GAUSSIAN MIXTURE MODELING OF VOWEL DURATIONS FOR AUTOMATED ASSESSMENT OF NON-NATIVE SPEECH

Xie Sun

Department of Computer Science
University of Missouri, Columbia, MO 65211, USA
xs8pb@mail.missouri.edu

Keelan Evanini

Educational Testing Service
Princeton, NJ 08541, USA
KEvanini@ets.org

ABSTRACT

This paper investigates using Gaussian Mixture Model (GMM) based vowel duration features for automated assessment of non-native speech. Two different types of models were compared: a single GMM trained on a reference corpus of native speech and separate GMMs for different proficiency levels trained on a large corpus of scored non-native speech. 13 vowel categories were evaluated separately (after normalization by rate of speech), and a multiple regression model was used to evaluate the performance of all vowel categories combined. Experiments on an English language proficiency assessment show that the non-native speech GMMs outperform the native speech GMMs, and that all 13 vowels have significant correlations with human scores when the non-native speech GMMs are used. The multiple regression combination obtained correlations with human scores of 0.71 when transcriptions were used to extract the vowel durations and 0.64 when the Automatic Speech Recognition (ASR) output was used. The experiments demonstrate that the vowel duration feature based on non-native speech GMMs is a useful predictor of L2 proficiency and is robust to different datasets and situations.

Index Terms—automated pronunciation assessment, vowel duration, Computer Assisted Language Learning, Gaussian Mixture Model

1. INTRODUCTION

In the last two decades, many studies have used automatic speech recognition (ASR) technology to assess non-native pronunciation in read speech [1, 2, 3]. In these studies, pronunciation features calculated from the recognition process, such as average ASR confidence values, have consistently shown moderate to high correlations with human proficiency scores. Several studies have also used segment duration scores as a predictor of pronunciation proficiency. In a simple approach, [4] correlated mean raw segment duration values with pronunciation scores (although this metric was also highly correlated with speaking rate). [5, 6] calculated the average deviations between a non-native speaker's segments and mean segment durations trained on a native speaker corpus. In another approach, [2, 6] used the average log probability of segment durations based on native speaker distributions of durations for each phone (after normalizing for rate of speech). [7] used a similar approach to calculate duration scores for individual phones, but concluded that individual phone duration scores were almost uncorrelated with the corresponding human ratings.

Most of the previous research on using segment duration scores for non-native speech assessment has used native speech as a reference and made comparisons between native and non-native speech. However, vowel duration features using non-native speech data as a reference have not been studied yet. Intuitively, the speech dissimilarity between native and non-native speakers is much larger than among non-native speakers themselves. A study of vowel space characteristics between native and non-native speakers [8] provides support for this intuition. So it may be meaningful to use non-native speaker data to assess non-native speech. Another important advantage of non-native speech data is the presence of data from speakers with different proficiency levels that can be used to train classification models for different levels, which can further help the discrimination of speakers with differing language proficiency. However, for native speakers, we assume all of them have the same language proficiency. In addition, instead of using single Gaussian models, we propose to use GMMs to model the vowel sound durations since they have two valuable properties: 1) GMMs can divide the vowel sound durations from different kinds of speakers into different clusters; 2) GMMs can partition the durations of stressed and unstressed vowels automatically, which helps to build more powerful and precise classifiers. Even though the GMM training of non-native speech requires a relatively large amount of training data for each score level, the approach has the potential to better discriminate speakers of different proficiency levels. Therefore, in this paper, we compare the performance of GMMs trained using both native and non-native speech data to evaluate the effect on segment duration scores. We focus our investigation specifically on vowel sounds, since they are perceptually more salient.

This paper is organized as follows: Section 2 describes the vowel categories and the data sets used in the experiments; Section 3 presents the methodology used for computing the features; Section 4contains the experimental results; and Section 5 discusses the findings and proposes future research directions.

2. DATABASE DESCRIPTION

2.1 Vowel Category Introduction

Based on the CMU pronouncing dictionary, there are a total of 15 vowel categories: AY, EH, IH, UW, IY, AE, AW, ER, AO, AH, OW, EY, AA, OY and UH. In our study, we exclude the vowels OY and UH since they have too few tokens for training a GMM. In addition, we combined all stressed and unstressed tokens for each vowel in the training data. Most of the vowel categories contain many more stressed tokens than unstressed ones; the vowels AH, ER, IH, and IY are exceptions to this generalization. In our training and evaluation sets, AH and ER have more unstressed tokens than stressed ones, and IH and IY have an almost equal number of stressed and unstressed tokens.

