Exploring goodness of prosody by diverse matching templates

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

In automatic speech grading systems, rare research is followed through addressing the issue of GOR (*Goodness Of pRosody)*. In this paper we propose a novel method by taking the advantage of our QBH (*Query By Humming*) techniques in 2008 MIREX evaluation task. A set of standard samples related to the top-cream students are initially picked up as templates, a cascade QBH structure is then taken from two metrics: the MOMEL stylization followed by DTW distance; the Fujisaki model followed by EMD distance. Sentence GOR is obtained by the fused confidence between target and each template, and forms a weighted sum as the goodness in the passage level. Experiment results indicate that performance increases with the count of template, and Fujisaki-EMD metric outperforms MOMEL-DTW one in terms of correlation. Their combination can be treated as template based GOR score, compensated with our previous feature based GOR score, the approach can achieve 0.432 in correlation and 17.90% in EER in our corpus.
Index Terms: speech prosody, query by humming

1. Introduction

The tidal wave of technologies in speech recognition and Computer Aided Language Learning has no doubt lead to a new era of automatic assessment of speech proficiency. Although a surge of researches concerning the Goodness Of Pronunciation or Fluency (GOP, GOF) have been approached by many studies, the automatic grading of Goodness Of pRosody (GOR) has been proved to be more challenging, ascribed to such reasons: 1) Pronunciation and fluency are two basic dimensions of speech, which can generally rank speech in a whole skeleton of views, but when it comes to the high level, prosody plays a more important role; 2) There are ample researches in GOP and GOF since a series of algorithms can be learned from acoustic, language model in speech recognition.

Prosody is a crucial part of speech. People convey their emotions dominantly by prosodic variation and round-about expression in communication. In L2 learner, speech proficiency is more vulnerable to prosodic errors, such as monotonous prosody, unnecessary tonal change, etc. Although some work involved in extraction of large dimension of prosodic features and score is computed by classifiers [1], it is lack of structural explanation of prosodic production, other work engaged in comparative analysis of upper-lower boundary between the templates and target with negligence of a more fine-grained scoring process in details of prosodic units [2]. In our previous work we modeled GOR by prosodic representation, production and impact [3], but how to explain goodness of prosody from comparative analysis with good patterns is still yet to be solved.

We specifically concern the inherent nature of human raters. In acoustic view, prosodic expression in high level presents a melodious wave of phoneme duration and pitch, which is then composed in various manners, among which pitch is more significant yet more complicated to model. Here we abstract it by Momel stylization and Fujisaki model in speech synthesis, and suppose that target speech with good prosody can be covered by the standard templates. When seeking it in matching, we introduce two distance metrics, namely DTW and EMD in our system in 2008 MIREX QBH evaluation task. (rank 1st in www.music-ir.org/mirex/2008/index.php/MIREX2008_Results)

Next, Section 2 presents our prosody model and describes the metrics used in our state-of-the-art QBH system. Corpus used in the experiment is presented and evaluation based on correlation and EER contrary to human judges is demonstrated in Section 3, which is followed by conclusion in Section 4.

2. Algorithms

Fig.1 Illustrates the flowchart of the proposed algorithm, Pitch and recognition results of speech are first acquired by our automatic speech recognition engine, then error correction and stylization in pitch are then adopted to remove micro-prosodic disturbance and hypothesize the pitch level of voiceless stretches of speech, which is then followed by Momel stylization [4] and Fujisaki model extractor [5] of pitch both in training and test data. After that DTW and EMD metric in QBH is performed respectively to the stylized pitch curve and Fujisaki model. The computation is processed several times according to the template count. Final GOR is measured by the fusions of confidence score of each standard template.

One of the recent fields of content based music retrieval is QBH, the basic idea of which is to make a robust comparison metric of target and template pitch in database regardless of level shift, time warping, addition or miss of data. Sophisticated QBH methods concentrate on framed based methods such as DTW [9], which is designed to overcome the above problems based on distance measures in pitch frames. Nevertheless, the disadvantages of DTW in QBH are time-consuming and rigidly patterned. Other majorities of approach are note-based, which abstractly converts pitch contour to the imagination of music staff. In our previous work, the proposed note-based algorithm with computer vision [8] achieved comparative performance to DTW but yielded a decrease in time. More importantly, note based metric brings about the possibility to match abstract models in various spaces, like Fujisaki model in this study.

