakers, the phoneme duration has been shown to be an effective feature [4]. In this work, we further explored some phoneme duration based features. These hand-crafted and specialized features are promising in this task. Given subsystems from different features, score level fusion was employed to further improve the overall performance.

Thre remainder of the paper is organized as follows. The corpus is explained in Section II and the methods are in Section III. Experimental results and discussions are presented in Section IV while conclusions are provided in Section V.

II. Corpus

The data set in this work is provided by the INTERSPEECH 2015 Computational Paralinguistics Challenge organizers [11].

* This research was funded in part by the National Natural Science Foundation of China (61401524), Natural Science Foundation of Guangdong Province (2014A030313123), the Fundamental Research Funds for the Central Universities(15lgjc10), CMU-SYSU Collaborative Innovation research Center Foundation(35321.2.8.1011475) and Guangdong Shunde SYSU-CMU Joint Research Institute.

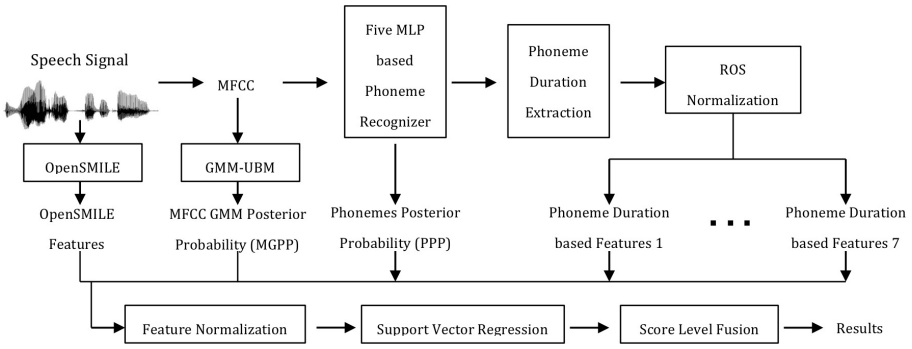

Fig. 1. The System Overview

The training set comes from the AUWL [4] and ISLE [12] corpora. For the AUWL, there are 3732 speech files (5.5 hours, 432 distinct sentences/phrases) from 31 speakers (16 German, 4 Italian, 3 Chinese, 3 Japanese, 5 other). From ISLE, there are 158 speech files (0.3 hours, 5 distinct sentences) from 36 speakers (20 German, 16 Italian). The development set is a subset of the C-AuDiT database [13] and there are 999 speech files (2.7 hours, 19 distinct sentences) from 58 speakers (26 German, 10 French, 10 Spanish, 10 Italian, 2 Hindi). The test set has 594 speech files (1.4 hours, 11 distinct sentences) from 54 speakers (23 German, 12 Chinese, 19 others). The scoring scale of the training set and the testing set are the same that range from 1 for normal to 5 for unusual, while development set ranges from 0 for good to 2 for bad. To better model the real-life situation, all three data sets are disjunct from each other with respect to both speakers and contents [11].

III. Methods

The overview of the proposed system is demonstrated in Fig. 1. In our proposed system, there are multiple subsystems, each using a specific feature, namely, openSMILE feature, MFCC-GMM posterior probability (MGPP) feature, phoneme posterior probability (PPP) feature, and seven phoneme duration based features. We first present the the proposed features in section III-A. Then section III-B describes the regression and score level fusion methods.

A. Features

1) The openSMILE feature: The utterance level 6373 dimensional openSMILE features was extracted by the openSMILE toolkit and provided by the 2015 Paralinguistic Challenge organizers [11]. Since various kinds of features, such as MFCC, loudness, auditory spectrum, voicing probability, F0, and F0 envelop, etc., are included, this feature set can capture both the acoustic and prosodic level information to evaluate non-native accent. In our system, it serves as a baseline.

