TR-IT-0304

Objective Distance Measures for Assessing Concatenative Speech Synthesis

Jing-Dong Chen & Nick Campbell

Abstract

This report contains two parts. In the first part, Several different acoustic transformations of the speech signal are compared for use in the assessment and evaluation of concatenative speech synthesis. The transformations tested include the LPC, LSP, MFCC, residual MFCC, bispectrum, Mellin transform of the log spectrum, Wigner-Ville distribution, etc. The computed distances between a synthesized utterance and a naturally spoken version of the same sentence are compared by correlation with perceptually-based scores obtained from a MOS evaluation. The results show that the distances computed using the bispectrum have the highest degree of correlation with the MOS score. Both the RMFCC and the LPC outperform the MFCC and the LPCC. The Wigner-Ville distribution based cepstrum is found to behave poorly in this task.

A five-level-score evaluation method based on a technique called sluggish coding is proposed in the second part. The experimental results show that with the use of sluggish coding, the method can change a distance obtained from the DTW to a five-level-score which is revealed to have high correlation with the MOS score.

Acknowledgement

This work is supported by the Trust Agency of the Japan Key Technology Center (KTC) and is carried out in Department II of ATR-ITL. Our gratitude must be expressed to many officials in the KTC for supporting this work.

We have benefited from advice and suggestion of many people. Among whom we especially thank Seiichi Yamamoto, the president of ATR-ITL, Yoshinori Sagisaka, the head of Department I of ATR-ITL, Professor Kuldip K. Paliwal in Griffith University, Yoshinori Michijiri, Kazuyuki Ashimura, Masahiro Nishimura, Ken Fujisawa, Shuwu Zhang and all members in Department II of ATR-ITL.

Part I. Objective Distance Measures by Using Different Acoustic Transformations and DTW

I. Introduction

With advances in the technology of computer memory, computing power, and speech synthesis, the quality of synthetic speech is continually improving. Much attention is now being devoted to the assessment and evaluation of the quality of the synthetic speech.

The evaluation measures can be generally classified into subjective and objective methods. Subjective methods require human listeners to judge the speech quality, which may evaluated for intelligibility, naturalness, voice pleasantness, liveliness, friendliness, etc., but individual subjects can perform differently when attempting the same task of synthetic speech assessment.

The Mean Opinion Score (MOS) is a standard method for evaluating speech coding, and is also being used to measure the quality of synthetic speech. However, the fact that the MOS score needs the expertise of human listeners causes the subjective evaluation process to be lengthy and expensive. This motivates many researchers to investigate automatic objective measures which are expected to provide results in agreement with the subjective measures. Physical distance measures between speech waveforms are often used for this purpose.

Previous works compared several different distance measures for speech recognition and showed that recognition performance varied according to the features used [1][11]. For speech synthesis, relatively little research has been performed on this topic, but recent developments in concatenative speech synthesis have used objective distance measures when segments are selected from a large speech corpus. Since the purpose of the unit selection is to locate segments that will make the synthetic speech sound natural, much effort has been devoted to finding the relation between objective distance measures and the perceptual impressions. In a recent study [2], several commonly-used distance measures, such as FFT-based cepstra, LPC-based cepstra, line-spectral pairs (LSP), log area ratios (LAR), and a symmetrized Itakura distance were compared. The results revealed that transforms which use frequency warping had a higher degree of correlation with the perceptual data than those did

not, and that the MFCC and Itakura distance achieved the highest correlation coefficient among the measures investigated.

This report will focus on the comparison of different distance measures for evaluating the quality of synthetic speech. The evaluation framework is based on comparison between synthesized utterances and original speech waveforms using feature representation and dynamic time warping.

The correlation between the computed distances and the MOS score is used to determine which transform performs the best discrimination. The results show that a measure based on the bispectrum has the highest degree of correlation with the perceptual data. While the distance measure based on the Wigner-Ville distribution is revealed to perform poorly for our task.

II. Different features of speech signal

1. Linear prediction coefficients

A linear predictive (LP) analysis of a speech signal is based on the all-pole model. This model assumes that a speech sample is a weighted linear combination of P previous samples, i. e.,

$$s(n) = \sum_{i=1}^{P} a_i s(n-i) + e(n) \tag{1}$$

where $s(n)$ is speech signal, $e(n)$ is the prediction error, and a_i are weights applied to previous speech samples. The weights correspond to the coefficients of an all-pole filter with transfer function

$$H(Z) = \frac{1}{A(z)} = \frac{1}{1 - \sum_{i=1}^{P} a_i z^{-i}} \tag{2}$$

where $A(z)$ is a nonrecursive LP analysis filter. $H(Z)$ captures the vocal tract information, hence the linear prediction coefficients (LPC) a_i are widely used in speech processing such speech recognition and speech synthesis.

The LPCs are estimated by minimizing the mean square of the prediction error. In this report, the autocorrelation method is used to perform such task.

2. Linear prediction cepstral coefficients

For speech recognition and distance measures, the linear prediction cepstrum is more often used than linear prediction coefficients themselves. The cepstrum is defined by the inverse transform of log-spectrum

$$c(n) = Z^{-1}[\log S(Z)] = Z^{-1}[\log \frac{1}{1-\sum_{i=1}^{P} a_i z^{-i}}] \tag{3}$$

where $c(n)$ is linear prediction cepstral coefficient (LPCC). It can be proven that the LPCC is a function of the linear prediction coefficients

$$c(n) = -a_n + \frac{1}{n}\sum_{i=1}^{n-1}(n-i)a_i c(n-i) \tag{4}$$

thus, the LPCC can be efficiently computed using an recursive process.

3. Line-spectrum pairs

Linear prediction coefficients and linear prediction cepstral coefficients are the most frequently used parameters for speech processing. However, line-spectrum pairs (LSP) are also gained some interest as alternative parameters since LSPs have many useful properties[3][4]. First, an error in one line-spectrum affects the all-pole spectrum near that frequency, thus LSPs may reflect some properties of auditory perception. Second, LSPs encode speech spectral information more efficiently than the other parameters.

