Automatic reference independent evaluation of prosody quality using multiple knowledge fusions

Shen Huang, Hongyan Li, Shijin Wang, Jiaen Liang, Bo Xu

Digital Content Technology Research Center, Institute of Automation
Chinese Academy of Science, Beijing 100190, China
{shenhuang,hyli,sjwang,jeliang,xubo}@hitic.ia.ac.cn

Abstract

Automatic evaluation of GOR (*Goodness Of pRosody*) is a more advanced and challenging task in CALL (*Computer Aided Language Learning*) system. Apart from traditional prosodic features, we develop a method based on multiple knowledge sources without any prior condition of reading text. After speech recognition, apart from most state-of-the-art features in prosodic analysis, we cultivate more concise and effective feature set from the generation of prosody based on Fujisaki model, and influence of tempo in prosody—the variability of prosodic components based on PVI method. We also propose methods of boosting training without any annotation by mining larger corpus. Results in experiment investigate the GOR score on 1297 speech samples of excellent group of Chinese students aging from 14-16, we can draw several conclusions: On the one hand, adding the knowledge sources from generation and impact of prosody can contribute to 1.76% reduction in EER and 0.036 promotion in correlation than prosodic features alone; On the other hand, final result can be considerably improved by boosting training approach and topic-dependent scheme.
Index Terms: speech prosody, PVI, Fujisaki model

1. Introduction

The performance of prosody when evaluating L2 language learners plays an important role in computer aided language learning systems [2][3][4], especially when characterizing the quality of speech uttered by *excellent* group of people. Segmental analysis such as GOP (*Goodness Of Pronunciation*) [1] and GOF (*Goodness of Fluency*) [12] can no longer be effective enough. In linguistic and educational views, people who are emotional and vibrantly place and choose phrases and accents can yield a more relevant impression of hearing. Further, people convey their emotions dominantly by skills of prosodic variation and round-about expression. For L2 learners, speech proficiency is more vulnerable to prosodic errors, such as monotonous prosody, unnecessary tonal change, etc.

Tremendous researches have focused on fully automatic evaluation of goodness in pronunciation and fluency [1][12]. But when it comes to the evaluation of GOR, rarer research is put forward. Some work involve in extraction of large dimension of prosodic features which are later selected in ad hoc manner and score is acquired by classifiers [2][3]. However, it is lack of structural explanation of prosody production and impact from intonation units [2][3], or the performance of the reference independent part compared to segmental cues is unfavorable [2]. Other works [4] introduce intonation-model based scoring by training HMM for categorical intonation units on continuous f0 and energy contours from native speech. In [5] the authors narrow the scope of research to the liveliness of GOR defined by pitch fluctuation in speech, which turned out to be a relative reliable result to distinguish speech between good and excellent prosody. In this paper we will extend a different thought and investigate prosody by three parts of knowledge: 1) large dimensions of currently used prosodic features; 2) how speech prosody is generated; 3) the impact of prosody. In the second part, we propose to use a well-know method—Fujisaki models [10], for characterizing prosody by phrase and accent. For the third part, PVI model, a recently used tool in distinction of language type (stress & syllable timed, etc), is introduced in pitch, energy and duration levels. Next section will present our multiple knowledge bases of GOR feature set. In Section 3 we will briefly introduce our corpus in training and test, and a boosting scheme in data training is also proposed before evaluation and conclusion of results.

2. Algorithms

2.1 Expression of prosody—baseline prosodic features

In order to determine which set of features are relevant to specify GOR of speech, we first introduce prevalently used prosodic features. Like many previous study in prosody [2][3], there is no consensus on how to formulate, but we try to attain a comprehensive approach by exploring as much available information as possible. Features are extracted in four components, namely pitch, intensity, duration and formant. In contrast, we take a fusion of similar reference independent features proposed in [2] and [3]. like mean, variance, maximum statistics of duration, energy, pitch, and formant of different group of vowels, consonants; number and duration of pauses; Mean Length Run (MLR); Articulation RaTe (ART); rate of speech; duration, count and ratio of the voiced and unvoiced segments; different order of moments, maximum, minimum, onset and offset in pitch, energy, duration and formant both in voiced and unvoiced segments, etc. A total of 153 features are acquired, 62 of which is chosen at the minimized error by method of SFFS feature selection algorithm [6], the object function of which is by minimizing the overall error rate of test set generated by a SVM classifier illustrated later in experiment.