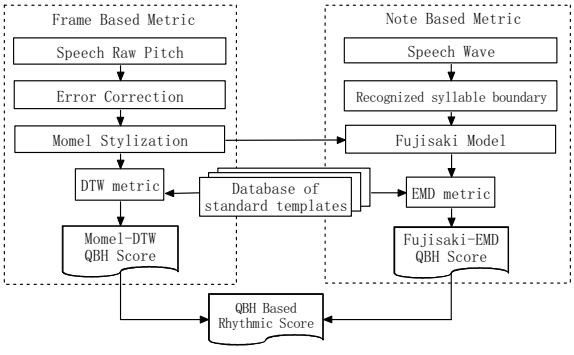

Fig.1 *Flowchart of the proposed approach*

2.1 Frame based: Momel-DTW metric

It is known to all that pitch is prerequisite for modeling rhythm, various techniques related to pitch post-process are towards reducing redundancy, thus stylization serves as such purpose to remove the errors between original and reconstructed F0.

The consideration of stylization is that when evaluating goodness of rhythm, listeners tend to intuitively appreciate the overlook of pitch tendency and rhythm skeleton, they seem to ignore unvoiced speech and perception unconsciously bridges the silent gap by filling in the missing part of the pitch contour. A popular pitch stylization method is MOMEL (*modeling melody*) by Hirst [4], which is a micro-prosody filter proved to be better than simple interpolation. It lies on the acceptation that melodic curve can be, by pieces, approximated with a best second degree polynomial. A quadratic spline aligned to target vertex along the F0 contour is reconstructed as the smoothed version that is perceptually indistinguishable from the original.

After the pitch reconstruction by MOMEL both in each templates and target, DTW is adopted to measure the similarity between the two time series, as is shown in such formula:

$$D(i,j) = d(i,j) + \min \begin{cases} D(i-2, j-1) \\ D(i-1, j-1), \\ D(i-1, j-2) \end{cases}$$

Suppose that the target pitch is represented by $t(i)$, $i = 1..\ m$, and the template pitch by $r(j)$, $j = 1..n$. $D(i, j)$ is the minimum distance starting from the left-most side ($i = 0$) of the DTW table to the current position (i, j). $d(i, j)$ is the node cost associated with $t(i)$ and $r(j)$ [9]. DTW searches the path with the least global distance from the beginning $D(0,0)$ to the ending $D(M,N)$. The best path is the one with the least global distance, which is the sum of cells alone the path.

2.2 Note based: Fujisaki-EMD metric

The extensive used Fujisaki model [5] reconstructs a given pitch contour by superimposing three components in the log F0 domain: A speaker-individual base frequency F_b, a phrase component, which results from impulse responses to impulse-wise phrase commands associated with prosodic breaks. and is described by onset time T_0, magnitude A_p and time constant α, The accent component, which results form step-wise accent commands associated with accented syllables, and is described by on- and offset times T_1 and T_2, amplitude A_a and time constant β. Typical values for α and β are 3 and 20/s respectively. The main attraction of the Fujisaki-model lies in the fact that it offers a physiological interpretation connecting F0 movements with the dynamics of the larynx, a viewpoint not inherent in any other current used pitch models.

We deem that no matter how different good rhythm patterns of target varies, as long as there is an enough coverage of templates with good rhythm, it is inevitable to find a closest candidate with a fair resemblance between their Fujisaki model. Problem is how to define such metric to take the advantages of model abstraction. We find that such homology and distinction can be directly observed by visualized graph of Fujisaki model. So thinking tools in computer vision may be a proper solution.

