An Objective Measure for Estimating MOS of Synthesized Speech

Min Chu and Hu Peng
Microsoft Research China
Beijing, P.R.C
minchu@microsoft.com; h-peng@microsoft.com

Abstract
This paper proposes an average concatenative cost function as the objective measure for naturalness of synthesized speech. All its seven component-costs can be derived directly from the input text and the scripts of speech database. A formal Mean Opinion Score (MOS) experiment shows that the average concatenative cost and its seven components are all highly correlated with MOS obtained subjectively. The correlation coefficient between the objective measure and subjective measure is -0.872. The mean of errors in MOS estimation for individual waveforms is 0.32 with 0.40 RMSE. When estimating the overall MOS for TTS systems, the mean error is smaller than 0.05. With the proposed objective measure, it becomes possible and easy for us to track the performance in naturalness regularly. The proposed cost function could also serve as criteria for optimizing the algorithms for unit selecting and speech database pruning.

1. Introduction
Evaluating the quality of synthesized speech contains two aspects, intelligibility and naturalness. Intelligibility is not a big issue for most current TTS systems. Yet, the naturalness of synthesized speech is still far from people’s expectation. To have regular evaluations on the improvements in naturalness becomes very important for system development, particularly for research on high quality systems. The Mean opinion score (MOS) is one of the most popular and widely accepted subjective measures for naturalness. However, running a formal MOS evaluation is expensive and time consuming. It is almost impossible to run MOS assessments for every algorithmic change or idea. Several objective measures[1], such as Signal to Noise Ratio, Bark Spectral Distortion and Mel Scale Distance, had been used to estimate subjective quality in speech coder[2] and voice communication systems[3][4]. However, they are not suitable for evaluating the quality of speech from TTS systems, since no original waveforms exist as references when synthesizing speech from text. In this paper, the average concatenative cost, \( C_o \), which can be directly derived from the input text and the scripts of speech database, is proposed as a new objective measure for speech quality. Results from simulation experiments reveal that \( C_o \) is highly correlated with MOS obtained subjectively. The correlation coefficient between the two is -0.872.

The paper is organized as follows. Section 2 gives the definition of the average concatenative cost function. A formal MOS experiment, the results and discussions, the application of the COST-MOS curve and the optimizing of weights in the cost function are presented in Section 3. Finally, conclusion of our major findings and the outline for the future studies are given in Section 4.

2. Average concatenative cost
In conventional concatenative TTS systems, pitch and duration modification algorithm, such as PSOLA, is applied to pre-stored units to guarantee that the prosodic features of synthetic speech meet the predicted target values. These systems have the advantages of flexibility in controlling the prosody. Yet, they often suffer from significant quality decrease in naturalness. In the Mandarin TTS system developed at MSRCN[5], speech is generated by directly concatenating syllable segments without any pitch or duration modification under the assumption that the speech database contains enough prosodic and spectral varieties for all synthetic units and the best fitting segments can always be found.

For constructing a high quality Mandarin TTS system, a very large speech corpus, containing 12,000 sentences and 190,000 syllables, has been collected. It covers about 64% of possible variations of all units[5]. Using a good criterion for finding the best fitting unit from the speech corpus becomes crucial for generating high quality speech. In our case, units that minimize the concatenative cost, \( C_i \), of a whole utterance are selected. \( C_i \) is given by (1),

\[
C_i = W_1 \sum_{s} D_s(l) + W_2 \sum_{c} C_c(l)
\]

where \( L \) is the number of syllables in the utterance. \( C_c(l) \) is the smoothness cost between the \( l \)th syllables and its succeeding neighbor. If two concatenated units are continuous segments in source corpus, 0 is assigned to \( C_c \). Otherwise, 1 is assigned. \( D_s(l) \) is the contextual distance for the \( l \)th syllable and is defined by (2),

\[
D_s(l) = \sum_{t} W_t D_t(l)
\]

where \( D_t(l) \) is the distance between the \( l \)th factor in source contextual vector and that in the target vector. \( W_t \) is the weight for \( D_t(l) \). I is the total number of factors considered. In our case, six factors, which are position in phrase, position in word, left phonetic context, right phonetic context, left tone context and right tone context, are considered as factors in the contextual vector. Since all these factors take only categorial values, the distance between categories are predefined according to expert knowledge.