2.2 Generation of prosody—Fujisaki models

Having analyzed the production of pitch, people recently tend to reverse the thought and develop in a number of studies based on speech synthesis, especially the algorithms for modeling speech prosody [9] such as ToBI, Tilt, INTSINT, and Fujisaki model etc, from which Fujisaki model is selected in our work due to its robustness, physiological interpretation connecting pitch movements with the dynamics of the larynx [10]. The following approach supplies cues concerning the relationship between statistical properties of Fujisaki model and GOR score.

After pitch contour of speech is extracted in 10ms steps and pre-process is performed, we use the well-known procedure for extracting Fujisaki model by Mixdorff [11], a multistep version that exploits Fujisaki parameters by structure of filters applied

in pitch stylization. Final parameter is obtained by a hill-climb search for local mini- and maximum in filtered contour, which reproduces a given F0 contour by superimposing three components in the log F0 domain: 1) A speaker-individual base frequency *Fb*; 2) The phrase component, which results from impulse responses to impulse-wise phrase commands associated with prosodic breaks; 3) The accent component, which results from step-wise accent commands associated with accented syllables. After a series of boundaries of the two commands are generated for each speech, we exploit such features related to components of Fujisaki model: 1) number of phrase command; 2) number of accent command; 3) mean and standard variance of amplitude, duration and envelop of accent command; Totally 8 features are generated for analysis.

2.3 Impact of tempo in prosody—PVI operator

Duration and pitch variability poses a great diversity in different levels of prosodic speech. Intuitively speaking, people who are delicate and prominent in prosody tend to pronounce melodious wave of duration and pitch, which is then composed in various manners, yielding an uncertain form of prosodic features in syllables. These syllables are then combined with one another to form a sound rhythm. In Hincks's work [5], this phenomenon was simply modeled by F0 variance, which seems unilateral. Nevertheless, Investigations [2][3][4] have shown that the effects of variation and uncertainty on other prosodic components need to be taken seriously. In this study, for a better reflection of the auditory impression of different GOR, several measurements have been augmented from the output of an improved version of our HMM based automatic speech recognition system [13].

Ramus Class: A routine measure in speech prosody [7], taking %V (percentage of vocalic duration in speech). ΔC, ΔV, ΔS (standard deviation of duration and pitch in consonantal, vocalic, and syllabic levels) as the scope of study. This measure is calculated for each sentence, which is then averaged to compose a passage measure. Further, since the above measures have been demonstrated to interact with the average segment prosody, we apply a normalization procedure as follows:

VacroC = ΔC / mean consonantal duration (pitch) *100
VacroV = ΔV / mean vocalic duration (pitch) * 100
VacroS = ΔS / mean syllabic duration (pitch) * 100

Grabe and Low Class: Previous work in PVI [8] took only the duration of consonants in rPVI and vowels in nPVI as the basis for clustering stress- or syllable- timed languages, which provided ample evidences for dialectal variations in study of prosody. However we extend it in other prosodic components. First tsylb2.1 (*http://www.nist.gov/speech/index.htm*) is used to separate phoneme sequences of word into syllables, and rPVI and nPVI is computed in syllabic, vocalic, and consonantal levels (e.g. s-nPVI, v-nPVI, c-nPVI). All the measures are calculated in pair-wise steps through both global and IPS (Inter Pause Segments) levels. IPS [8] is introduced to segment speech with pause intervals and to avoid extreme values from phrase-final lengthening. The result rPVI of it is averaged later. e.g, the two PVI formulae of duration are computed as follows:

1) Raw PVI of duration:

$$rPVI = \sum_{k=1}^{m-1} \left| d_k - d_{k+1} \right| / (m-1)$$

2) Normalized PVI of duration:

$$nPVI = 100 \times \sum_{k=1}^{m-1} \left| \frac{d_k - d_{k+1}}{(d_k + d_{k+1})/2} \right| / (m-1)$$

Where m is the number of intervals. *dk* is the duration of *k* th interval. Notice that the normalization in 2) is claimed to be necessary in order to counteract variation in tempos of different type of phoneme. The duration ratio of vowel versus consonant is included, too. Totally 19 features are generated.

