INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Objective evaluation methods for Chinese Text-To-Speech systems

Teng Zhang[1], *Zhipeng Chen*[1], *Ji Wu*[1], *Sam Lai*[2], *Wenhui Lei*[2], *Carsten Isert*[2]

[1]Tsinghua University, Beijing, P.R.China
[2]BMW Group Technology Office China, Shanghai, P.R.China
teng-zhang10@mails.tsinghua.edu.cn, {*sam.lai*, *wenhui.lei*}@bmw.com, carsten.isert@bmw.de

Abstract

To objectively evaluate the performance of text-to-speech (TTS) systems, many studies have been conducted in the straightforward way to compare synthesized speech and natural speech with the alignment. However, in most situations, there is no natural speech can be used. In this paper, we focus on machine learning approaches for the TTS evaluation. We exploit a subspace decomposition method to separate different components in speech, which generates distinctive acoustic features automatically. Furthermore, a pairwise based Support Vector Machine (SVM) model is used to evaluate TTS systems. With the original prosodic acoustic features and Support Vector Regression model, we obtain a ranking relevance of 0.7709. Meanwhile, with the proposed oblique matrix projection method and pairwise SVM model, we achieve a much better result of 0.9115.

Index Terms: TTS evaluation, oblique matrix projection, pairwise SVM, Chinese Language

1. Introduction

TTS systems have achieved high-level maturity, which allows them to be used in daily spoken dialogue applications such as Interactive Voice Response (IVR), voice broadcasting and in-vehicle voice assistant systems. Evaluating the performance of TTS systems is not only crucial to the automobile manufacturers and users, but also a great help to TTS development.

Traditional evaluation methods for TTS systems mainly investigate synthesized speech with subjective scores given by people[1]. There are some disadvantages utilizing such methods. On one hand, if a subjective test is executed by a small number of people, the result will not be highly reliable. On the other hand, it will be time-consuming and extremely expensive if a test is required by a large number of people. As a result, an automatic judgmental approach is more feasible and practical.

Some objective methods have been developed to evaluate speech quality. One idea is to collect natural speech samples from the same speaker which is used for the synthesis inventory, and a perceptually weighted distance between the synthesized and the naturally-produced samples of this speaker can then be used as an index of the quality degradation[2]. Mariniak proposed to extract perception-based features such as Mel Frequency Cepstrum Coefficients (MFCCs) from the synthesized speech material and then compare them with features extracted from (other) natural speakers[3]. Another approach is to extract parameters from the speech signal which are related to the degradation expected for TTS[4]. A German research team proved naturalness, disturbances and temporal distortions to be the 3 significant perceptual quality dimensions of TTS systems[5].

The main shortcoming of these methods is that there is no natural speech that can be used to evaluate TTS systems in most situations. On the other hand speech quality depends more on high-layer characteristics (prosodic linguistic features such as rhythm and intonation, and subjective attribute like naturalness) rather than bottom-layer acoustic characteristics, however it's really difficult to extract those high-layer features.

In recent years, machine learning algorithms have been proposed for many difficult tasks and have achieved some outstanding results. However, there is very few machine learning based research conducted in the field of objective evaluation of TTS systems. In this paper, we propose a novel objective evaluation framework for TTS engines based on machine learning. Firstly, we generate a data set from different TTS systems and manually label it for our research, then we extract quality related features, finally a regression model can be constructed to complete the evaluation task.

Subspace decomposition method has long been used in signal analysis domain[7]. By decomposing the speech signal into four types of information, we can extract the quality related information which is extremely important to the TTS evaluation task. We exploit a subspace decomposition method based on the oblique projection algorithm and then use a pairwise based Support Vector Machine (SVM) model to evaluate TTS systems. The result show great improvement compared with the original prosodic acoustic features and Support Vector Regression model.