It has been proved that transportation distance has its remarkable merit in computer version [7]. However, we adopt a new structure of method called MSEMD (*Multi Scaled Earth Mover Distance*) for query by humming [8], which is based on an improved version of this measure. When it comes to Fujisaki model, as Fig.2 portrays, we suppose that each phrase and accent command of Fujisaki model of standard template can be represented by "suppliers" which is a number of hillocks, while the commands of target can be deemed as set of "demander", which is a mass of holes with a certain amount of capacities. But GOR problem is simpler than QBH in that target voice of one sentence and its corresponding template share the *same* information, without the necessity to "multi-scaled" with many trials to find the optimal segment. We can simply "single-scaled" target sentence with each template to the same length in order to eliminate the variation of rate of speech.

More specifically, the model abstracted in both QBH and in Fujisaki model and their differences can be seen from Table 1.

Tab. 1 *Configuration of EMD clusters in QBH and GOR task*

Parameters of EMD cluster	QBH (Note)	GOR（Fujisaki Model）
X axis in image	time (s)	time (s)
Y axis in image	music semitone	1.Ap in phrase command 2.Aa in accent command
Transportation weight occupied	beat of each music note	1. constant in phrase 2. duration of the accent

Matching score is computed by EMD, which measures the minimum flow work needed to transport earth of the hillocks to fill the total amount of holes, with which the cost considering both capacities and two dimensional positions. Such kinds of problem can be transformed into solving problem of linear programming [7]. We followed such steps in GOR problem.

Step1 (Model Extraction): Extract the Fujisaki model of both target sample and each template. Here we used a well known method by Mixdorff [6], a multistep version that exploits the Fujisaki parameter 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.

Step2 (Optimize parameter): To solve the graphic search problem in EMD, there should be an optimal balance of Y and X axis in EMD parameter. After several trails in one passage of our test corpus, we find that multiplying about 25.00 to Y axis in EMD parameter of the target is an optimal choice.

Step3 (Normalize tempo): To eliminate the impact of different rate of speech, Scale is performed in target voice based on each standard template so that they share the same information and tempo. The scale factor can be obtained from result of the ASR engine, Let N be the number of template, T_{Rki} is the duration of i th word in the k th template. T_{Dj} is the duration of j th word in the target. $N_{RecogRk}$ and N_{RecogD} is the total count of recognized word of template and target respectively:

$$scale_k = \sum_{i=1}^{N_{RecogRk}} T_{R_{ki}} \Big/ \sum_{j=1}^{N_{RecogD}} T_{D_j} \ .\ \ k = 1,2,3....N$$

Then we update all EMD clusters $(x_i, y_i, w_i)^T$ in target. The scale process of target Fujisaki model parameter is as follows:

$$x_{phrase_i} = x_{phrase_i} \cdot scale_k . \quad i = 1,2,3....N_{phrase}$$
$$x_{accent_i} = x_{accent_i} \cdot scale_k . \quad i = 1,2,3....N_{accent}$$
$$w_{accent_i} = w_{accent_i} \cdot scale_k . \quad i = 1,2,3....N_{accent}$$

Where N_{phrase} and N_{accent} are the total amount of phrase and accent commands, $X_{phrasei}$ and $X_{accenti}$ are the horizontal axis of EMD cluster in i th phrase and accent command respectively. $W_{accenti}$ is the weight parameter. Notice that we *don't* introduce any scale process in weight parameter of *phrase command.*

Step4 (Resemblance): After the above configuration of parameters, we use transportation simplex method to solve the problem and compute initial feasible solution by Russel's method [7]. After linear programming, resemblance to k th template is represented by such formula defined as the work normalized by the total flow:

$$EMD(R_k, D) = \frac{\sum_{i=1}^{N_{R_k}} \sum_{j=1}^{N_D} d_{ij} f_{ij}}{\sum_{i=1}^{N_{R_k}} \sum_{j=1}^{N_D} f_{ij}}$$

Where N_{Rk} and N_D are the total amount of Fujisaki clusters of k th template and target, d_{ij} and f_{ij} are the ground distance and flow from EMD cluster i in k th template to j in the target.