One way to define the LSP is through the decomposition of the LPC analysis filter in (2) into even and odd functions. This can be accomplished by taking a difference and sum between $A(z)$ and its conjugate function

$$A_o(z) = A(z) - z^{-(P+1)}A(z^{-1}) = (1-z^{-1})\prod_{k=1}^{P/2}(1-2\cos(2\pi f_k T)z^{-1}+z^{-2})$$

and

$$A_e(z) = A(z) + z^{-(P+1)}A(z^{-1}) = (1+z^{-1})\prod_{k=1}^{P/2}(1-2\cos(2\pi f_k' T)z^{-1}+z^{-2}) \tag{5}$$

where $A_o(z)$ is an odd symmetrical function and $A_e(z)$ is an even symmetrical function. T is the sampling time interval. f_k and f_k' are the lower and upper line-spectra of the kth LSP.

4. Mel-scale frequency cepstrum

The auditory spectrum has confirmed the fact that the human ear resolves frequencies non-linearly. Experiments in speech recognition suggest that using a similar non-linear manner in feature representation improve the recognition performance. A commonly used non-linear scale is mel-scale which is defined by

$$Mel(f) = 2595 \log_{10}(1 + f/700) \tag{6}$$

Since in speech recognition, cepstrum is more often used than spectrum, a cepstrum derived from mel-scale spectrum is defined and is computed as follows. The spectrum of the speech signal is smoothed by a mel-scale triangle filterbank which is designed to give approximately equal resolution on a mel-scale. A log operation is then used to compress the dynamic range of the outputs of filter bins. Finally, the mel-scale frequency cepstral coefficients (MFCC) are calculated from the log filterbank amplitudes using the discrete cosine transform (DCT). Several works reported that in speech recognition the MFCCs outperform the LPCCs, thus this parameters are widely used in state-of-the-art speech recognizer. Some researchers have employed the MFCC for assessing the quality of synthetic or coded speech and concluded that the use of mel-scale frequency improves the correlation between objective measures and the perceptual results [2].

5. LPC residual

As shown in (1) and (2), passing the speech signal through LP filter $A(z)$ results in the removal of the near-sample correlation and produces the LP residual $e(n)$. The LP residual represents all the information not captured by the LPC, such as pitch, phase, etc. In speech recognition and speaker recognition, only LPC or some derived features are used and the LP residual is usually ignored. However, in the evaluation of the concatenative speech synthesis, the LP residual is perhaps more useful than LPC and LPCC since the concatenation process is based on the minimization of the cepstral distances between two concatenative segments.

The LP residual, like speech, is a time-domain signal. before the use of LP residual, it is necessary to represent it into lower order coefficients. In this paper, the LP residual is converted to mel-frequency cepstral coefficients and the resulted feature is called Residual MFCC or shorted for RMFCC.

6. Modified Mellin transform of log spectrum

The modified Mellin transform (MMT) of a signal *f(t)* is defined by the relation

$$M(s) = s\int_0^\infty f(t)t^{s-1}dt \tag{7}$$

where s is the Laplace factor. Just as the Fourier transform has a property of time delay invariance, the modified Mellin transform has a property of scale invariance. Many methods can be used to implement the MMT, while a direct form is employed in this paper. The resulted MMT is referred as to the modified direct Mellin transform (MDMT).

In [5], the MMT, combined with Fourier transform, has been used to represent speech signal for speech recognition. The proposed feature is calculated as shown in Fig. 1. The log spectrum is directly estimated by using an log operation to the amplitude spectrum. The modified direct Mellin transform is then used to remove the vocal tract length factor contained in the log spectrum. Finally, parameters are computed from the transformed sequence using DCT. Since the feature is actually the modified Mellin transform of log-spectrum, it is short for MMTLS. Due to the scale invariance property of the MMT, the MMTLS feature is proven to be insensitive to the vocal tract length factors among speakers.

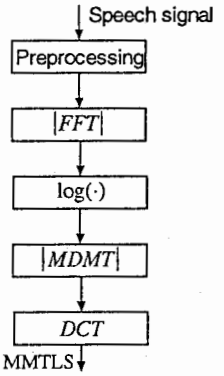

Fig. 1 MMTLS

In this paper, the same MMTLS feature is employed to represent speech signal for assessing the quality of synthesized speech. However, as far as the voice personality and intra-speaker voice quality are concerned in speech synthesis, the VTL information can not be removed. When only the naturalness and intelligence are concerned, such kind of information can be removed.

7. Segmental modified Mellin transform of log spectrum

For uniform lossless tube model of vocal tract, the MMTLS feature is proven insensitive to the vocal tract length. However, this is just a simplest ideal case. A real vocal tract model should consider many other effects, such as vibration losses and thermal losses. An improved MMTLS feature called segmental modified Mellin transform of log spectrum (SMMTLS) [6] is used to remove the vocal

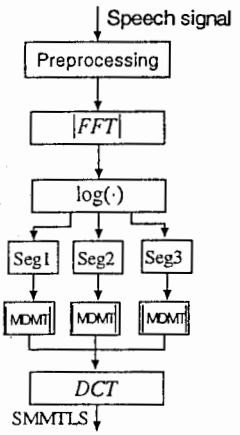

Fig. 2 SMMTLS

tract length effect for a more complicate vocal tract model. The SMMTLS is quite similar with the MMTLS. The only difference is that in the estimation of SMMTLS, the log spectrum is divided into several segments along the frequency axis, and the MDMT operation is used for each segment of log spectrum.

8. Bispectrum

All the speech features described above are determined from the amplitude spectrum or power spectrum (or autocorrelation function). The phase information of speech signal is thus neglected. However, the phase information is proven to play a very important role in speech naturalness and in general signal quality. Also, the higher order information is ignored since the power spectrum only determined by the second order statistics. If the speech process is a Gaussian process, the second order statistics are suffice for the complete description. However, much evidence appears to indicate that in general speech is non-Gaussian. The above two reasons motivate us to use higher order spectrum in the evaluation of speech synthesis.

Higher order spectrum is defined in terms of the higher order autocorrelation of the process [7]. For example, the third-order spectrum, also called bispectrum, by definition, is the Fourier transform of the third-order autocorrelation sequence. Similarly, the Fourier transform of the fourth-order autocorrelation function is called fourth-order spectrum or trispectrum. In this paper, only the bispectrum is investigated.