The rest of this paper is organized as follows. In section 2, we state the data set used in our experiment and its acquisition. Next, a baseline system using traditional acoustic features and Support Vector Regression (SVR) model is established in Section 3. Section 4 discusses the subspace decomposition method we employ to extract the quality related information in speech. Then a pairwise SVM approach is introduced to implement evaluation tasks in section 5. Section 6 conducts some experiments and evaluates the performance of the proposed system. At last, we conclude this paper and present our future work in Section 7.

2. Data acquisition

The data set used in this study is generated from different TTS systems and manually labeled for our study.

Firstly, we pick out 720 sentences from a mass of texts that are frequently used in automotive application scenarios. Chinese is referred to as monosyllabic language, which has about 400 syllables[8]. Considering some important phonetic phenomena in the synthetic speech, a greedy algorithm with random initial value is used in text selection. Compared to the completely random selection, we get a great increase in the coverage for syllable, syllable links and other linguistic phenomena. In

http://dx.doi.org/10.21437/Interspeech.2016-421

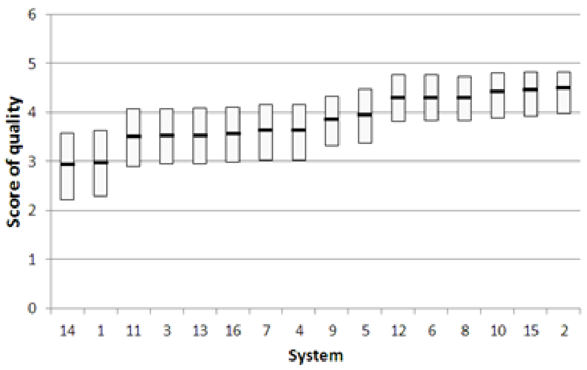

Figure 1: *Quality of TTS/person systems.*

this way, the selected sentences can easily reveal the stand or fall of the speech synthetized by TTS systems.

Table 1: *Information of TTS systems.*

No.	TTS system	No.	TTS system
1	Nuance web-tingting	6	iFLY Intp-Xiao Qi
2	iFLY vivi-Xiao Qi	7	iFLY vivi-Xiao Yu
3	Nuance 64 bit-lili	8	Nuance 32-tiantian
4	iFLY Intp-Xiao Yu	9	Microsoft-lili
5	Sino Voice	10	Nuance 64-tiantian

Then, speech samples are generated by different TTS systems including $nuance$, $iflytek$, $microsoft$, $sinovoice$ and their different versions which are shown in Table 1. In order to increase the coverage of different speech quality, we introduce human speech. 10 different TTS systems and 6 persons are used to generate a total of 11520 speech samples, whose average length is 7.34s.

Next, we design a labelling scheme to obtain the mean opinion scores (MOS) for all speech samples. In this procedure, 160 persons are chosen to label the quality of all speech samples from their subjective feelings. In Figure 1, TTS systems and real persons have been ranked in the order of smallest to largest in MOS scores. No.2, 6, 8, 10, 12 and 15 are real persons, they get the highest scores, which meets our intuition and prove the credibility of the labelling result. TTS system with the lowest scores are No.14, and the best one is No.5.

3. Baseline system

Based on previous studies on acoustic features associated with speech understanding, we employ the 1582-dimensionality feature set which have been used in the processing of emotion in speech [9]. The feature set consists of pitch related features, volume related features, tremble related features, frequency spectrum related features, Mel-scale filter bank and some other features, and functions of their mean value, variance, etc.

Using the acoustic features as input and the quality scores as output, we can construct a regression model. From the idea of machine learning, it's a straightforward way to use SVR model to achieve the evaluation goal.

However, the high-dimensional 1582-dimensionality feature set results in an obvious degradation of performance in the out-of-set test, indicating that there may be some acoustic features that are useless in the evaluating procedure. We put forward to a relevancy-based feature selection method. We first calculate the relevancy between each feature of the 1582-dimensionality set and labelling MOS scores, then select the highest N-dimensionality (N is from 100 to 1582) features as an extraction result and input them to SVR model for testing. Our experiments show that a 600-dimensionality feature set is more distinctive.