3. Database

3.1 Corpus description

Corpus is taken from our collection of the most excellent group of Chinese students with good English speaking skills from age 14 to 16. A direct thought for selecting excellent students is that we try to maximally eliminate the influence of other quality of speech, e.g. integrity, pronunciation, fluency etc. All subjects are provided with opportunities to read and practice the English texts beforehand in order to be able to concentrate on reading as naturally as possible. It takes 90 sec to record each speech.

Reading materials cover 8 different topics of passage, each contains about 110 normal words. Finally there are 14'880 (1860×8) students recorded with sample rate of 16khz, 16bit. 1297 samples of which are chosen topic-equally and are annotated by 7 linguists in terms of their *overall speech proficiency* with 4.0 (good) - 5.0 (excellent) interval (step=0.1). All the linguists are adept in English teaching and are trained to unify their standard as closely as possible. Each speech is tagged by 2 linguists alternatively with inter-rater correlation ranging from 0.306 to 0.525, 0.415 in average. A recheck by a third linguist is needed in those with score distances above 0.3. Final score is obtained through simple average. For GOR analysis, these 1297 samples are also rated by *impression of prosody*. In view of the elite student group, we find that it is appropriate to tag it with *Excellent* and *Good* by 2 linguists alternatively, a third recheck is also implemented, The average agreement rate between linguists is 78.51%. These annotated speeches are randomly split into fixed set of 649 and 648 samples in training and test for 10 times, and we take the average result of the 10 experiments. From Fig 1 we can see the distribution of *overall speech proficiency* in terms of *Good* and *Excellent GOR* in this corpus. Obviously GOR is a distinctive and effective quality to grade speech with high performance.

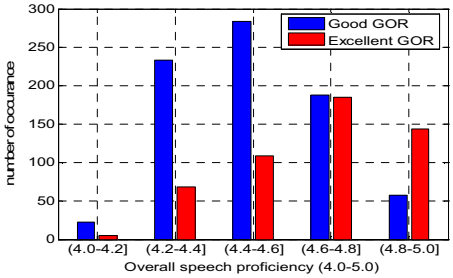

Fig.1 *Distribution of overall score of different level of GOR*

3.2 Boosting of training data

As many data-driven systems such as automatic speech recognition, boosting of training data can inevitably and effectively enhance the performance. In our selected corpus there are 14'880 speeches. How to utilize the tremendous amount of data as developed knowledge is vitally important. The key problem is that to arrange so many linguists in the work of annotation is labor consuming. Here we propose a hierarchical algorithm to filter out data with various prosodies. As depicted in the flowchart Fig.2, first a raw evaluation of pronunciation [1] and fluency [12] is implemented in 13'583 rest untagged speeches in order to discard some bad performed speeches with low GOP and GOF score, then a second classifier trained by 648 training samples of various levels of

GOR is treated as raw GOR engine to the rest samples. According to the output confidence of classifier, each sample is attached by a continuous GOR score ranging from 0 to 1. Next, we simply sort these samples in ascending order, and take those speeches with raw GOR score above 0.8 as positive training samples, less then 0.25 as negative training samples. Finally there are 3108 positive and 3551 negative samples respectively.

