OBJECTIVE DISTANCE MEASURES FOR ASSESSING CONCATENATIVE SPEECH SYNTHESIS

Jing-Dong Chen & Nick Campbell

ATR Interpreting Telecommunications Research Labs
Kyoto, 619-02 Japan. ({jchen,nick}@itl.atr.co.jp)

Abstract

Several different acoustic transforms of the speech signal are compared for use in the assessment and evaluation of concatenative speech synthesis. The transforms tested include LPC, LSP, MFCC, bispectrum, Mellin transform of the log spectrum, Wigner-Ville distribution (WVD), etc. The computed distances between a synthesised 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 WVD-based cepstrum is found to behave poorly in this task.

1. 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 quality in the synthetic speech.

The evaluation measures can be generally classified into subjective and objective methods. Subjective methods require human listeners to judge speech quality, which may be 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, a standard method for evaluating speech coding, can also be used to determine 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, which have been widely employed in speech recognition technology, can be used for this purpose.

Previous work (e.g., [1]) compared several different distance measures for speech recognition, and showed that recognition performance varies according to the features used. For speech synthesis however, relatively little research has been performed on this topic, although recent developments in concatenative speech synthesis have used objective distance measures for selection units from a large speech corpus [2]. Since the purpose of this 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 perceptual impressions.

In a recent study investigating the perceptual relevance of distance measures for unit selection [3], some commonly-used distance measures such as FFT-based cepstra, LPC-based cepstra, line-spectrum 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 among the measures investigated.

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

The correlation between the computed distances and a previously obtained 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.

2. REPRESENTATIONS OF THE SPEECH WAVEFORM

In all, nine different acoustic transforms are compared for use in the assessment of concatenative speech synthesis. They are the linear prediction coefficients (LPC [4]), linear prediction cepstral coefficients (LPCC

[5]), line-spectrum pairs (LSP [6]), mel-scale frequency cepstral coefficients (MFCC [7]), residual MFCC (RMFCC, see below), bispectrum [8], modified Mellin transform of the log spectrum (MMTLS [9]), segmental modified Mellin transform of the log spectrum (SMMTLS [10]), and Wigner-Ville distribution based cepstrum (WVD [11]). The RMFCC is computed via two steps: first, the LP residual is produced by passing the speech signal through a twelve-order LP analysis filter, then the RMFCCs are calculated by converting the LP residual into mel-frequency cepstral coefficients.

The linear-prediction analysis and Fourier analysis based speech features are determined from the amplitude spectrum or power spectrum (or autocorrelation function). The phase information of the speech signal is thus neglected. However, phase information has been proven to play an important role in speech naturalness and signal quality in general. Furthermore, the higher order information is ignored since the power spectrum is only determined by second order statistics. If speech were a Gaussian process, then the second order statistics would suffice for a complete description. However, evidence appears to indicate that in general speech is non-Gaussian. The above two reasons motivate us to use the bispectrum for the evaluation of speech synthesis. The bispectrum, by definition, is

$$B(\omega_1,\omega_2) = \sum_{l=-\infty}^{+\infty}\sum_{d=-\infty}^{+\infty} R(l,m)W(l,m) \; exp(-j\omega_1 l - j\omega_2 m) \quad (1)$$

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. For further details refer to [12]. In this paper, a two-dimensional Fourier transform was used to compute the bispectrum. Since the dimension of the bispectrum is generally high, a two-dimensional DCT was used to decompress the bispectrum into lower order coefficients.

Cohen's class of time-frequency distribution (TFD [13]) is defined by

$$P(t,\omega) = \frac{1}{4\pi^2}\int\int\int \phi(\theta,\tau) f(u+\frac{\tau}{2}) f^*(u-\frac{\tau}{2}) \; e^{j\theta u - j\theta t - j\omega\tau} du d\tau d\theta \quad (2)$$

where $f(u)$ is the signal,$f^*(u)$ its complex conjugate and $\phi(\theta,\tau)$ the kernel defining a particular distribution. Different choices for the kernel yield quite different TFDs. For example, a kernel taking a constant value, say 1, will yield a well-known TFD called the Wigner distribution (WD). Cohen's class TFDs satisfy a long list of properties yet this varies with the different choice of kernel. 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 trade-off in the short-time analysis. In this paper, only the Wigner-Ville distribution is used and the feature extraction scheme is the same as that described in [11].

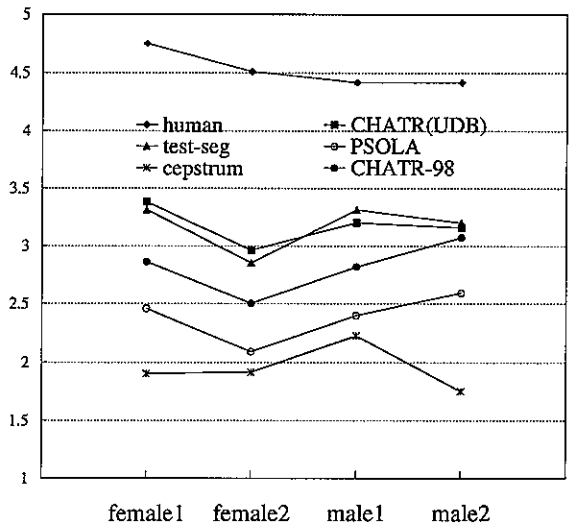

Figure 1: Results of MOS evaluation, comparing natural human speech and several different synthesis methods using four speakers' voices

3. EVALUATION METHOD

The speech waveforms were synthesized using the CHATR system [14] 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 Figure 1). The comparisons reported in this paper use the results from three methods: (a)'Test_Seg', (b)'PSOLA', and (c)'UDB' (see Table 1).