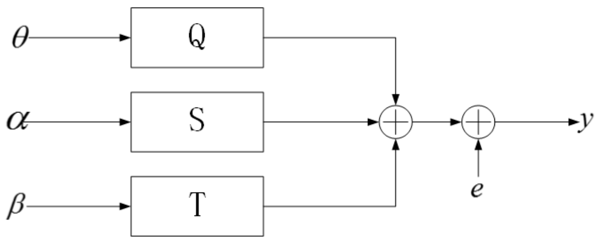

Figure 2: *The observation model of speech.*

4. Speech subspace decomposition

Subspace decompositon methods have been employed in solving various statistical problems in array signal processing[10], blind channel estimation[11] and code-division multiple access(CDMA) communications[12] to distinguish various perturbation sources.

Synthetic speech signals contain a wealth of information such as text, speaker, mood, quality and so on. It is easy to understand that quality related information which is most meaningful for TTS evaluation is only a small part of the signal. Therefore, the extraction of quality related information from synthetic speech is similar to the problem of blind source separation. In this situation, subspace decomposition method is a feasible solution. In this section, we try to use an oblique projection algorithm to extract the speech quality related information from the signal level.

We assume that the synthetic speech signal y consists of four types of information, the quality related information θ is generated from the quality coefficient matrix Q, the timbre related information α is imported through a timbre coefficient matrix S, the text related information β can be represented by a text coefficient matrix T, and of course, the additive noise e. Figure 2 is the observation model of speech signal.

$$y = Q\theta + S\alpha + T\beta + e \quad (1)$$

Assuming that Q, S and T are orthogonal with the noise signal e, but they themselves are not necessarily orthogonal between each other. In order to extract θ from y, we propose an estimation procedure based on the oblique projection algorithm. The idea of oblique projection and its application in signal processing are well known[13][14]. For the sake of establishing notations, we briefly present a few necessary definitions.

The oblique projection of a vector obtains the component of a vector in a particular direction while eliminating the component of the vector along a different direction. Specifically, for the speech observation model described as Formula (1), giving the matrices Q, S and T, the oblique projection $y_{Z_S|Z_T,Z_Q}$ of signal y can be obtained from Formula (2), θ^* is the least square solution of θ and it can be estimated as follows.

$$min_{\theta,\alpha,\beta}||y - Q\theta - S\alpha - T\beta||^2 \rightarrow y_{Z_S|Z_T,Z_Q} = Q\theta^* \quad (2)$$

Firstly we calculate $P_S^{\perp}$, the orthogonal projection matrix of S, using Formula (3).

$$P_S^{\perp} = I - S(S^TS)^{-1}S^T \tag{3}$$

Then multiply Formula (1) by $P_S^{\perp}$, we can get Formula (4).

$$P_S^{\perp}y = P_S^{\perp}Q\theta + P_S^{\perp}T\beta \tag{4}$$

Next we calculate $P_T^{\perp}$, the orthogonal projection matrix of $P_S^{\perp}T$, using Formula (5).

$$P_T^{\perp} = I - P_S^{\perp}T((P_S^{\perp}T)^TP_S^{\perp}T)^{-1}(P_S^{\perp}T)^T \tag{5}$$

Finally multiply Formula (4) by $P_T^{\perp}$, we can get θ^* as Formula (6).

$$\theta^* = (Q^TP_T^{\perp}P_S^{\perp}Q)^{-1}Q^TP_T^{\perp}P_S^{\perp}y \tag{6}$$

However, the matrices Q, S and T are not easy to get, so we propose an iterative process to estimate them.