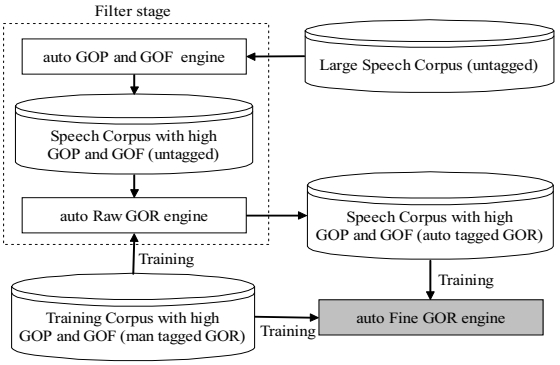

Fig.2 *hierarchical step of training by boosting untagged data (man tagged: tag by linguists, auto tagged: tag by classifier)*

After acquisition of boosting samples, there are two kinds of training resource, "Base" is the 648 training corpus with tagged GOR by linguists, and "Boost" is the boosting samples with 3108 positive and 3551 negative samples respectively. In Tab.1, the first two systems are uni-model oriented, of which system 1 is the baseline system. Provided that there are 8 different topics covering the training and testing data, system 3 and 4 are trained and tested in accordance with their corresponding topics. Another problem to be tackled in system 3 and 4 is that there is an unbalanced distribution of training examples per topic, one or two topic may contain inadequate numbers of training resource. To overcome this drawback in system 3 and 4, if the amount of training samples of a specific topic is less than 500, we simply use the "Base" training set combined with the boosting samples of that topic for training instead.

Tab.1 *Four different training schemes*
Base - 648 training corpus with tagged GOR by linguist
Boost - boosting samples (3108 positive and 3551 negative)

	System	Training resources	Number of Model	Topic Relevant
1	baseline	Base	1	NO
2	boost scheme 1	Base+Boost	1	NO
3	boost scheme 2	Boost	8	YES
4	boost scheme 2	Base+Boost	8	YES

4. Evaluation and Discussion

Performance of a CALL system can be usually justified on the grounds of two criteria:
1). EER (Equal Error Rate): Obtained from DET curves, which describes the performance of GOR for two-class (Good & Excellent in this work) discrimination when a fixable threshold varies. The equal probability of false-detect and false-alarm is the EER working point.
2). Correlation: In this work the coefficient comes from the relationship between *automatic GOR score* and the *overall speech proficiency score* ranked by linguists.

For generating confidence of GOR, we use SVM classifier, a prevalent and effective data mining tool, to construct classifier. RBF kernel method with best grid searched parameters in training set is used (C=32768, g=0.00488). Models are trained independent of genders after a normal feature scale ranging from -1 to 1. In this section we will take three experiments to test GOR performance in a reference independent context.

The first experiment aims to gain the insight into the interplay of different knowledge sources and GOR performance, we perform a two-tailed, paired t-test ($p<0.05$) on each feature to examine their significance and Tab.2 lists the p-value of the top 5 best performed features in different knowledge groups.

Tab.2 *Two-tailed, paired t-test of the best 5 performed features in each feature set*

Knowledge group	Feature	p-value
Prosodic feature (prosody representation)	Pitch Variance	8.42E-15
	Right Pause Duration	1.11E-15
	Rate of Speech	2.22E-15
	1st order formant variance in vowels	7.45E-14
	Mean Length Run	2.95E-12
Fujisaki Model (prosody generation)	Accent Amp Mean	4.93E-14
	Accent Envelop Mean	2.68E-14
	Accent Envelop Var	9.00E-13
	Accent Amp Var	1.82E-11
	Accent Count	0.188
PVI (prosody impact in variation)	IPS Pitch s-nPVI	7.89E-19
	IPS Pitch s-rPVI	6.37E-18
	IPS duration v-nPVI	3.78E-11
	IPS duration v-rPVI	6.10E-11
	Global duration v-rPVI	1.47E-10

From the analysis above, we can see that traditional prosodic features related to pitch and pause manifest more separating capacities. In the way to represent pitch variability, it can be observed that PVI operators related to pitch (IPS Pitch s-nPVI, rPVI) achieve superior performance than other prosodic features. The result of prosody generation using Fujisaki models shows that accent command coveys more information than phrase based one, hence a smaller p-value is gained. Another interesting find shows that people's impression of GOR depends more on vocalic and syllabic variation of different prosodic components than consonants'.