2.2 Native Speech Corpus

We used the spoken responses from the Atlas of North American English (ANAE) as the native speech corpus. This corpus includes 437 speakers sampled from every major dialect region in North America [9]. Each speaker participated in an interview of

978-1-4577-0539-7/11/$26.00 ©2011 IEEE

approximately 30 minutes consisting of spontaneous speech and targeted elicitation of specific lexical items. A subset of the words bearing phrasal stress were manually extracted by the ANAE annotators; Table 1 provides the number of tokens for each vowel contained in this data set.

Table 1 Statistics of vowel tokens for the ANAE data

Vowel	# of tokens	Vowel	# of tokens
AH	20405	AA	11627
AW	17043	AO	10709
OW	16657	UW	8778
EY	15827	AY	8269
AE	14527	IY	8223
IH	14440	ER	6540
EH	12751	Total	165796

2.3 Non-native Speech Corpus

The non-native speech corpus consists of responses to Read Aloud items drawn from an English proficiency assessment. The training corpus consists of 737 speakers from two countries: 448 from China and 289 from India (representing a range of L1 backgrounds). The evaluation corpus contains 99 Chinese speakers and 155 Indian speakers. Each speaker responded to four Read Aloud items (45 sec. in duration) in which they were instructed to read a paragraph out loud. Each response was scored by two different expert raters for pronunciation proficiency using the following three-point scale: score 1 (low-level) is used for non-native speech that is not generally intelligible; score 2 (medium-level) for speech that is generally intelligible with some lapses; and score 3 (high-level) for speech that is highly intelligible. For each speaker, a speaker-level score is calculated by adding all the scores from the four items. The Pearson correlation between the two speaker-level scores (each comprising scores from only 4 individual items) for the two human raters is 0.75.

Due to the design of the assessment, there were three distinct sets of four Read Aloud items, meaning that the speakers did not all produce the same lexical items. Thus, the total number of tokens used to calculate the vowel duration features for each speaker varied. Table 2 shows the statistics for different vowel categories in each score level for the mixture of India and China training data. As the table shows, the data is very unbalanced: the number of tokens for score 2 is around 2 times and 3 times the number of tokens for score 3 and score 1, respectively. Compared with the counts for the native speech corpus in Table 1, the non-native speech training corpus has many fewer tokens for each vowel in each score level. The training and evaluation sets were sampled to reflect the overall score distribution of the corpus, and the evaluation set includes the same 12 Read Aloud items from the training set (with no speaker overlap).

3. PRONUNCIATION SCORING

3.1 Vowel Duration Extraction and Normalization

Vowel durations in the native speech corpus were obtained by using a forced alignment system to provide phone-level boundaries for the words in [9]. For the non-native speech corpus, the ASR based vowel duration extraction uses a two-pass method [5]: first the utterance is recognized using a non-native speech acoustic model (AM); then a native speech AM is used for forced alignment. This study also examines the performance of the vowel duration features when human transcriptions are used as input to the forced alignment instead of the ASR hypotheses; in this case, the ASR step is bypassed.

Previous research has showed that Rate of Speech (ROS) is highly correlated with proficiency scores, e.g. [1]. In order to remove the effect of ROS so that the actual effect of the vowel duration feature can be studied, the vowel durations are usually normalized to compensate for the ROS of a particular speaker [2]. The simplest approach to ROS normalization is to compute the average number of phones per unit of time for a given speaker. The normalized duration, x, is then computed as:

$$x = d_i \times ROS_s \qquad (1)$$

where ROS_s is the estimated ROS for speaker s and d_i is the duration for the vowel i.

Table 2 Statistics of vowel tokens of different score levels for the mixture of Indian and China training data

Vowel	# of tokens		
	score1	score2	score3
AH	9052	28860	13659
AW	618	2134	970
OW	1310	4009	1980
EY	1701	5463	2597
AE	2520	8514	4001
IH	4608	15528	7335
EH	2371	7428	3595
AA	2063	6460	3081
AO	1605	4834	2231
UW	1941	6002	2735
AY	912	3215	1638
IY	3801	12190	5919
ER	1781	6373	3036
Total	34283	111010	52777