In the beginning, we initialize them in a simple way. For example, the timbre of each speech represents which TTS system it belongs to, so we assume that the timbre related information α is a supervised embedding low dimensional vector according to the system labels of all samples. Implementing the Linear Discriminant Analysis (LDA) method[15] with Formula (7) and system labels, we can get the generalized inverse matrix of S_{init} just using the projection matrix of LDA result. The same process can be done to the quality matrix Q_{init} and text matrix T_{init} according to the quality scores and textual information as class labels. Then the corresponding information θ, α and β can be estimated using Formula (6).

$$S_{init}\alpha^{'} = y \tag{7}$$

Next S, T and Q can be re-estimated using Formula (8) and LDA method. When the iterative process ends, θ^* is the quality related information we extract from the synthetic speech signal y. The algorithm can be seen below.

$$S_{next}\alpha^{'} = y - Q_{now}\theta^* - T_{now}\beta^* \tag{8}$$

Algorithm 1 Speech subspace decomposition algorithm

```
init $S_{init}$, $Q_{init}$ and $T_{init}$ as formula (7) using LDA method
for each $i \in [1, iteration\_number]$ do
    estimate $\theta^*$, $\alpha^*$ and $\beta^*$ as formula (6) using oblique projection method
    estimate $S$, $T$ and $Q$ as formula (8) using LDA method
end for
return $\theta^*$
```

5. Pairwise SVM model

The traditional machine learning based method to evaluate the quality of TTS systems is to predict the speech quality using the global acoustic features and a regression model, experiments in [6] have shown that the Support Vector Regression (SVR) model is the best. However, the performance of SVR model is mainly restricted by the following two aspects. The acquisition and labelling of synthesized speech data requires heavy workloads, which reduces the expression ability of data-driven machine learning methods. Meanwhile, the labelling results mainly depend on persons' subjective feeling, which is not that credible. In order to reduce the negative influence of these two phenomena, we exploit the pairwise idea.

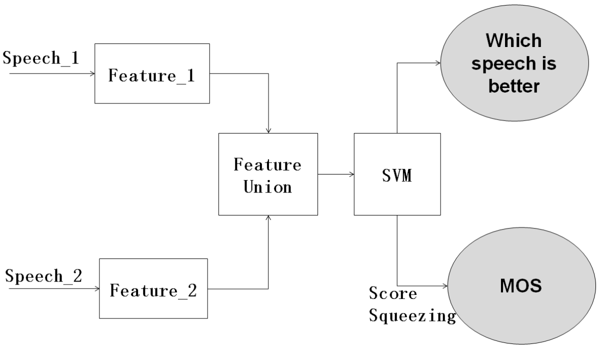

Figure 3: *The pairwise SVM model.*

As mentioned in section 2, we generate speech samples from 16 different sources (including 10 different TTS systems and 6 persons) and 720 different sentences. For 16 synthesized speech samples generated from different sources but for the same textual sentence, we can get all combinations of pairs, which is 16*15=240 in total. This procedure can be done to all the 720 sentences, and finally we get a total of 172800 speech pairs, which can be used as our training corpus.

In the Feature Union step, there are two methods that can be used. The difference method means that we use the difference of two input utterances as the input features of the SVM model, which can be shown as $\theta = \theta_1 - \theta_2$; the combination method means that we connect two input utterances as the input features of the SVM model, which can be shown as $\theta = [\theta_1; \theta_2]$.

To improve the credibility of our labels, we change the scores to be binary labels, which is defined to be 1 if the former speech is better than the latter one, and 0 if the former speech is worse than the latter one. Thus a pairwise SVM model is constructed as figure 3.

In the pairwise SVM model, we input two speech utterances, and output which one is better. Using a straightfoward score squeezing method, we can finally get the MOS of each input speech from the scores of giving speech samples and the output of the pairwise SVM model.

6. Experiments

We conduct several experiments to verify the performance of our methods. In the feature extraction step, we use the 1582-dimensionality feature set and the subspace decomposition features separately. In the model training step, we contrast the straightfoward SVR model and the pairwise SVM model.

The database used for this study consists of 11520 utterances generated from 10 TTS systems and 6 persons using 720 text sentences. In our experiments, we design a round-robin test as the out-of-set test. Each time, we select one system as the out-of-set system. The other 15 systems are used for training. When testing, the trained model is utilized for testing of all 16 systems.

6.1. Evaluation Index

We use the following evaluation index to measure the performance of our models.