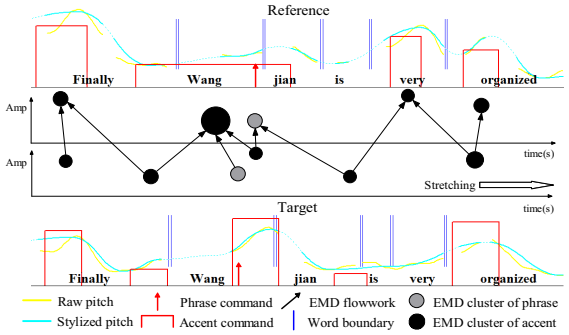

Fig.2 *Illustration of Fujisaki model applied to EMD parameter*

2.3 Sentence and passage GOR

Since such a group of templates present, it is necessary to fuse their scores to get a final decision. There are many ways in fusion, like linear regression, wrapper, etc. For our objective lies in metric comparison, we only take two simple fusion schemes. Final GOR is determined by 1) MEAN: an overall "impression" to all the templates, means every template judges the GOR of target by parallel voting strategy; 2) MIN: the best "impression" of one candidate in all the templates. To the former one, we simply use average score of QBH metric; while to the latter, the minimum metric of the templates is employed:

1) MEAN: $P(GOR_{sent} \mid D_i)_1 = \sum_{k=1}^{N} Dist_{R_{ki}, D_i} \Big/ N$

2) MIN: $P(GOR_{sent} \mid D_i)_2 = \min_k (Dist_{R_{ki}, D_i}), \quad k = 1, 2....N$

$$Dist_{R_{ki}, D_i} = \alpha_s \cdot DTW(R_{ki}, D_i) + (1 - \alpha_s) \cdot EMD(R_{ki}, D_i)$$

Where α_s is a balance factor of the above two metrics, N is the number of template, R_{ki} is i-th sentence of k-th template. D_i is i-th sentence of the target.

Even within one passage, the EER performances of its sentences differ a lot, due to the fact that position, length, vocabulary, materials vary. More importantly, some sentences are more melodic-prone to read and are more likely to perform, e.g. interrogative, imperative sentence. When taking a reversed role, one may tend to judge the GOR of passage by some representative sentences, so the overall passage GOR shouldn't be equally balanced voting of all the sentences, a feasible way we propose to overcome the drawback is to introduce a PRI (*Potential Rhythmic Importance*) weighting factor, Which is usually empirically set by linguists, A more representative sentence is potential to be assigned with a higher PRI weight, then the passage GOR is obtained as follows:

$$P(GOR_{passage} \mid D) = \sum_{i=1}^{M} W_{PRI_i} \cdot P(GOR_{sent_i} \mid D_i)$$

$$P(GOR_{sent_i} \mid D_i) = \alpha_p P(GOR_{sent_i} \mid D_i)_{QBH} + (1 - \alpha_p) P(GOR_{sent_i} \mid D_i)_{Feature}$$

Where $P(GOR_{passage} \mid D)$ is the GOR of passage D, M is the total amount of sentences in the passage, and W_{PRSi} is the PRI weight of i th sentence, $P(GOR_{sent_i} \mid D_i)$ is a weighted GOR of sentence from QBH metric based on diverse template matching and feature based method in previous work, and α_p is the balance factor.

3. Results and Discussion

3.1 Corpus and Annotation

Corpus [3] is taken from our collection of the most excellent group of Chinese students with good English speaking skills from age 14 to 16. Reading passages cover 8 topics with about 110 normal English words each. 1297 of 14880 speech samples (1 sample per student) with 90 sec are subjectively annotated by 7 linguists, all of whom are proficient in English teaching and are trained to unify their tagging standard as closely as possible. Two different scores are annotated for each sample:

1). *Overall proficiency*: interval of 4.0-5.0 (step=0.1). Each speech is annotated by 2 linguists alternatively with inter-rater correlation from 0.306 to 0.525 (0.415 in average), a recheck by a third linguist is needed in those with score distance>0.3. Final ground truth score is obtained through simple average.

2). *GOR*: these 1297 samples are also rated by prosodic impression by 2 linguists. In view of the elite group of students, Our linguists find that it is difficult to distinguish GOR with more detailed levels, but separate *Excellent(1)* from *Good(0)* is an appropriate choice. The average agreement rate between linguists is 78.51%, a third recheck is also introduced.