3.2 Native Speech GMM Training and Scoring

A GMM is the linear combination of several Gaussian models. It is able to approximate any distribution and deals very well with diverse data. We train the GMMs for 13 vowel categories of native speech. The log-likelihood scores of 13 vowel categories in non-native speech are calculated using the corresponding native speech GMMs. These are then used as the vowel duration scores to differentiate between non-native speakers with different proficiency levels by seeing how far away from the native speaker distribution each non-native speaker is. A GMM density function for the durations of a vowel category is described as:

$$G_{vowel}(x) = \sum_{i=1}^{M} w_i \, g\left(x \middle| \mu_i, \sigma_i^2\right), \quad \sum_{i=1}^{M} w_i = 1 \qquad (2)$$

where w_i is the component weight,M is number of Gaussian components, and$g\left(x \middle| \mu_i, \sigma_i^2\right)$ is a Gaussian density function which is defined as:

$$g\left(x \middle| \mu_i, \sigma_i^2\right) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{-\frac{\left((x-\mu_i)\right)^2}{2\sigma_i^2}\right\} \qquad (3)$$

where μ_i is the component mean and σ_i^2 is the component variance. The predicted vowel duration score for a vowel category using native speech GMMs is calculated as:

$$S_{vowel} = \frac{1}{n}\sum_{i=1}^{n} \log G_{vowel}(x_i) \qquad (4)$$

where x_i is a vowel duration instance in a vowel category from non-native speech and$n$ is the number of vowel instances.

3.3 Non-native Speech GMM Training and Scoring

In addition to using native speech data to train the GMMs, we also used non-native speech data to do so. The difference in this case is that there are different proficiency levels for the non-native speakers. So, for the non-native speech data, we trained three GMMs for each of the 13 vowel categories (one GMM for each score level). We then use the non-native speech GMMs for each score level to differentiate non-native speakers of different proficiency levels. To do this, we combine the scores for each level with the associated posterior probabilities to obtain the final predicted score. The posterior probability is defined for a vowel duration instance x_j in a vowel category as:

$$P^{j}_{vowel} = P(Vowel_k|x_j) = \frac{p(x_j|Vowel_k)}{\sum_{i=1}^{l} p(x_j|Vowel_i)} \qquad (5)$$

where l represents the number of different language proficiency scores and $Vowel_k$ is one of the non-native speaker vowel duration GMMs for a vowel category. We define the predicted score for a vowel instance j as:

$$S^{j}_{vowel} = \sum_{k=1}^{l} P^{j}_{vowel} \times S_k \qquad (6)$$

where S_k is an item level score (1, 2, or 3).We define the predicted vowel duration score for a vowel category as:

$$S_{vowel} = \frac{1}{n}\sum_{j=1}^{n} S^{j}_{vowel} \qquad (7)$$

where n is the number of vowel instances.

In Eq. (6), we combine different level non-native speaker models together in order to reduce the prediction errors. This method of combining different models generally performs better than using the 1-best model [10].

3.4 Multiple Regression Model for Ensemble Features

In order to evaluate the performance of combing all 13 vowel duration features, a multiple regression model is used. It is defined as:

$$S_{ensemble} = \sum_{i=1}^{13} \alpha_i S_i + \beta \qquad (8)$$

where α_i is the weight and S_i is the predicted score of a vowel category. β is the intercept. In our experiment in Section 4, we partitioned a multiple regression training data to obtain the weight α_i and intercept β and tested them on the evaluation dataset.

4. EXPERIMENTS

In this part, we compare the performance of native and non-native speech GMMs. In addition, we also evaluate the non-native speech model on different datasets and situations. Since an item usually includes a relatively limited number of vowel tokens, in which not all of different vowel categories are represented well, we use the tokens produced by each speaker in each vowel category from all four items to calculate the speaker level predicted scores in order to facilitate the comparison among speakers who read different items in our experiments. Pearson correlations for speaker level scores between one human rater and machine scores were calculated. We also did the experiments for non-native speaker models based on the 1-best model mentioned in Section 3.3. However, the correlations decreased around 0.04 on average for all 13 vowel categories. So the related subsequent experiments used the method of combined non-native speaker models. In addition, single Gaussian models were also tried on the same dataset, but the performance was much worse than when GMMs were used.

4.1 Comparison of NSG and NNSG

Figure 1 presents a performance comparison between the native speech GMMs (NSG) and non-native speech GMMs (NNSG) for all 13 vowel categories using the transcription based India training and evaluation data. Except for the vowel AE, all vowel categories show better performance using NNSG than using NSG. Vowel AH has the highest correlation in both cases: 0.64 for NNSG and 0.41 for NSG. From Tables 1 and 2, we can see that NSG has many more tokens to train the more robust GMMs than NNSG does. However, in general, the performance for NSG is much lower than NNSG. Based on this conclusion, the following experiments were performed only using NNSG.