- Relevancy of system ranks (hereinafter referred to as R):

$R = \frac{Cov(\overline{predict_i}, r_i)}{|\overline{predict_i}||r_i|}$, where $\overline{predict_i}$ is the rank result of the ith TTS system given by our algorithm, and r_i is the system rank based on the labelling MOS for the same TTS system. When $\overline{predict_i}$ and r_i are exactly the same, R equals 1, when they are completely reverse, R equals -1. Larger R indicates that $\overline{predict_i}$ is a better prediction of r_i.

6.2. Feature selection

We use the $OpenSmileToolkit$[16] to extract the fundamental 1582-dimensionality features (named $features\_1582$) and then train a SVR model to predict the quality scores of speech utterances. Next, we select a more distinctive 600-dimensionality feature set (named $features\_600$) and train the SVR model again. This is our baseline system.

6.3. Model Design

From Formula (1), we assume that the speech signal consists of four types of information, in which only the quality information θ is what we want to extract. However, these four types of information are more or less overlapped, thus the subspace decomposition procedure may eliminate useful information unknowingly. Using different oblique projection strategies, we can eliminate different interference components. For example the Formula (6) eliminates the timbre information and text information in turn, we can also eliminate the timbre information only using Formula (5). In this section, we conduct four correlative experiments to verify whether our assumption is reasonable.

- eliminates the timbre information only (named $features\_s$): $\theta = (Q^T P_S^{\perp} Q)^{-1} Q^T P_S^{\perp} y$
- eliminates the text information only (named $features\_t$): $\theta = (Q^T P_T^{\perp} Q)^{-1} Q^T P_S{}^{\perp} y$
- eliminates the text and timbre information in turn (named $features\_ts$): $\theta = (Q^T P_S^{\perp} P_T^{\perp} Q)^{-1} Q^T P_S^{\perp} P_T^{\perp} y$

In the pairwise SVM model, there are two methods can be used to combine features of two input speech utterances, the difference method and the combination method, which has been mentioned in section 5.

In summary, we conduct five contrast experiments using different features and different models. More details can be seen in Table 2.

Table 2: *Details of all experiments.*

No.	Features	Model	Pairwise method
$baseline$	features_600	SVR	
1	features_600	pairwise SVM	difference
2	features_s	pairwise SVM	difference
3	features_t	pairwise SVM	difference
4	features_ts	pairwise SVM	difference
5	features_ts	pairwise SVM	combination

6.4. Test Results

Using the evaluation index described in Section 5.1, we can get the results of our experiments as Table 3 and Table 4.

Table 3 show that the selected 600-dimensionality feature set performs much better than the original 1582-dimensionality feature set, with a relative improvement of 18.0%.

Table 3: *Results of Feature Selection.*

Method No.	R	relative improvement
$features\_1582$	0.6535	
$features\_600$	0.7709	$+18.0\%$

Table 4: *Results of different methods.*

Method No.	R	relative improvement
$baseline$	0.7709	
1	0.7830	$+1.6\%$
3	0.7964	$+3.3\%$
3	0.8054	$+4.5\%$
4	0.9006	$+16.8\%$
5	0.9115	$+18.2\%$

Table 4 show the results of our contrast experiments. The results of experiment 1 are slightly better than the baseline with an improvement of 1.6% relatively, suggesting that the pairwise model is better than the SVR model. The results of experiment 2 and 3 show that the oblique projections on timbre and text are all helpful for the extraction of quality related information, achieving a relative improvement of 3.3% and 4.5% respectively. In experiment 4, we use the oblique projection features that eliminate both the text and timbre information in turn, and achieve a significant improvement of 16.8% relatively. When we use the combination method instead of the difference method in the pairwise SVM model, the relative improvement increases to 18.2%. These results show the rationality of our assumption. The oblique projection method and pairwise SVM model work pretty well in the TTS evaluation task.