From Fig 3 we can see the distribution of overall score of *Good* and *Excellent* GOR in this corpus. Obviously GOR is a distinctive quality to grade speech with high performance.

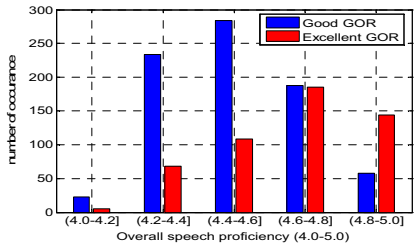

Fig.3 *Distribution of overall score of different levels of GOR*

3.2 Evaluation

Evaluation of performance of goodness of speech is usually based on two criteria: 1) Correlation: In this work we measure it between automatic GOR and ground truth overall proficiency score by linguists. 2) EER (*Equal Error Rate*): Derived from DET curve, describes the overall performance for two-class (*Good* & *Excellent* in this work) when a fixable threshold varies. Experiment is based on both sentence and passage level.

In sentence level, we plan to seek the relationship between template count and correlation. Because of the unbalanced distribution of samples per topic, linguists discuss together to choose 10 diversely styled speeches with *Excellent* GOR from the rest untagged speeches per topic as the standard templates, the 1297 tagged samples are used for evaluation. Further, we also compare the performance of different metrics in QBH. In the following results in the sentence level, passage 1 which contains 9 sentences with 162 test samples is used. 4 sentences with more than 10 words are objects of our study. The ground-truth sentence GOR is deemed to be consistent with passage GOR. e.g. If the passage GOR is tagged as *Excellent(1)*, then all its sentences belong to an *Excellent* GOR. Experiment result of sentence level in correlation is demonstrated in Fig 4.

After analyzing monotonicity in result contour we can see that the MIN score scheme achieves better performance than MEAN scheme, which indicates that it is more reasonable to consider an "optimal resemblance" in template rather than get a round impression from all the standard judges. But there are still slight turbulences in MIN scheme when template count is less than six, e.g. Sentence 3, 4. The reason may be: In QBH, all the similarity metric in each template is treated equally, but some templates can represent more targets than others, so there

is an order problem in template. If we choose a more general template in first several steps, its performance will be higher and a precarious result is gain, but after template count exceeds six, performance steadily increases. This rough increase also reflects the diversity of prosody in target. Along with the increase of template count, the chances of exploring the best resemblance enlarge. Theoretically, we can ultimately find an optimal one provided that there are enough templates equipped with various speaking styles, but it is unrealistic and a false hit of a bad performed target will occur if template count is too large. However, it is still a challenging task that concerns us all.

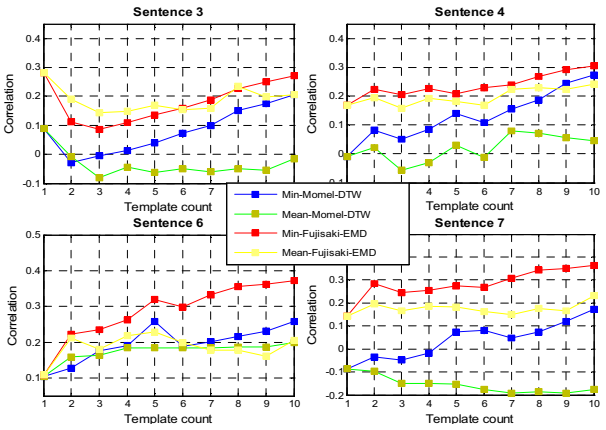

Fig.4 *Sentence GOR performances in passage 1 varies with different QBH metric, template count, and score schemes*

It is obvious that Fujisaki-EMD outperforms Momel-DTW, the reason is that EMD abstracts Fujisaki model in several clusters and avoids distance computation in pitch details, which has some drawbacks that a slight disturbance will impact the whole dynamic path. From prosodic view, DTW merely takes use of intonation information in speech, but Fujisaki model, according to its definition, is an implicit yet integral combine of phrase, intonation and even stress information.