For a speech signal $s(n)$, its bispectrum, by definition, is

$$B(\omega_1,\omega_2)=\sum_{l=-\infty}^{+\infty}\sum_{m=-\infty}^{+\infty}R(l,m)W(l,m)\exp(-j\omega_1 l-j\omega_2 m) \tag{8}$$

where ω_1 and ω_2 are angle frequencies, $W(l,m)$ is a two dimensional window function which is used for good bispectral estimation, and $R(l,m)$ is the third-order autocorrelation function. $R(l,m)$, by definition, is

$$R(l,m)=\varepsilon[x(n)x(n+l)x(n+m)] \tag{9}$$

where ε is an expectation operator. For a stationary or quasi-stationary segment of discrete speech signal $\{s(0),s(1),\Lambda\ ,s(N)\}$, the third-order autocorrelation sequence is estimated via

$$R(l,m)=\tfrac{1}{N}\sum_{i=B_l}^{B_u}s(i)s(i+l)s(+m) \tag{10}$$

where $B_l = \max(0,-l,-m)$, and $B_u = \min(N-1, N-1-l, N-1-m)$.

Many approaches are available for estimating the bispectrum. A two-dimensional Fourier transform based method is adopted in this paper. Since the dimension of the bispectrum is generally high, a two-dimensional DCT is used for decompressing.

9. Bilinear time-frequency distribution

Bilinear time-frequency distribution (TFD), proposed by L. Cohen, is defined as the two dimensional Fourier transform of a weighted version of the symmetrical ambiguity function (AF) of the signal.

$$P(t,\omega) = \frac{1}{4\pi}\iint A(\theta,\tau)\phi(\theta,\tau)e^{-j\theta t - j\omega\tau}d\tau d\theta \tag{11}$$

where $A(\theta,\tau)$ is the symmetrical AF of the signal to be analyzed and is given by

$$A(\theta,\tau) = \int s(t+\frac{\tau}{2})s^*(t-\frac{\tau}{2})e^{j\theta t}dt \tag{12}$$

$\phi(\theta,\tau)$ is a weighting function called kernel. Different choices for the kernel yield quite different TFDs. For example, if a kernel is taking a constant value, say 1, i. e.,

$$\phi(\theta,\tau) = 1 \tag{13}$$

with this choice of the kernel, a well-known TFD called Wigner distribution (WD) is yielded.

Bilinear TFDs satisfy a long list of properties yet this varies with the different choice of the kernel. Details please refer to [8]. The utility of the bilinear TFDs in speech processing derive from their ability to provide simultaneously high time and frequency resolutions and thus avoid the well-known TF tradeoff in the short-time analysis. In this paper, only Wigner distribution is used and the feature extraction scheme is the same as that described in [9].

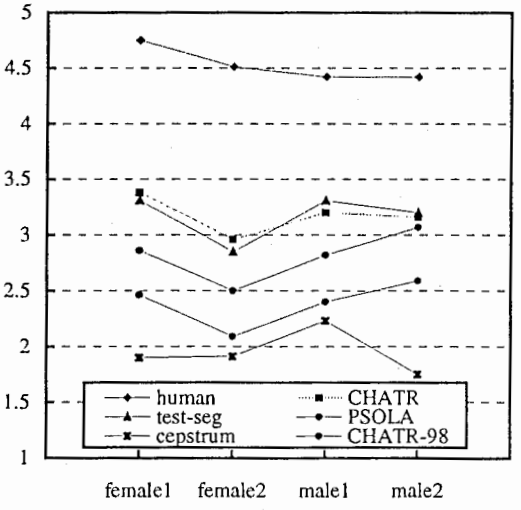

Fig.3 Results of MOS evaluation

III. Systems and database

The speech waveforms were synthesized using the CHATR system using raw-waveform concatenation. MOS tests had been previously performed to determine the benefits of using signal processing to modify the prosody of waveform to the predicted targets (see Fig. 3). The comparisons reported in this paper use the results from three methods: (a) Test_Seg, (b) PSOLA, and (c) UDB. (see Table1).

Method	(a)	(b)	(c)
Signal processing	no	yes	no
Natural Prosodic targets	yes	yes	no
Predicted prosodic targets	no	no	yes

Table 1. Configuration of test data

The basic evaluation principle used in this paper is to compare the synthetic speech to its natural recorded counterpart. The architecture of the system is shown in Fig. 4. The speech signal is first segmented into frames. Then the feature extraction will use the acoustic transformation described above to change the speech signal into feature vectors. Since the synthetic speech and the natural counterpart may have different duration, and therefore unequal numbers of frames, dynamic time warping (DTW) is used for alignment. Finally, a simple transform is used to transfer the distance to an objective score.

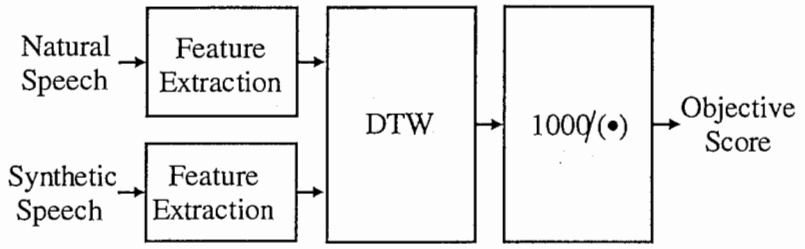

Fig. 4 The evaluation system

The data used in this comparison come from 40 listeners' evaluations of six Japanese sentences. For each sentence, natural speech and three types of synthetic speech from four voices were heard. Each sentence was scored twice by the listeners according to five level MOS evaluation score. Listeners were asked to judge naturalness and intelligibility, where 'naturalness' was defined to include both prosodic and voice-quality features.

IV. Experiment and results

Both the original natural speech and the synthetic speech were sampled at 16kHz. For all transforms except WVD, the speech signal is segmented into frames of 20ms. Each frame was converted into 12 feature coefficients according to each of the transforms to be tested. The WVD is an exception. It makes no implicit short-time stationary assumption and therefore it is not necessary to segment the speech signal into frames. However, the WV distribution of a discrete speech signal is a two-dimensional data matrix. In order to change the two-dimensional data matrix into a feature vector sequence like MFCC, the WV distribution is sampled along the time axis with an interval of tens of milliseconds. For each time point, a slice of the WVD, as the frequency domain input, is converted into a few MFCCs.

The computed scores and the MOS scores are in fig. 5 ~Fig. 10. From the results, it can be confirmed that the different measures we tested do indeed discriminate differently.

From sentence s1 we can see that the quality of the synthetic speech increases linearly from PSOLA, through test_seg to UDBp according the MOS score. However, no objective measure except the bispectrum and the MFCC was able to detect this improvement. All the measures were able to discriminate the difference in quality between test_seg and UDBp, and the bispectrum, LPC, LPCC, and MFCC have a better correlation with the subjective test results.