In order to make the concatenative cost comparable among utterances with variable number of syllables, the average
concatenative cost of an utterance is proposed in this paper. The definition of average concatenative cost is given by (3) to (5).

$$C = \sum_{i=1}^{7} W_i C_w$$  \hspace{1cm} (3)

$$C_w = \frac{1}{L} \sum_{i=1}^{L} D_i, \quad i = 1, \ldots, I$$  \hspace{1cm} (4)

$$W_i = \begin{cases} W_1, & i = 1, \ldots, I \\ W_2, & i = I + 1 \end{cases}$$  \hspace{1cm} (5)

where, $C_w$ is the average concatenative cost and $C_w$ ($i = 1, \ldots, 7$) are the seven factors contribute to $C_w$, which are the average costs for position in phrase, position in word, left phonetic context, right phonetic context, left tone context, right tone context and smoothness. $W_i$ are weights for the seven component-costs and all are set to 1.

Since the MSRCN Mandarin TTS system based on above unit selecting criterion has demonstrated good performances on speech quality, $C_w$ is believed to highly relate with the naturalness of synthetic utterance. It is proposed as an evaluation criterion for naturalness in this paper.

3. COST-MOS curve

Since $C_w$ is believed to reflect the naturalness of synthesized speech, a formal subjective experiment is done to investigate the relationship between $C_w$ and MOS.

3.1. MOS experiment

For getting a good picture for the relationship between $C_w$ and MOS, 100 sentences are carefully selected from a 200MB text corpus so that $C_w$ and $C_w$ ($i = 1, \ldots, 7$) of them are scattered into wide spans. The average lengths of these sentences are 16.6 syllables. Four waveforms are generated for each sentence by the MSRCN Mandarin TTS engine with four speech databases, whose sizes are 1.36GB, 0.9GB, 0.38GB and 0.1GB respectively. $C_w$ and $C_w$ of each waveform are calculated.

All the 400 synthesized waveforms, together with some waveforms generated from other TTS systems and waveforms uttered by a professional announcer, are randomly played to 30 subjects. The subjects are asked to score the naturalness of each waveform from 1 to 5 (1=Bad, 2=Poor, 3=Fair, 4=Good, 5=Excellent). The mean of the 30 scores for a given waveform represents its naturalness in MOS.

50 original waveforms uttered by the speaker who provides voice for the speech database are used in the MOS experiments. The average MOS for these waveforms is 4.54, which provides an upper bound for MOS of synthetic voice. Providing subjects a wide range of speech quality by adding waveforms from other systems in the experiment is helpful for subjects to make good judgments on the naturalness. However, only the MOS for the 400 waveforms generated from MSRCN TTS system are considered in our study.

3.2. COST-MOS curve

The 400 waveforms are plotted as small points in Fig. 1, where the horizontal axis is the objective measure, $C_w$, and the vertical axis is the subjective measure, MOS. The correlation coefficient between the two dimensions is –0.822, which reveals that the cost function replicates, to a great extent, the perceptual behavior of human beings. The minus sign of the coefficient means that the two dimensions are negatively correlated. The larger $C_w$ is, the smaller the corresponding MOS will be. A linear regression trend line is estimated by calculating the least squares fit through all points and is plotted in Figure 1 as the solid curve. The curve is denoted as COST-MOS curve and its expression is also given in Figure 1.