The second experiment investigates the result of different methods in passage level. 80 samples (10 per topic) in Exp.1 are used as templates. The same 10-fold 649 samples in 1297 samples are used for test [3]. Score process is topic dependent and final result is gain by the average of the 10-fold correlation. We take the MIN scheme in both metrics, and find α_S=0.13 is an optimal choice learned by linear regression. In passage GOR computation, we also take two score schemes: 1) treat each sentence equally; 2) weighted sum by PRI of each sentence multiplied by sentence GOR. Jia's upper-lower bound method [2] is introduced in contrast. Final results are shown in Fig.5.

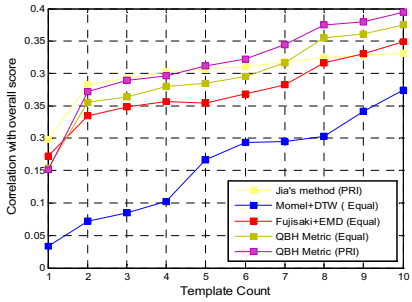

Fig.5 *Passage GOR performance in correlation*

Thanks to the average, the increase of correlation is steadier than sentence level. The fusion of both DTW and EMD metric (abbreviated as QBH metric) achieves better performance than each individual, which confirms the same result that frame-based and note-based metric can complement to each other [9].

And, passage GOR via a PRI factor attached in each sentence outperforms equal average. In comparison, Jia's method looks better than QBH metric when less templates are provided, but when template count exceeds 5, QBH metric takes the lead. A reasonable explanation is that Jia's method considers only the boundary information in template, which is coarser but gives an outline of prosody so that a few templates will take effect. However, it is more limited in that QBH seizes the inherent nature of prosodic production by investigating both intonation and stress in a dynamic time sequential way and hence a more detailed comparison is applied.

We take the best template based system (QBH Metric+PRI, 10 templates), and investigate the results of feature based method in our previous study [3] of the same test set. Although template based method lags behind (EER=21.18%, Corr=0.382) partially because of its lack of other information such as duration, amplitude, formant, etc. It is encouraging that their combination using α_P =0.22 by linear regression in training set can improve EER to 17.90% and correlation to 0.432, which is comparable to inter-rater correlation (0.415) of linguists.

Tab. 2 *EER and correlation of various algorithms*

Method	EER	Corr
Feature based	18.27%	0.417
Template based	21.18%	0.382
Feature + Template based	**17.90%**	**0.432**

4. Conclusion

Automatic evaluation of GOR is a more advanced level in CALL system. We investigate GOR in a query by humming perspective of view, sentence GOR is estimated by two metrics: Momel-DTW and Fujisaki-EMD. Their weighted combination is deemed as QBH metric and passage GOR is obtained by weighted average of sentence GOR. The performance in correlation shows an increase trend with the number of templates. After parameter setting, the best performed system achieves 0.382 in correlation and 20.18% in EER. When complemented in our previous feature based GOR score, the final EER can be reduced to 17.90% with 0.432 in correlation. Further study will focus on sentence GOR in different structure of sentence, and a more template count will also be investigated.

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

5. References

[1] Teixeira, C, Franco, H, etal. "Prosodic features for automatic text-independent evaluation of degree of nativeness for language learners", ICSLP. 2000

[2] Huibin Jia. etal. "Prosody Variation: Application to Automatic Prosody Evaluation of Mandarin Speech", Speech prosody 2008

[3] Shen Huang, etal. "Automatic reference independent evaluation of prosody using multiple knowledge fusions", Interspeech 2010

[4] Hirst, D. Espesser, R. etal. "Automatic modeling of fundamental frequency using a quadratic spline function". Travaux de l`Institut de Phon´etique d'Aix 15. 1993:71-85. Univ. de Provence.

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

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

[7] Rubner Y. "The Earth Mover's Distance as a Metric for Image Retrieval Retrieval", Tech Rep. Stanford Univ. 1998.

[8] Shen Huang, Lei wang, Bo,xu, etal . "Query By Humming Via Multiscale Transportation Distance In Random Query Occurrence Context", ICME.2008

[9] Lei Wang, Shen Huang, Bo Xu, etal. "An Effective and Efficient Method for Query by Humming System Based on Multi-Similarity Measurement Fusion", ICALIP 2008