For sentence S2, the perceptual data shows a great increase in the quality of synthetic speech from PSOLA to test_seg, while again only the bispectrum can detect the improvement. From test_seg to UDBp, the quality still improves a lot but not as much as between PSOLA and test_seg. All objective measures can detect this improvement. Among them, the bispectrum, LPC, and MFCC perform slightly better than the other methods.

For S3, the quality from PSOLA, test_seg to UDBp increases linearly from a MOS score of 2.7 to a MOS score of 3.9. While only LPCC can detect the improvement from PSOLA to test_seg. From test_seg to UDBp, all methods detect an improvement; the bispectrum and the MFCC outperform the other methods.

In Fig.8, for sentence s4, the perceptual data shows a significant increase in the quality of synthetic speech from PSOLA to test_seg. However no methods can detect the improvement in this case. From test_seg to UDBp, there is still an improvement in the quality but not so greatly than that from PSOLA to test_seg. All the methods can detect the

improvement while the LPCC, LPC, MFCC, bispectrum and SMMTLS performs better than the rest methods.

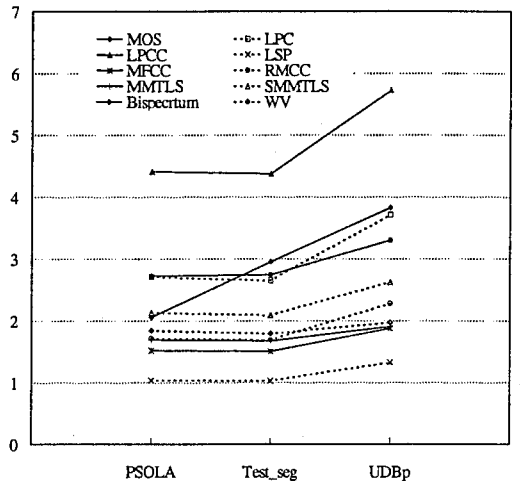

Fig. 5 Average score for s1

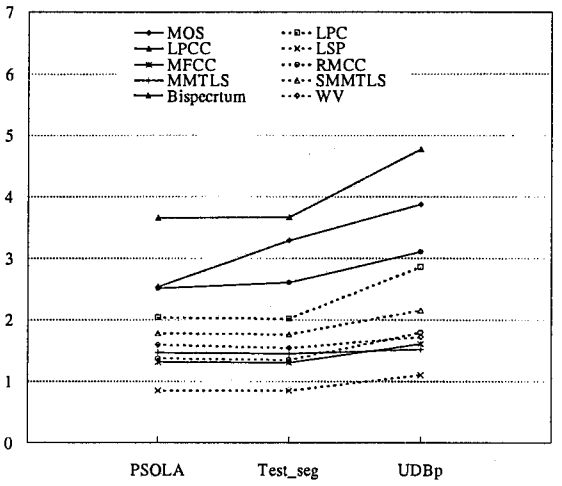

Fig. 6 Average score for s2

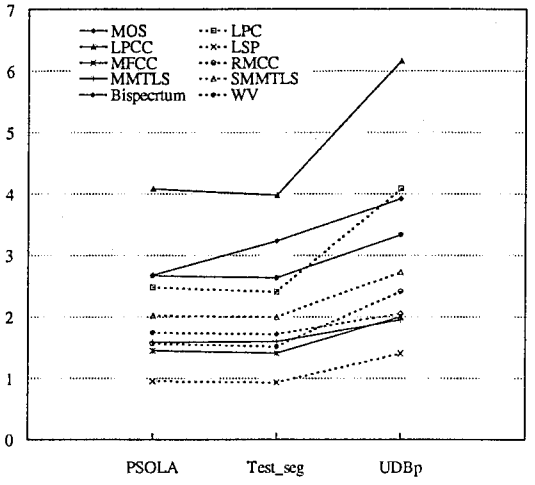

Fig. 7 Average score for s3

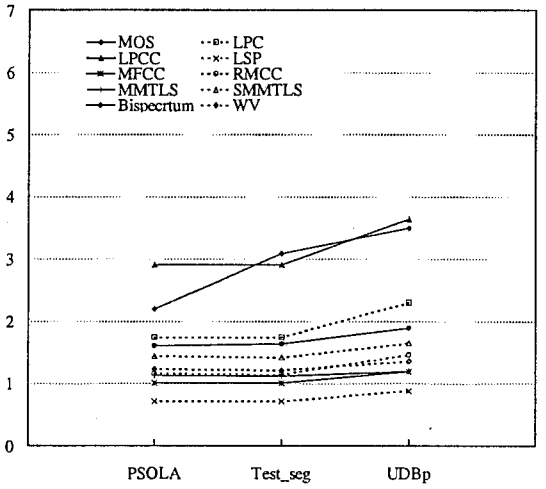

Fig. 8 Average score for s4

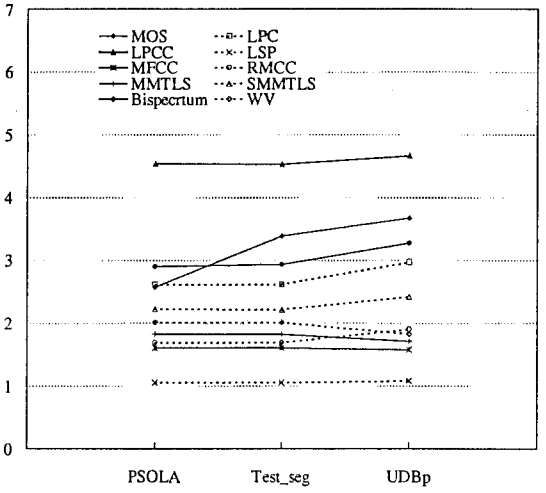

Fig. 9 Average score for s5

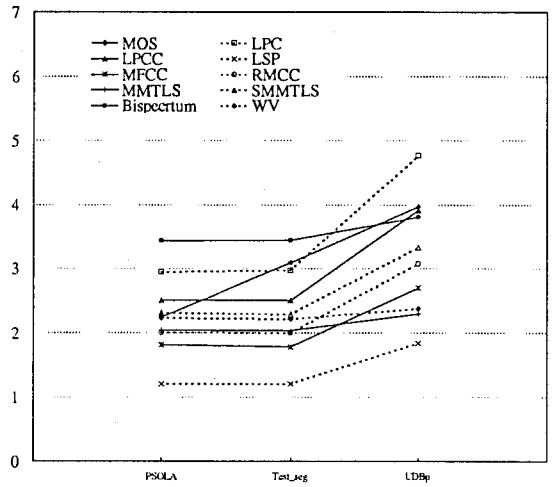

Fig. 10 Average score for s6

As for sentence s5, the MOS score reveals that the quality of test_seg is much better than the PSOLA synthesis, but only the bispectrum and LPCC show a slight improvement. From test_seg to UDBp, all the objective measures except Wigner-Ville distribution can identify the improvement. Among them, LPCC, bispectum, LPC and SMMTLS performs a little bit better.