![Figure 1: Objective measure vs. subjective measure for the 400 waveforms and the trend-line between the two measures.](image)

According to the COST-MOS curve, an estimated MOS, denoted as $\hat{MOS}$, can be derived from each $C_w$. The error for estimation ($\hat{E}_{es}$) and root mean squared error (RMSE) are defined by (6) and (7)

$$E_{es} = \left| MOS - \hat{MOS} \right|$$  \hspace{1cm} (6)

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} E_{es}^2}$$  \hspace{1cm} (7)

where N is the total number of waveforms.

The distribution of $E_{es}$ is shown in Figure 2. The mean error for all samples is 0.36 and RMSE is 0.47. 73.3% of samples have errors smaller than 0.5 and 96.5% of samples have errors smaller than 1.

3.3. Application of the COST-MOS curve

Though, there may have large errors for individual waveforms when estimating MOS from the COST-MOS curve, the curve performs well when estimating the overall MOS of a TTS system.

As mentioned in Section 3.1, four waveforms are generated for each of the 100 sentences by changing the size of speech database. The four speech databases used are labeled as Db1, Db2, Db3 and Db4 respectively. Db1 keeps all segments in the large speech corpus. When some of non-frequently used
segments are pruned out of database step by step, three other databases are formed. The difficulty for finding suitable units for concatenation increases when the size of speech database is reduced. As a result, the naturalness of synthesized speech decreases accordingly.

The average of MOS for the 100 waveforms generated from each database and the average of the estimated MOS for each database are list in Table 1. The errors for estimation are also list in Table 1. All these errors are smaller than 0.05 (in MOS).

![Figure 2: The distribution of errors in estimating MOS from the average concatenative cost.](image)

**Table 1:** The size, average MOS, estimated MOS and error for estimation for the four speech databases.

<table>
<thead>
<tr>
<th>Database label</th>
<th>Db1</th>
<th>Db2</th>
<th>Db3</th>
<th>Db4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size (GB)</td>
<td>1.38</td>
<td>0.9</td>
<td>0.38</td>
<td>0.1</td>
</tr>
<tr>
<td>Average MOS</td>
<td>2.93</td>
<td>2.91</td>
<td>2.86</td>
<td>2.36</td>
</tr>
<tr>
<td>Estimated MOS</td>
<td>2.94</td>
<td>2.94</td>
<td>2.87</td>
<td>2.31</td>
</tr>
<tr>
<td>Error</td>
<td>0.01</td>
<td>0.03</td>
<td>0.01</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Since the 100 sentences used in the MOS experiment are deliberately selected to cover a wide range of variance in $C_s$, the distribution of $C_s$ over the 100 sentences is not identical to the distribution of $C_s$ when synthesizing real text. The average $C_s$ for the four databases are calculated over a 183MB text corpus (containing 2558349 sentences), which are randomly selected from newspapers. The average $C_s$ for the four databases and the MOS derived from them are list in table 2. The estimated MOS in table 2 are higher than those in table 1, since the percentage of low cost sentences in real text is higher than that in the selected sentences in the MOS experiment. These MOS should represent the naturalness of speech synthesized from real text based on the four databases. To minimize the average $C_s$ over a large amount of text corpus could serve as a good criterion for evaluating algorithmic changes and new ideas during system development.

According to table 2, when the size of database changes from 1.38GB to 0.38GB, very small decrease of naturalness is perceived. However, when the size of database is reduced to 0.1GB, significant decrease in naturalness is perceived. Our current pruning algorithm for speech database is still simple. Farther improvements are needed. Minimizing the average $C_s$ over a large amount of text corpus could also serve as the criterion for evaluating database-pruning algorithms.