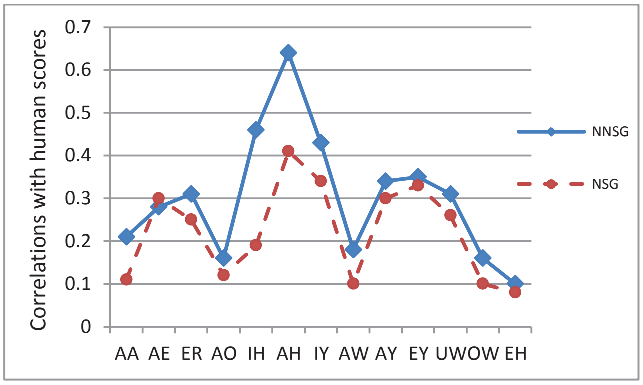

Fig. 1 Correlation comparison between NSG and NNSG with Indian data

4.2 Robustness and Effectiveness of NNSG for Different Datasets and Situations

Table 3 shows the correlations based on the transcriptions for the data from India and China. We used the Indian training data to train Indian non-native speech GMMs (INNSG) and used the Indian evaluation data as test set. We also used China training data to train China non-native speech GMMs (CNNSG) and used the China evaluation data as test set. All vowel categories have significant correlations with human scores. Vowel AH has the highest correlations at 0.64 and 0.71 for the India and China datasets, respectively. Table 4 compares the results based on the transcriptions and ASR outputs using a mixture of India and China training and evaluation data. We only used ASR outputs for the evaluation data and the GMM were still trained using the transcription based India and China mixed training data. The word error rates for the ASR outputs were 0.20 and 0.30 for the India and China evaluation sets, respectively. We partitioned the mixed evaluation data into two equal sets which each included the same number of Read Aloud items and had the same speaker level score distribution. One set was used to train the multiple regression model and the other was used as the final mixed evaluation set. The set used to build the regression model was also used to tune the optimal number of components in all GMMs. The results based on ASR outputs do not decrease very much compared to the results based on transcriptions. The highest and lowest correlations for these two different cases are exhibited by AH and ER. We also used the multiple regression model to evaluate the combined results from all vowel categories. The overall performance combining all 13 vowel durations is 0.71 when the transcriptions were used and 0.64 for the ASR output based results.

In order to validate how robust this approach is, we also conducted experiments using non-native speech from different L1 backgrounds for the training and the evaluation data. Figure 2 compares the performance of INNSG and CNNSG using India evaluation data and Figure 3 compares the performance of INNSG and CNNSG using China evaluation data. Even though most of the individual vowel correlations are lower when training data from a different L1 background is used, the average correlations decrease by only 0.047 and 0.079 for the different situations in Figures 2 and 3, respectively. The comparisons in Figures 2 and 3 show that the non-native speech data from different L1 backgrounds can complement each other when GMMs are trained. This is a very valuable property, especially when the amount of training data from different non-native speech resources is relatively small. It could also be a good explanation why the general performance for the mixed data from India and China in Table 4 is better than that of using either the India or China datasets alone in Table 3.

Table 3 Different vowel duration Pearson correlations with human scores based on transcriptions for India and China data separately

Vowel	Correlations (China)	Correlations (India)
AH	0.71	0.64
AA	0.50	0.21
EH	0.42	0.10
AE	0.41	0.28
AY	0.39	0.34
AO	0.30	0.16
EY	0.30	0.33
IH	0.27	0.46
ER	0.27	0.31
OW	0.25	0.16
AW	0.24	0.18
IY	0.23	0.43
UW	0.21	0.29
Average	0.35	0.30

All 13 vowel category correlations are significant at α=0.05

Table 4 Different vowel duration Pearson correlations with human scores based on transcriptions and ASR outputs for the mixture of India and China evaluation data

Vowel	Correlations (transcriptions)	Correlations (ASR)
AH	0.68	0.60
IH	0.51	0.45
AA	0.48	0.42
AE	0.42	0.45
IY	0.41	0.40
ER	0.41	0.37
AO	0.32	0.26
AW	0.31	0.37
AY	0.30	0.36
EY	0.27	0.25
UW	0.26	0.23
OW	0.25	0.28
EH	0.24	0.22
Average	0.37	0.36
Regression	0.71	0.64