For S6, once again the MOS score shows a linear increase in the quality of synthetic speech from PSOLA, test _seg to UDBp. From PSOLA to test_seg, four methods can detect the improvement. They are the LPCC, the LPC, LSP and the Bispectrum. From test_seg to UDBp, all methods can detect the improvement, but the SMMTLS, RMFCC and LSP have the higher discrimination.

From the results, It can be seen that different measures will yield quite different discrimination. The correlation between the computed distance and the MOS score is often used as a benchmark to evaluate the distance measures themselves.

If take the MOS score and the computed score as different set of observations, the degree of correlation between the computed distances and the MOS score can be measured by the population correlation coefficients. The population correlation coefficients between variable X and variable Y is defined by [10]

$$\rho_{XY} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{14}$$

where σ_X and σ_Y are the standard errors, and $\mathrm{cov}(X,Y)$ is the covariance. With this definition, the correlation of objective measures with the perceptual impressions are computed and shown in Table 2. It can be seen that from the sentence to sentence, the correlation varies greatly. For sentence s1, s3 and s6, the bispectrum has the highest correlation with perceptual data. The corresponding population correlation coefficients reach 0.811, 0.866 and 0.86 respectively. For s2, the MFCC has the highest correlation with the MOS score. LPCC gets the highest correlation in sentence s4, slightly higher than bispectrum. RMFCC is in the top of the correlation rank for sentences s5.

Sentence / Feature	A30	A46	A50	B42	C33	C47
LPC	0.798667	0.781024	0.855588	0.656487	0.701831	0.791018
LPCC	0.788331	0.784294	0.837507	0.688673	0.358695	0.744202
LSP	0.771741	0.776364	0.837119	0.684135	0.357572	0.760300
MFCC	0.793592	0.811214	0.813205	0.667289	0.204887	0.711478
RMFCC	0.803915	0.756281	0.846353	0.672757	0.744299	0.770692
MMTLS	0.648626	0.487465	0.861299	0.436815	0.227143	0.676662
SMMTLS	0.778885	0.732756	0.857973	0.632813	0.540633	0.751239
Bispectrum	0.811420	0.793365	0.866005	0.688405	0.676086	0.859643
WVD	0.622009	0.637378	0.643757	0.402021	-0.205478	0.256517

Table 2 The correlation coefficients for six test sentence and different objective measures

LPC	LPCC	LSP	MFCC	RMFCC	MMTLS	SMMTLS	Bispecrtum	WVD
0.76410	0.70028	0.69787	0.66694	0.76572	0.55634	0.71572	0.78249	0.42695

Table 3 The average correlation coefficients for different objective measures

The average correlation coefficients of objective measures with MOS score is shown in Table 3. It can be seen that overall, the bispectrum has the highest degree of correlation with MOS scores. This may possibly come from the two useful properties of bispectrum. First, the bispectrum encapsulates some phase information which is revealed to play a very important role in speech naturalness or in general speech quality . Second, all three synthesis methods to be evaluated in this paper employ the concatenation of segments of recorded natural speech. The waveform concatenation is performed with minimizing the MFCCs of the two successive units. In another words, the second order statistics is taken into account in the synthesis process. So in the evaluation process, the higher order information contained in the bispectrum may be more efficient.

The LPC feature performs better than the LPCC and the LSP. While the cause of the result is not clear. Actually, from Table 2, it can be seen that in most cases, the LPC, LPCC and LSP perform similarly. There is an only exception for sentence C33. In that case, LPCC and LSP get correlation coefficients about 0.35, while the LPC get a value of 0.7. We need to perform further experiments with a larger database before we can draw a conclusion for the evaluation among LPC, LPCC and LSP transforms.

The MFCC transform performs poorly in comparison with the bispectrum, the LPC, LSP, LPCC and SMMTLS. This result was unexpected, since many papers have reported that the MFCC has a higher correlation than the linear prediction based features. The cause may possibly be that the synthesis selection is based on minimizing the MFCCs between two successive segments. Hence the MFCC may not be reliable for assessment.

The WVD performed poorly in this comparison. One reason may be that we have not yet found an optimal way to convert the Wigner-Ville distribution into parameter vectors. Although the WVD can achieve very high time and frequency resolution simultaneously, it has some cross-term properties which may affect the performance.

Finally, The RMFCC has a correlation which is poorer than the bispectrum while better than the other seven methods. This is a particularly interesting result. It may indicate that the LPC residue contains some information about the speech signal such as pitch information which affects the quality of speech greatly.

Part II. A Five-Level-Score Evaluation System Based on Sluggish Coding

I. Introduction

In the first part, we have compared several different objective measures for use in the assessment and evaluation of speech synthesis and found that a measure using bispectrum has the highest degree of correlation with perceptual data. However, the acute readers may find that the computed distance is not a fivel-level-score like MOS score. With the use of different transformations will yield quite different scores. For example, for the same sentence, the use of the LPCC yields a score which is much higher than scores computed from the other transformations. While the use of the LSP always generate the lowest score amongst the nine acoustic transformations investigated. Even with the use of same acoustic transformation, the computed scores vary greatly for different utterance since the different utterance may have different duration and phonetic composition. That means the computed distance not only contains the quality information, but also is dependent on the transform used, the phonetic composition of the speech, the duration, etc. However, the target of this research is to investigate automatic objective measures to assess the quality of synthetic speech, the objective measures are expected to estimate a score which is only dependent upon the quality regardless of the phonetic composition, duration and other factors.