**Table 3:** Correlation coefficients between MOS and $C_s$ and its the seven component-costs, $C_{ai}$.

<table>
<thead>
<tr>
<th>$C_{ai}$</th>
<th>$C_{aq}$</th>
<th>$C_{at}$</th>
<th>$C_{av}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.822</td>
<td>-0.453</td>
<td>-0.617</td>
<td>-0.694</td>
</tr>
<tr>
<td>-0.714</td>
<td>-0.685</td>
<td>-0.650</td>
<td>-0.847</td>
</tr>
</tbody>
</table>

**Table 4:** The optimized weights in cost function.

<table>
<thead>
<tr>
<th>$W_1$</th>
<th>$W_2$</th>
<th>$W_3$</th>
<th>$W_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.60</td>
<td>0.10</td>
<td>0.76</td>
</tr>
<tr>
<td>$W_5$</td>
<td>$W_6$</td>
<td>$W_7$</td>
<td></td>
</tr>
<tr>
<td>1.76</td>
<td>0.72</td>
<td>2.96</td>
<td></td>
</tr>
</tbody>
</table>

### 3.4. Optimizing weights in cost function

The correlation coefficients (denoted as CR) between MOS and $C_s$ and its seven component-costs are list in Table 3. $C_{aq}$, the cost for smoothness, has the highest CR. $C_{at}$, the cost for position in phrase, has the lowest CR. The $C_s$ calculated in previous sections is obtained by setting equal weights for all seven components. However, in order to get better correlation between MOS and $C_s$, it is reasonable to assign a larger weight to the component with higher CR and a smaller weight to the component with lower CR. In this section, the weights, $W_i$ in equation (3) are optimized by maximizing the correlation coefficients between $C_s$ and MOS. The optimized weights are listed in Table 4. After recalculating $C_s$ for the 400 waveforms using the new weights, the CR between $C_s$ and MOS raises to -0.872. Stronger correlation is obtained. Again, the minus sign means negative related.

It should be noted that the value of $C_s$ increases for all waveforms after the optimization, because the weight for $C_{aq}$.
increases a lot and \( C_{\alpha} \) often takes larger values than other components. However, the correlation between \( C_{\alpha} \) and MOS is increased so that the average error for estimating MOS is reduced. An update picture for \( C_{\alpha} \) vs. MOS of the 400 waveforms is plotted in Figure 3. Again, a new COST-MOS curve and its expression are plotted in the same figure. In estimating MOS from the new COST-MOS curve, the mean error for 400 waveforms is reduced to 0.32 and RMSE is reduced to 0.40. 81.3% of errors are smaller than 0.5 and 98.3% of errors are smaller than 1. So, the optimized cost function is a better replication of the perceptual behavior of human being.

\[
y = -0.8149x + 4.7385
\]

4. Conclusion

This paper proposes an average concatenative cost function, which is defined as the weighted sum of seven sub-costs, as an objective measure for naturalness of synthesized speech. All the seven sub-costs could be derived directly from the input text and the scripts of speech database. A formal MOS experiment shows that the average concatenative cost and its seven components are all highly correlated with MOS obtained subjectively. A linear regression trend-line is estimated by calculating the least squares fit through all points. The resulted COST-MOS curve can be used to estimate MOS from average concatenative cost. The mean of errors in MOS estimation for the 400 waveforms used in the MOS experiment is 0.36 and the RMSE is 0.47. Though, there may have large errors for individual waveforms when estimating MOS from the COST-MOS curve, the curve performs well when estimating the average MOS of a TTS system. All the errors are smaller than 0.05 (in MOS) when estimating the average MOS for the waveforms generated from the four databases. It becomes possible and easy for us to track the performance in naturalness regularly with the proposed objective measure. The proposed cost function could also serve as criteria for optimizing the algorithms for unit selecting and speech database pruning in future studies.

After optimizing the weights for the seven components in the cost function, the correlation coefficients between MOS and \( C_{\alpha} \) increased to -0.872. In addition, the average error and RMSE in MOS estimation is reduced to 0.32 and 0.40 respectively. Since the optimized cost function simulates the perceptual behavior of human being better, it should be a more suitable criterion for unit selection in the TTS system.

5. Acknowledgments

The authors would like to thank Lin He, Bei Wang and Rui Zhang for organizing the MOS experiment. Thanks also go to everyone who takes part in the experiment.

6. References