The second experiment is oriented in contrasting with different fusions of knowledge bases. In this section, traditional prosodic features, PVI based features, and Fujisaki model based one are calibrated to form a wide range of GOR confidence score from SVM classifier, which is examined by EER and correlation performance separately and collectively. Differences in fusion method is that "feature" uses all the generated features from different knowledge and one classifier, while "model" takes the average confidence score of different classifier trained by their corresponding knowledge. In Tab.3, some modest improvement in EER and correlation is clear with the addition of PVI based operator, and statistical qualities of Fujisaki model. An explanation for low EER and less feature count of PVI may contribute to its successful use in language clustering (stress-timed & syllable-timed), identification of native & nonnative speaker, proficient speakers with *excellent* GOR seek to demonstrate variations in pitch, energy and duration in levels of phoneme, syllable and phrase respectively, consequently large PVI is reflected. Further, various levels of PVI in different prosodic components are melted together and linguistic relevance between PVI and GOR is bridged via classifier mapping. In Fujisaki models, at first glance the 30.97% EER is the worst performance of all, but the neatly represented commands derived only from pitch and the fast computation show its advantages like efficiency in computation, less prior knowledge of input. And since Fujisaki model can reconstruct F0 from several parameters and seizes the macro and micro scope of pitch generation clue by both phrase and

accent components, it can complement to the other two systems. In fact, in our observations, speakers with good GOR tend to generate more amounts of command both in phrase and accent levels, yet more undulate accents appear, and the statistics of amplitude, duration and envelop of these commands vary.

We also compare three traditional measures derived from *integrity (*Match Rate*)*, *pronunciation* (GOP) [1], *fluency* (GOF) [12], which are all reference dependent way of scoring. Notice that in this corpus, all the speeches read by students achieve full integrity, so Match Rate between speech recognizer and reference is insignificant. Similar result can also be concluded from GOP and GOF score, which is widely used in most of state-of-the-art CALL systems, but little improvement can be seen when it comes to prosody. This is mainly due to fact that in advanced speaker levels, students in corpus have little impediment or dialect to ensure good qualities of fluency and pronunciation. When task transforms to explore and rank students between *good* and *excellent*, the pivotal role of prosody, its generation model (Fujisaki model), its compact and effective impact of tempo (PVI), have more significant effects.

Tab.3 Performance of different knowledge bases:
Corr (Correlation), RTF (Real Time Factor, including speech recognition time except for method of Fujisaki model)

Method	Fusion	EER	Corr	RTF
Match Rate (integrity)	Feature	0.434	0.058	
GOP (pronunciation)	Feature	0.399	0.127	
GOF (fluency)	Feature	0.342	0.148	
Prosodic (62 attributes)	Feature	0.200	0.381	0.029X
Fujisaki Model (8 attributes)	Feature	0.310	0.230	**0.006X**
PVI (19 attributes)	Feature	0.240	0.278	0.016X
Prosodic Feature+ Fujisaki Model (70 attributes)	Feature	0.207	0.351	0.031X
	Model	0.200	0.360	
Prosodic Feature + PVI (81 attributes)	Feature	0.188	0.401	0.030X
	Model	0.199	0.369	
PVI+ Fujisaki Model (27 attributes)	Feature	0.226	0.309	0.017X
	Model	0.268	0.279	
All (89 attributes)	Feature	**0.183**	**0.417**	0.034X
	Model	0.199	0.385	

The third experiment aims at comparing the performance of four different training schemes, and verifying whether the proposed boosting of samples and topic relevant training can take effect. In this section we simply choose "All" integrations of knowledge base in the second experiment. Fig.3 and Tab.4 depict DET curve and EER performance respectively and are comparable to the result of Tab.3 in the same way.