In this part, we will describe a five-level-score evaluation method. The evaluation process is divided into two steps. First, compare the synthetic speech with a natural spoken version of the same sentence by using "Acoustic transformation + DTW" described in the first part. The result of the comparison is a distance between the synthetic speech and its natural counterpart. Second, normalize the distance to a five-level-score similar to a MOS score by using a technique called sluggish coding. The preliminary results show that the computed score obtained by such method has high correlation with the perceptual data obtained from a MOS tests.

II. Evaluation paradigm

MOS evaluation of an utterance can be looked as an event which is associated with the evaluation space as shown in Fig. 11. Where each sample in the space corresponds to a score from one listener. Different listener gives different score and all the scores fall in the evaluation space. The expectation of all scores is the MOS score which is used as a standard for assessing the quality of speech.

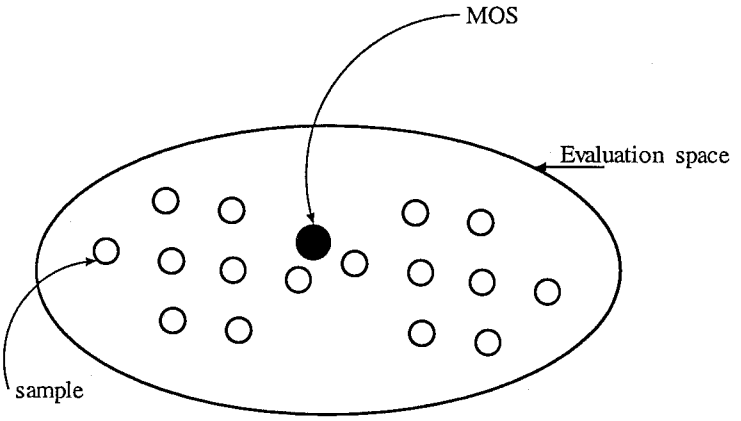

Fig. 11 stochastic model of an evaluation process

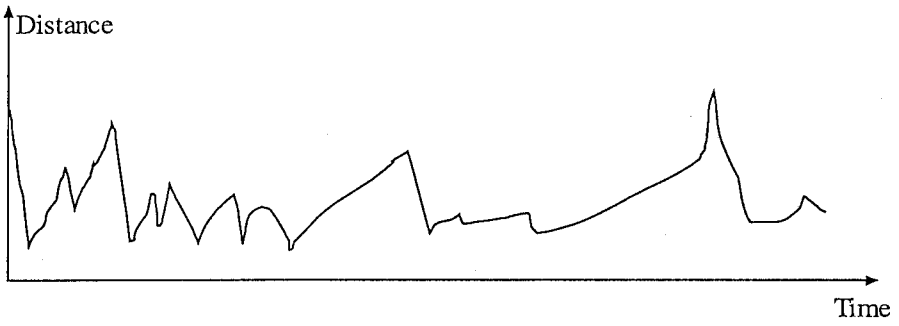

Fig. 12 Distance vs. time

However, for objective distance measures, we can not get different distance samples for one utterance. What we get is the computed distance which is a function of variable of time, as shown in Fig. 12.

If for each time point, we convert the distance to a score which represent the speech quality of current frame regardless of the phonetic composition and other factor, then the Fig. 12 may change as Fig. 13.

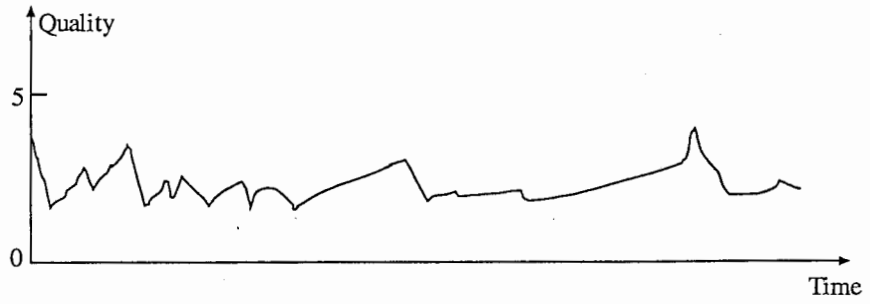

Fig. 13 Quality vs. time

If the quality is a stationary process, we then can use the time-domain average instead of the expectation in the evaluation space. Or in other wards, we use the time-domain average objective score as an alternate to a MOS score to evaluate the quality of synthesized speech. In the next section, we will introduce how to use sluggish coding to convert the distance in one time point to a score which is hopefully dependent on only the speech quality.

III. Sluggish coding

The sluggish coding is first proposed for SONAR target classification [12]. For an input sequence $X = [x_1, x_2, \Lambda\ , x_N]$, The sluggish coding will code the sequence into another binary sequence $Y = [y_1, y_2, \Lambda\ , y_N]$. Where y_k only takes a value of 1 or 0. The coding algorithm is defined as

$$y_k = \begin{cases} 1, & if\ |x_k| > T_u \\ 0, & if\ |x_k| < T_l \\ y_{k-1}, & else \end{cases} \tag{14}$$

where T_u and T_l are two thresholds. The reason to use sluggish coding is that comparing with the other binary coding methods, the sluggish coding is found to be more robust to noise.

We here expand the above definition to a five-level sluggish coding which is used to convert a distance to a five-level-score. If $d_{i,j}$ is a distance between the natural speech and synthetic speech in one node of DTW path, where *i* denotes the *i*th frame in the natural speech and *j* denotes the *j*th frame in the synthetic speech, the five-level sluggish coding converts the distance to a five-level score y_k by

$$y_k = \begin{cases} n, & if\ T_u^n < x_k < T_l^{n+1} \\ n, & if\ T_l^n < x_k < T_u^n and\ y_{k-1} \geq n \\ n-1, & if\ T_l^n < x_k < T_u^n and\ y_{k-1} < n \\ 5, & if\ x_k > T_u^5 \\ 0, & if\ x_k < T_l^1 \end{cases} \quad (15)$$

where $0 < n < 5$, T_u^n and T_u^n are different thresholds.