2) The MFCC-GMM posterior probability (MGPP) feature: For each utterance in the data set, the Universal Background Model (UBM) is applied to extract MGPP feature. Given the GMM-UBM λ with *M* Gaussian components,

$$\lambda_i = \{\omega_i, \mu_\mathbf{i}, \mathbf{\Sigma_i}\}, i = 1, ..., M \tag{1}$$

For each frame-based MFCC feature x_t, the occupancy posterior probability is calculated as follows:

$$P(\lambda_i|\mathbf{x_t}) = \frac{\omega_i p_i(\mathbf{x_t}|\mu_\mathbf{i}, \mathbf{\Sigma_i})}{\sum_{j=1}^{M} \omega_j p_j(\mathbf{x_t}|\mu_\mathbf{j}, \mathbf{\Sigma_j})} \tag{2}$$

This posterior probability reveals the fraction of this MFCC feature x_t belonging to the i^{th} Gaussian component [8]. The Gaussian component can better represent the feature vector if with larger posterior probability. Since the UBM is trained by both native and non-native speeches, this posterior probability can reveal the different distribution between native and non-native accents. Then the MGPP feature is defined as follows:

$$\mathbf{b} = [b_1, b_2, ..., b_M], b_i = \frac{y_i}{T} = \frac{1}{T}\sum_{1}^{T} P(\lambda_i|\mathbf{x_t}) \tag{3}$$

$$MGPP = \sqrt{\mathbf{b}} \tag{4}$$

Equation (3) calculates the zero-order Baum-Welch statistics and we adopt $\sqrt{b}$ as MGPP features in order to apply the Bhattacharyya probability product (BPP) kernel [14].

3) The phoneme posterior probability (PPP) feature: For each utterance in the data set, PPP feature extraction uses the phonetic phoneme states (three states per phoneme) instead of the MFCC trained GMM as tokens to calculate the posterior probability. Then both (3) and (4) are applied to extract PPP feature. We believe that this histogram style feature can provide the phoneme confidence information for distinguishing native and non-native accents.

In this work, we employed the multilayer perceptron (MLP) based phoneme recognizer [15] to calculate frame-based phoneme posterior probability as well as recognizing phoneme components for phoneme duration based features. As has already been mentioned, non-native pronunciation is orginated from the phenomenon that non-native speakers tend to keep his habits of mother language. To better capture non-native information, instead of only using acoustic model from English, we also used models from other four language, namely Czech, Hungarian, Russian and Mandarin to recognize phonemes and further extracted features based on these different languages' phonemes. English based model was trained with 1000 neurons in all nets using switchboard and fisher databases. Mandarin based model was trained by call friend,

call home databases. Other three language based models were trained on SpeechDat-E Databases.

4) Phoneme duration based features: An overview of phoneme duration based features are shown in Table I, which contains seven feature categories. Based on the recognized phonemes segments from five language MLP phoneme recognizers, each feature category generates five features. Note that we omit the silence at the beginning and the end of speech.

The reciprocal phoneme-based rate of speech (ROS) is actually the average phoneme duration. This is a basic but fundamental property of speech since it reveals how fast the speaker said. The second to the fourth feature categories reveal the speaker's hesitation during speech from different perspectives. Non-native speakers are more likely to hesitate during speech and thus there are more pause phonemes. The next two features directly compares the difference between observed phoneme duration and corresponding native or non-native phoneme duration (native phoneme statistics are calculated by training data with score less than 1.5 and non-native phoneme statistics are by training data with score larger than 2.5). Since non-native speakers intend to maintain some pronunciation rules of their mother languages, they might speak some specific phonemes frequently. The last feature, phoeme frequency, is used to reveal such information.

Note that except the first feature, reciprocal ROS, all other phoneme duration based features are normalized by ROS in order to remove the effect of speaking rate. On the other hand, good speakers (with good nativeness score) tend to speak faster than poor speakers. Hence, speaking rate itself is an useful feature but it may affect the aptness of other features.

B. Regression and fusion

We adopt the LIBLINEAR [16] for the Support Vector Regression with parameters tuned from the cross validation sets. As for the score fusion, we adopt two level SVR: the first level generates score for each subsystem and the second level further use these scores as input feature to generate the final score. The score fusion model is trained by another cross validation set which is different from the one used to tune the parameters. When evaluating the testing data set, since the training and development data are with different scales used for annotation, only the training data is used for modeling and the score fusion model is exactly the same as the one tuned on the cross validation set.

IV. Experimental results

The experimental results on the development set with different features for SVR are shown in the Table II. The performance is measured by Spearman's Correlation Coefficient (between the machine evaluated score and human evaluated score). First, the 6373 dimensional openSMILE baseline outperformed 256 dimensional MGPP feature, which might be because that openSMILE includes both acoustic and prosodic information. Only MFCC feature itself might be not powerful enough to differentiate native and non-native accent. Using prosodic contour features [6] together with GMM-UBM approach might be promising, which is a topic for future work.