All 13 vowel category correlations are significant at α=0.05

5. CONCLUSION AND FUTURE WORK

In this paper, we have investigated a GMM-based vowel duration feature for the assessment of non-native pronunciation. Two approaches were tried based on native and non-native speech models. From the experimental results, we conclude that non-native speech GMMs have better performance. The method is very robust to the different datasets and ASR outputs. In addition, it does not necessarily require speech from the same L1 background in the training and evaluation datasets, which can increase the speed of development for a practical application. A disadvantage for the native speech GMM training is that they were trained using the ANAE corpus [9] which is comprehensive, but has different speech styles and a different distribution of lexical items compared to the Read Aloud items in the assessment. So in the future work, we will use vowels associated with their context, which guarantees all vowels to have the same lexical structure. We also plan to use more non-native speech data to train more robust GMMs. We will also try the approach on speakers from a larger number of countries. Finally, we hope to evaluate the approach on spontaneous speech.

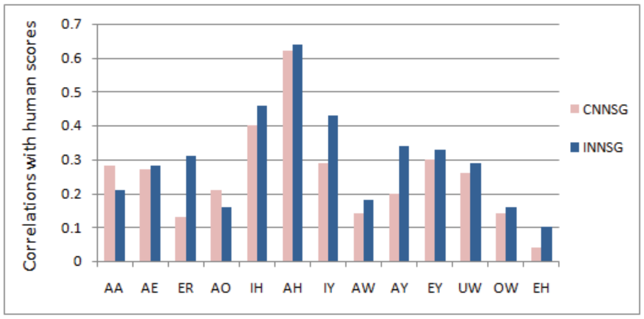

Fig.2 Correlation comparison between CNNSG and INNSG using Indian evaluation data

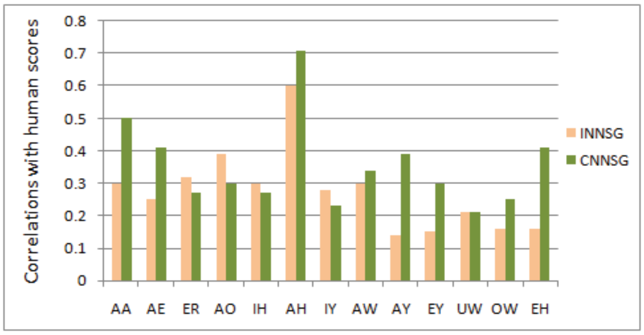

Fig.3 Correlation comparison between CNNSG and INNSG using China evaluation data

6. REFERENCES

[1] J. Mostow, S.F. Roth, A.G. Hauptmann, and M. Kane, “A prototype reading coach that listens,” in Proc. AAAI, 1994, pp. 785–792.

[2] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, “Automatic Scoring of Pronunciation Quality,” *Speech Communication*, vol. 30, pp.83–93, 2000.

[3] S. M. Witt, “Use of Speech Recognition in Computer assisted Language Learning,” Ph.D. thesis, University of Cambridge, 1999.

[4] Cucchiarini et al. (1997a) "Automatic evaluation of Dutch pronunciation by using speech recognition technology,” In: Furui, S., Juang, B.H.,Chou, W. (Eds.), Proceedings IEEE Workshop, ASRU, Santa Barbara, pp. 622–629.

[5] L. Chen, K. Zechner, and X Xi, “Improved pronunciation features for construct-driven assessment of nonnative spontaneous speech,” in NAACL-HLT, 2009.

[6] Hacker et al. (2005) "Pronunciation feature extraction," in Pattern Recognition, 27th DAGM Symposium , 2005, pp. 141–148.

[7] Kim et al. (1997) "Automatic pronunciation scoring of specific phone segments for language instruction," In: Proceedings of the European Conference on Speech Communication and Technology 1997. Rhodes, pp. 649–652.

[8] L. Chen, K. Evanini, and X. Sun. "Assessment of non-native speech using vowel space characteristics," in Proc. SLT 2010.

[9] W. Labov, S. Ash, and C. Boberg, *The Atlas of North American English*, Mouton de Gruyter, 2006.

[10] Y. Wang, J. Yang, and N. Peng, “Unsupervised color-texture segmentation based on soft criterion with adaptive mean-shift clustering,” *Pattern Recognition Letters*, vol. 27, no. 5, pp. 386–392, 2006.