IV. Experiment and results

The database used for the experiment is the same as that in Part I. The acoustic transformation used to convert the speech signal into feature vector is LPC. The synthesis method is the UDB (see table 1). The thresholds in the sluggish coding for different sentences are determined by

$$T_u^n = T_u^{n-1} + 1.1T \quad (16a)$$

$$T_l^n = T_l^{n-1} + T \quad (16b)$$

$$T = N_{AS} / 4 \quad (16c)$$

Where N_{AS} is the average norm of the feature vector of the natural speech. If the feature sequence of the natural speech denotes as $X_N = [\vec{x}_1, \vec{x}_2, \Lambda, \vec{x}_K]$, where K is the total frames of the speech signal, then the N_{AS} is calculated as

$$N_{AS} = \sqrt{\sum_{i=1}^{K} \|\vec{x}_i\|^2} \quad (17)$$

With the use of sluggish coding with the thresholds described above, the evaluation method can automatically convert a distance in every node of the DTW path to a five-level-score. As described above, taking the average along the DTW path will yield a five-level-score which is used to represent the quality of the synthesized speech.

The computed five-level-score for different speakers are shown in Fig. 14 ~ Fig. 17. For the comparison, the MOS score for the same speaker and same sentence is also shown in the figures.

From Fig. 14, It can be seen that for the six sentence of the speaker MYS, although the computed scores for sentence s1 and s2 are different from the MOS scores, the computed scores are quite similar to the MOS score.

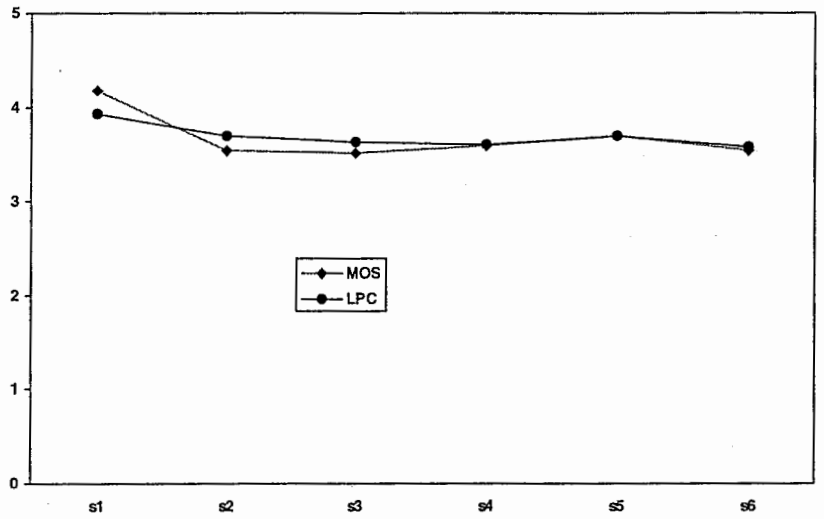

Fig. 14 The computed scores and MOS scores for speaker MYS

In the Fig. 15, the variation trend of the computed score is similarly to the MOS score except the sentence s4. For sentence s4, the MOS score suggests the quality of synthetic speech s4 is poor comparing with the other five sentences, while the objective measure get a relative high score. The reason is that the threshold for the sluggish coding for the sentence of this speaker is quite high.

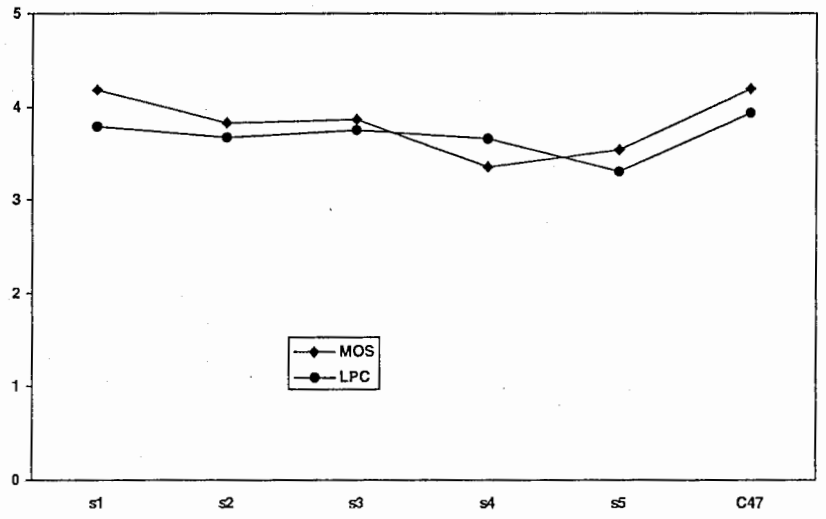

Fig. 15 The computed scores and MOS scores for speaker MYA

In Fig. 16, it is clear that the computed scores are quite similar to the MOS score.

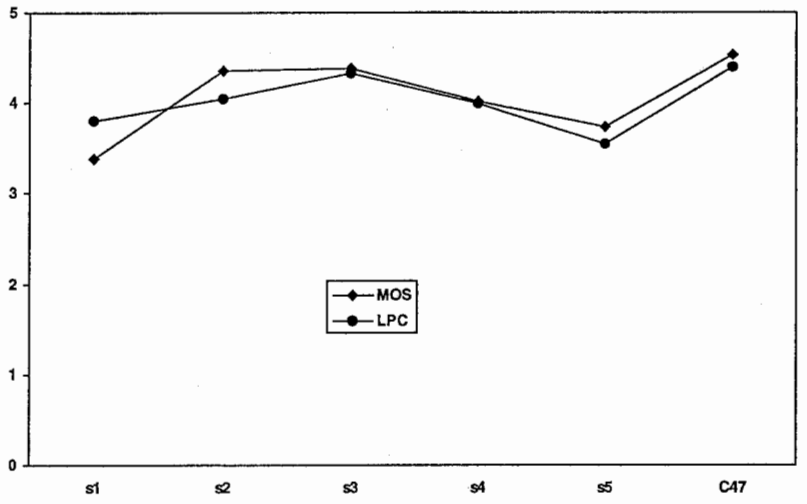

Fig. 16 The computed scores and MOS scores for speaker FMP

In Fig. 17, for all sentences but s4 and s5, the computed scores are quite similar to the MOS score. However, for s4 and s5, there are big differences between objective score and the MOS scores. We have listened both the original speech and synthesized speech and found that for this speaker and the two sentence, the prosody for the synthetic speech is quite different from that in the natural speech.