Second, the PPP feature achieved better performance than the openSMILE baseline. The underlying reason might be that, besides the acoustic information, it also included phonetic information from five different languages. From last column of Table III, we can find that all these single language based PPP features worked worse than openSMILE baseline. However, combining these single language based PPP features improved the result and beated the baseline.

Third, we can find the System 4, which is the fusion system of subsystems form seven phoneme duration based feature, achieved the best result. Table III gives detailed results of each phoneme-based feature on single language and five languages, and the best one among these languages is in bold font. We can find that first three features are very effective that even using only one feature from one single language can beat the openSMILE baseline. Furthermore, features based on only English language fail to perform best in any feature type. To some degree, it reveals that the distinguishing ability is limited if only using English phonemes. After combining features from five languages (the last row), most of these features are

TABLE I
AN OVERVIEW OF PHONEME DURATION BASED FEATURES

Feature Category	Definition
Reciprocal Rate of Speech (ReciROS)	The reciprocal of the number of phonemes per time.
Average pause phoneme duration (AvgPauDur)	The average pause phoneme duration of a speech.
Voiced phoneme duration ratio (VoiPhoDurRatio)	The ratio between the sum of voiced phoneme duration and the total duration.
Voiced phoneme number ratio (VoiPhoNumRatio)	The ratio between the number of voiced phonemes and all phonemes.
Phoneme duration native difference (PhoDurNatDiff)	The difference between the native phoneme duration and observed phoneme duration.
Phoneme duration non-native difference (PhoDurNonDiff)	The difference between the non-native phoneme duration and observed phoneme duration.
Phoneme frequency (PhoFreq)	The phonemes frequency within an utterance (set inexistent phoneme to zero).

TABLE II
PERFORMANCE ON THE DEVELOPMENT SET WITH DIFFERENT FUTURES FOR SVR AND SCORE LEVEL FUSION

System	Features	Parameter C	Correlation
1	openSMILE	0.002	0.4119
2	MGPP[a]	0.01	0.2846
3	PPP (5 languages)	0.1	0.4251
4	Phoneme duration based features (5 languages)	Refer to Table III[b]	0.5623[c]
5	Fusion 4+2		0.5614
6	Fusion 4+2+3		0.5694
7	Fusion 4+2+3+1		0.5617
8	Fusion 4+3		**0.5706**

[a] The size of GMM is 256.
[b] It depends on different features, please refer to Table III.
[c] This is the fusion score of seven phoneme duration based feature subsystems.

TABLE III
PERFORMANCE ON THE DEVELOPMENT SET USING FEATURES CALCULATED WITH TOKENS FROM DIFFERENT LANGUAGES

Features & Languages	ReciROS	AvgPauDur	VoiPhoDurRatio	VoiPhoNumRatio	PhoDurNatDiff	PhoDurNonDiff	PhoFreq	PPP
English	0.4861	0.4958	0.4845	0.3996	0.3028	0.3659	0.3643	0.3628
Mandarin	0.4752	0.5135	0.5635	0.4895	**0.4578**	**0.4581**	0.3478	0.3577
Czech	0.4993	0.5073	0.5423	0.5613	0.2998	0.2887	0.3676	0.3400
Hungarian	**0.5392**	0.5273	0.5666	0.5802	0.3744	0.3558	0.3907	0.**3704**
Russian	0.5195	**0.5462**	**0.5887**	**0.5995**	0.2532	0.3000	**0.4101**	0.3531
Parameter C	0.01	0.001	10	10	0.01	0.01	0.01	0.1
5 languages	0.5400	0.5368	0.5207	0.5892	0.4909	0.4903	0.4609	0.4251

TABLE IV
PERFORMANCE ON TEST SET WITH BASELINE AND SYSTEM 8

	openSMILE baseline	System 8
Correlation	0.4250[a]	**0.6089**

[a] The detailed result is presented in [11].