7. Conclusions

The difficulty of speech quality evaluation mainly lies in following two aspects. First, the speech signal contains a lot of speech quality irrelevant information, which will still be preserved after using traditional speech processing methods. Therefore, it is very difficult to extract the speech quality related information on signal level. Second, we only have 10 TTS systems, which is too few to cover the characteristics of synthesis speech generated by different TTS systems. It is a very big challenge for traditional machine learning methods to automatically learn the speech quality related global information from the synthesis speech of limited TTS systems and rough labels.

In this paper, we assumed that the speech signal consist of different types of information and used the oblique projection method to solve the problem of blind source separation. We also combined the pairwise idea and traditional machine learning methods to construct a pairwise SVM model to improve the performance of our system. And at last we got a ranking relevance of 0.9115, which was much better than the baseline result of 0.7709.

However, there is still a lot of work to do on signal level, which is more important but also be more difficult as well. Of course, the popular deep learning methods are also a potential solution if we can collect more data.

8. Acknowledgements

This work is a joint project of Tsinghua and BMW, and supported by BMW China Service Ltd.

9. References

[1] F. Hinterleitner, G. Neitzel, S. Möller, and C. Norrenbrock, "An evaluation protocol for the subjective assessment of text-to-speech in audiobook reading tasks," in *Proceedings of the Blizzard challenge workshop, Florence, Italy*. Citeseer, 2011.

[2] M. Cernak and M. Rusko, "An evaluation of synthetic speech using the pesq measure," in *Proc. European Congress on Acoustics*, 2005, pp. 2725–2728.

[3] A. Mariniak, "A global framework for the assessment of synthetic speech without subjects," in *Third European Conference on Speech Communication and Technology*, 1993.

[4] T. Falk and W. Chan, "Single ended method for objective speech quality assessment in narrowband telephony applications," 2004.

[5] F. Hinterleitner, S. Möller, C. Norrenbrock, and U. Heute, "Perceptual quality dimensions of text-to-speech systems," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[6] C. R. Norrenbrock, F. Hinterleitner, U. Heute, and S. Möller, "Towards perceptual quality modeling of synthesized audiobooks-blizzard challenge 2012," in *Blizzard Challenge Workshop 2012*. Citeseer, 2012.

[7] Z. Xu, "Perturbation analysis for subspace decomposition with applications in subspace-based algorithms," *Signal Processing, IEEE Transactions on*, vol. 50, no. 11, pp. 2820–2830, Nov 2002.

[8] D. V. Klein, "Foiling the cracker: A survey of, and improvements to, password security," in *Proceedings of the 2nd USENIX Security Workshop*, 1990, pp. 5–14.

[9] F. Eyben, A. Batliner, and B. Schuller, "Towards a standard set of acoustic features for the processing of emotion in speech." in *Proceedings of Meetings on Acoustics*, vol. 9, no. 1. Acoustical Society of America, 2012, p. 060006.

[10] R. Roy, A. Paulraj, and T. Kailath, "Estimation of signal parameters via rotational invariance techniques-esprit," in *30th Annual Technical Symposium*. International Society for Optics and Photonics, 1986, pp. 94–101.

[11] E. Moulines, P. Duhamel, J.-F. Cardoso, and S. Mayrargue, "Subspace methods for the blind identification of multichannel fir filters," *Signal Processing, IEEE Transactions on*, vol. 43, no. 2, pp. 516–525, 1995.

[12] S. E. Bensley and B. Aazhang, "Subspace-based channel estimation for code division multiple access communication systems," *Communications, IEEE Transactions on*, vol. 44, no. 8, pp. 1009–1020, 1996.

[13] R. T. Behrens and L. L. Scharf, "Signal processing applications of oblique projection operators," *Signal Processing, IEEE Transactions on*, vol. 42, no. 6, pp. 1413–1424, 1994.

[14] C. R. Rao, S. K. Mitra *et al.*, "Generalized inverse of a matrix and its applications."

[15] B. Scholkopft and K.-R. Mullert, "Fisher discriminant analysis with kernels," *Neural networks for signal processing IX*, vol. 1, no. 1, p. 1, 1999.

[16] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=944919.944968