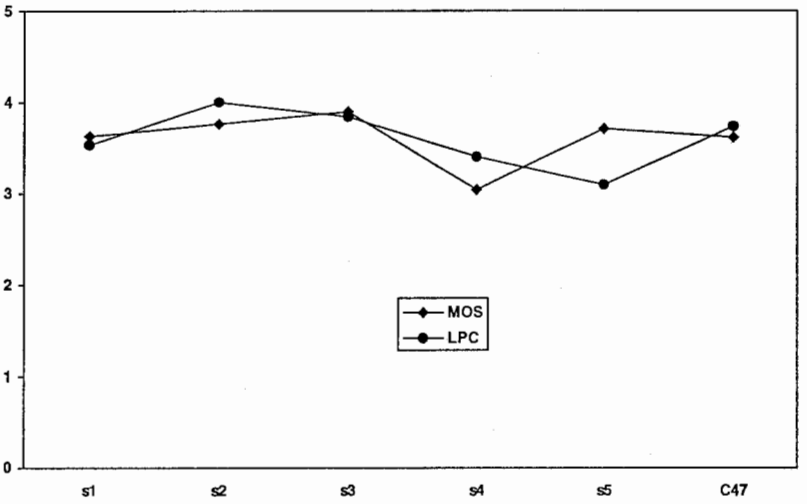

Fig. 17 The computed scores and MOS scores for speaker FOM

V. Discussion

This part has described a five-level-quality evaluation method based on sluggish coding. The preliminary experimental results show that the computed score has high degree of correlation with the MOS score.

In this part, only signal level features are considered for the evaluation of speech synthesis. However, in order to achieve reliable assessment results, information from at least two different levels needs to be taken into account, i.e., signal level features and prosodic features. Prosodic evaluation is currently being performed separately in depart II of ATR-ITL, we will not discuss in this report.

For the signal level evaluation, we are aware of various factors that may affect the performances of the acoustic measures.

1. Variable phoneme pronunciations

The current evaluation system directly compare the synthetic speech with the recorded natural counterpart. However, if a phoneme has different pronunciations in the natural and the synthetic speech, then the DTW may show a big difference for such phonemes. For example, in sentence s2, the transcription for speaker FOM in the natural speech has "nihon" pronounced as "nippon", while in the synthetic speech this word is pronounced as "nihon". Human listeners may not pay attention to such differences, thinking of them merely as acceptable alternative pronunciations in common use. However the physical measure shows a big difference which greatly affects the correlation between the computed distance and MOS score. We must assume that these differences, which can also carry nuance differences, are to be considered as errors in the synthesis.

2. Coarticulation effects

In natural spoken language, some words or phonemes may be repeated or deleted. For MOS scores, the human listeners may easily understand and neglect such a repetition or deletion. However, for objective distance measures, if such phenomena exist, then a big difference will be observed. Such problem can be resolved by carefully controlling the choice of utterance used for testing the physical measures.

3. Other factors

Many other factors may affect the distance measures. For example, breath noise, or different emphasis patterns. Breath noise is often heard in a long spoken sentence. If the

breath does not appear in the synthetic speech, a difference will be observed. Emphasis, on the other hand, can be looked upon as one dimension of prosodic evaluation. As with pronunciation differences, the difference in implied meaning (or nuance) may be better considered as an error in synthesized speech.

V. Conclusion

In this report, several acoustical transformations were investigated for assessment and evaluation of speech synthesis. The result showed that the bispectrum had the highest degree of correlation with the MOS score.

A five-level-score evaluation method based on sluggish coding was proposed. Preliminary results showed that the computed five-level-score had very high correlation with the MOS score.

In this report, only signal level features were considered. However, quality assessment and evaluation of synthetic speech is a multi-dimensional problem. Information from at least two different levels need to be taken into account, i. e., integrating both signal level features and prosodic features. Prosodic evaluation is currently being performed separately and is a much more subjective task. Work is in progress to combine the automatic assessment of prosodic features for a more comprehensive evaluation result.

The distance measures discussed in this report were based on the comparison between synthetic speech and a spoken version of the same sentence. However, in most case, there were only synthetic speech available and thus the DTW based method can not be used. We have tentatively used HMM in such case and found that the preliminary result is not so promise as what we expected. The work in progress will further our research in this area.

References

[1] A. H. Gray, J. D. Markel, "Distance Measures for Speech Processing", IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. 24, No. 5, PP. 380-391, 1976.

[2] Johan Wouters and Michael W. Macon, "A perceptual Evaluation of Distance Measures for Concatenative Speech Synthesis," in the proceeding of ICSLP'98, pp. 2747-2750, Dec. 1998.

[3] George S. Kang and Lawrence J. Fransen, "Application of Line-Spectrum Pairs to Low-Bit-Rate Speech Encodes", in the proceedings of ICASSP'85, Vol. I, PP. 244-247, Tampa, Florida, 1985.

[4] G. A. Mian and G. Riccardi, "A Localization Property of Line Spectrum Frequencies," IEEE Trans. on Speech and Audio Processing, Vol. 2, No. 4, PP. 536-539, October 1994.

[5] J. Chen, Bo Xu and Taiyi Huang, "A Novel Robust Feature of Speech Signal Based on the Mellin Transform for Speaker-independent Speech Recognition", in the proceeding of ICASSP'98, Seattle, USA, May 1998.

[6] J. Chen, Bo Xu, and Taiyi Huan, "An Improved Speech Feature Based on the Modified Mellin Transform for Speech Recognition", in the proceedings of ICCSLP'98, Singapore, Nov. 1998.

[7] T. Matsuoka and T. J. Ulrych "Phase estimation using the bispectrum", *Proceedings of the IEEE, Vol.* 72, PP. 1403-1411, 1984.

[8] Leon Cohen, "Time-Frequency Distributions-A review," Proceedings of the IEEE, VOl. 77, No. 7, July 1989.

[9] Adam, B. Fineberg and Kevin C. Yu, "Time-frequency Representation Based Cepstral Processing for Speech Recognition ," In the proceeding of ICASSP'96, Atlanta, Georgia, USA, May 1996, PP. 25-28.

[10] William W. Hines and Douglas C. Montgomery, "Probability and Statistics in Engineering and Management Science," third edition, John Wiley and Sons, 1990.

[11] Hynek Hermansky and Jean Claude Junqua, "Optimization of perceptually-based ASR front-end," in the proceeding of ICASSP'88, Vol. I, PP.219-222, 1988.

[12] J. Chen and Y. Xie, "An Improved Feature Extraction and Classification Technique of Underwater Targets", Proceedings of IEEE international Conference on Neural Networks and Signal Processing, 1995.