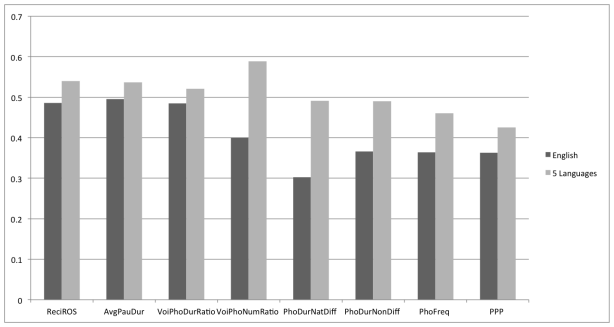

Fig. 2. Performance on development set with phoneme duration based features from single English language or five languages

improved or achieve the result close to the best one among five single language. When compared with features from only English language, all fused features from five languages achieved better results, which is shown in the Fig. 2.

With respect to the systems with score fusion, as shown in the Table II, the System 8 which combining PPP feature and phoneme duration based features achieved the best performance 0.5706 on the development set. On the other hand, Table IV provides the results of openSMILE baseline and System 8 on test set (we cannot provide the result of each system on test set because that the label of test set is not provided and we only have 10 trails to submit the result). Our proposed System 8 achieved 0.6089 on test set which outperformed the openSMILE baseline (0.4250) by 43.3%.

V. CONCLUSIONS

This paper presents an automatic non-native accent assessment approach using phoneme posterior probability and phoneme duration based features from multiple languages. To better capture non-native information, phoneme recognizers from five different languages are employed. Experimental results show that, features based on these five languages phonemes are complementary for capturing non-native information. Phoneme duration based features are most effective features and the performance is further improved after score fusion with PPP System which provides more information bout phoneme confidence. MGPP feature, which is based on GMM-UBM approach, does not perform well in this task. However, instead of using MFCC, other features such as prosodic contour features might be promising and this is a topic for our future work.

REFERENCES

[1] M. Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, vol. 51, no. 10, pp. 832–844, 2009.
[2] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2, pp. 95–108, 2000.
[3] F. H02nig, A. Batliner, K. Weilhammer, and E. N02th, "Automatic assessment of non-native prosody for english as l2," *Proc of Speech Prosody*, 2010.
[4] F. Hönig, A. Batliner, and E. Nöth, "Automatic assessment of non-native prosody annotation, modelling and evaluation," in *Proc. of ISADEPT–International Symposium on Automatic Detection of Errors in Pronunciation Training*, 2012, pp. 6–8.
[5] C. ACucchiarini, H. Strik, and L. Boves, "Quantitative assessment of second language learners fluency: Comparisons between read and spontaneous speech," *the Journal of the Acoustical Society of America*, vol. 111, no. 6, pp. 2862–2873, 2002.
[6] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Computer Speech & Language*, vol. 27, no. 1, pp. 151–167, 2013.
[7] D. Bone, M. Li, M. P. Black, and S. S. Narayanan, "Intoxicated speech detection: A fusion framework with speaker-normalized hierarchical functionals and gmm supervectors," *Computer Speech & Language*, vol. 28, no. 2, pp. 375–391, 2014.
[8] M. Li, "Automatic recognition of speaker physical load using posterior probability based features from acoustic and phonetic tokens," in *Proc. Interspeech*, 2014.
[9] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using gmm supervectors for speaker verification," *Signal Processing Letters, IEEE*, vol. 13, no. 5, pp. 308–311, 2006.
[10] J. Leather, "Second-language pronunciation learning and teaching," *Language Teaching*, vol. 16, no. 03, pp. 198–219, 1983.
[11] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger, "The interspeech 2015 computational paralinguistics challenge: Nativeness, parkinsons & eating condition," in *Proc. Interspeech*, 2015.
[12] W. Menzel, E. Atwell, P. Bonaventura, D. Herron, P. Howarth, R. Morton, and C. Souter, "The isle corpus of non-native spoken english," in *Proceedings of LREC 2000: Language Resources and Evaluation Conference, vol. 2*. European Language Resources Association, 2000, pp. 957–964.
[13] F. Hönig, A. Batliner, K. Weilhammer, and E. Nöth, "Islands of failure: employing word accent information for pronunciation quality assessment of english l2 learners." in *Proc. SLaTE*, 2009, pp. 41–44.
[14] T. Jebara, R. Kondor, and A. Howard, "Probability product kernels," *The Journal of Machine Learning Research*, vol. 5, pp. 819–844, 2004.
[15] P. Schwarz, P. Matejka, and J. Cernocky, "Hierarchical structures of neural networks for phoneme," in *Proc. ICASSP*, 2006, pp. 325–328.
[16] R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang, and C. J. Lin, "Liblinear: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, no. 12, pp. 1871–1874, 2008.