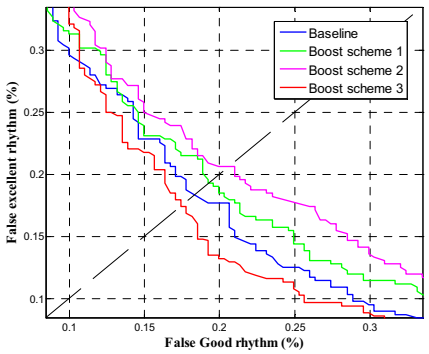

Fig.3 *DET Curve for different data boosting scheme*

From the above result we can notice that even expanding large amount of data for training (boost scheme 1), a slightly increase of EER can be observed, which mainly attribute to its *auto-tagged* part of s*amples*. This result is more significant in boost scheme 2 without "Base" training sources. But because of the diversity of the topic passages, only taking use of the "Base" data we can't attain a topic relevant training. In boost scheme 3, however, the combination of *man-tagged* and *auto-tagged* boosting data (scheme 3) brings about improvement of 0.86% in EER and 0.026 in correlation, which is a multi-model training scheme relevant to each individual topic.

Tab.4 Performance of data boosting schemes in training

Method	EER	Corr
baseline (All)	18.27%	0.417
boost scheme 1	19.26%	0.412
boost scheme 2	20.53%	0.364
boost scheme 3	**17.41%**	**0.431**

5. Conclusion

When relying simply on pronunciation and fluency score in most state-of-the-art systems, it's still difficult to achieve high correlation with human perception in advanced speakers. Due to the inherent nature of prosody, we proposed a reference independent method of multiple knowledge fusions based on its subjective impression: its representation, production model, impact of tempo variability in prosodic components. The fusion system achieved 18.27% in EER and 0.417 in correlation better than integrity, pronunciation, and fluency score. It's also noticeable that the proposed automatic boost of training method achieves another 17.41% in EER and 0.431 in correlation, which is comparable to inter-rater correlation (0.415) by linguists. Future improvement will be devoted to a more close analysis of GOR in phrase and sentence level, and mining different styles of elite patterns by co-training is also scheduled.

This work was supported by a grant from the National Natural Science Foundation of China （No. 90820303）

6. References

[1] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub. "automatic Scoring of Pronunciation Quality," Speech Communication, 30(2-3): 83 -94,1999.

[2] C. Teixeira, H. Franco, etal. "Evaluation of speaker's degree of nativeness using text-independent prosodic features". Workshop on Multilingual Speech and Language Processing, 2001.

[3] A. Maier, etal. "A Language Independent Feature Set for the Automatic Evaluation of Prosody", Interspeech 2009

[4] Tepperman, A. Kazemzadeh, and S. Narayanan, "A Text-free Approach to Assessing Nonnative Intonation", in Proc of InterSpeech ICSLP, Antwerp, August 2007.

[5] Hincks, R.. "Measuring Liveliness in Presentation Speech". Interspeech: 765-768. 2005

[6] Pudil, P., Ferri. "Floating search methods for feature selection with nonmonotonic criterion functions". In: Proc. of the 12th IAPR Intern Conf. on Pattern Recognition: 279–283, 1994

[7] Ramus, F, Nespor, M, Mehler, J. "Correlates of linguistic rhythm", Cognition, 73: 265-292. 1999

[8] Grabe, E, et al. "Durational variability in speech and the rhythm class hypothesis". In Laboratory Phonology VII, Gussenhoven C.; Warner, N. (eds.). Berlin: 515-546, 2002

[9] Antonis Botinis**.** "Intonation: analysis, modelling and technology". Text, Speech and Language Technology, 15. 2000

[10] Fujisaki, H, "Information, prosody, and modeling with emphasis on tonal features of speech", Speech Prosody 2004.

[11] Mixdorff, H. "A novel approach to the fully automatic extraction of Fujisaki model parameters". ICASSP: 1281-1284, 2000

[12] Cucchiarini, etal. "Quantitative assessment of second language learner's fluency by means of automatic speech recognition", Acoustical Society of America, 107(2): 989-999, 2000

[13] Sheng Gao, Bo Xu, "A New Framework for Mandarin LVCSR based on onepass decoder", ISCSLP:49-52, 2000