The basic evaluation principle used in this paper is to compare the synthetic speech to its natural recorded counterpart. The original speech signal and the synthetic speech signal are first segmented into frames. Each frame is then represented by several feature coefficients. Since the synthetic speech and the natural counterpart may have different durations, and therefore unequal numbers of frames,

Table 1: Configuration of test data

method:	(a)	(b)	(c)
signal processing	no	yes	no
natural prosodic targets	yes	yes	no
predicted prosodic targets	no	no	yes

Table 2: Correlation coefficients for six test sentences (and overall average) and different objective measures

	s1	s2	s3	s4	s5	s6	Ave
lpc	.79	.78	.85	.65	.70	.79	.76
lpcc	.78	.78	.83	.68	.35	.74	.70
lsp	.77	.77	.83	.68	.35	.76	.69
mfcc	.79	.81	.81	.66	.20	.71	.67
rmfcc	.80	.75	.84	.67	.74	.77	.76
mmtls	.64	.48	.86	.43	.22	.67	.56
smmtls	.77	.73	.85	.63	.54	.75	.72
bispec	.81	.79	.87	.68	.67	.86	.78
wvd	.62	.63	.64	.40	.20	.25	.43

dynamic time warping (DTW) is used for alignment. The result of DTW is a distance between the natural and the synthetic speech. The higher the distance value the poorer the quality of the synthetic speech, so the quality of speech synthesis corresponds to the inverse distance profile for each feature. To obtain an objective score which has a direct proportion to the speech quality, a transform is used to change the distance to a score

$$S_o(D) = \frac{1000}{D} \tag{3}$$

where D is the distance and S_o is the objective score.

However, even after the transformation, the resulting score is not a five-level quality score. We therefore use the correlation with the MOS score as a benchmark to assess the distance measure.

The data used in this comparison come from 40 listeners' evaluations of six Japanese sentences. For each sentence, natural speech and five types of synthetic speech from four voices were heard. Each sentence was scored twice by the listeners, using a 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.

4. EXPERIMENTAL 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 20 ms. Each frame was converted into 12 feature coefficients according to each of the transforms to be tested. 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 20ms. For each time point, a slice of the WVD, as the frequency domain input, is converted into 12 MFCCs.

The computed distance scores and the MOS scores are described in more detail in [15]. In this paper, we will use the population correlation coefficients as a benchmark to judge which transform is more appropriate for the assessment of speech synthesis. The results are 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 0.859 respectively. For s2, the MFCC has the highest correlation with the MOS score. LPCC and bispectrum get the highest correlation in sentence s4. RMFCC gives the highest correlation for sentence s5. The average population correlation of objective measures with the MOS scores is shown in the last column.

5. DISCUSSION

It can be seen that overall, the bispectrum has the highest degree of correlation with MOS scores. This may possibly come from two useful properties of the bispectrum. First, the bispectrum encapsulates some phase information, which is thought to play an 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 by minimizing the MFCCs of the pairs of successive units. In another words, the second order statistics are taken into account in the synthesis process, so in the evaluation process the higher order information, represented by the bispectrum, may be more efficient.

The LPC transform performs better than the LPCC and the LSP, although the reason for this is not clear. From Table 2 it can be seen that in most cases the LPC, LPCC and LSP perform similarly. There is one exception for sentence s5. In that case, LPCC and LSP have correlation coefficients of 0.35, while the LPC has 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, as mentioned above, the unit selection in the synthesis 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.

6. CONCLUSION

In this paper, several acoustic features were investigated for assessment and evaluation of synthetic speech. They are the LPC, LPCC, LSP, MFCC, RMFCC, MMTLS, SMMTLS, the bispectrum and the Wigner-Ville distribution. A comparison between computed distances and the MOS score was performed. The database used contains six Japanese sentences. Each 'sentence' includes the natural speech and three kinds of synthetic speech from two male and two female speakers. In all, 24 natural speech sentences and 72 synthetic speech sentences were compared. The results show that distances computed from the bispectrum had the highest correlation with the MOS score. Objective distance measures using the RMFCC also had a high correlation with the perceptual data, which reveals that the LPC residue also contains some important information about the speech signal that should be considered in assessing the speech quality.

In this paper, only signal level features are considered, but quality assessment and evaluation of synthetic speech is a multi-dimensional problem. Information from at least two different levels needs 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. Our future plans also include extending the current signal-level research using more extensive test data.

Acknowledgement

This work is supported by the Trust Agency of the Japan Key Technology Center and is being carried out in Department II of ATR-ITL.

REFERENCES

[1] Hynek Hermansky and Jean Claude Junqua, "Optimization of perceptually-based ASR front-end", in the proceeding of ICASSP88, Vol. I, PP.219-222, 1988.

[2] Wen Ding and Nick Campbell, "Optimsing Unit Selection with Voice Source and Formants in the CHATR Speech Synthesis System", in the proceeding of EUROSPEECH'97, pp. 537-540, Rhodes, Greece, 1997.

[3] 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.

[4] Shuzo Saito, "Speech Science and Technology", IOS PRESS, 1992.

[5] 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.

[6] George S. Kang and Lawrence J. Fransen, "Application of Line-Spectrum Pairs to Low-Bit-Rate Speech Encodes", in Proc ICASSP85, Vol. I, pp. 244-247, Tampa, Florida, 1985.

[7] Steve Young, "Large Vocabulary Speech Recognition: A Review", IEEE Sig Proc Magazine, 1996.

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

[9] 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 Proc ICASSP-98, Seattle, USA, May 1998.

[10] J. Chen, Bo Xu, and Taiyi Huan, "An Improved Speech Feature Based on the Modified Mellin Transform for Speech Recognition", in Proc ICCSLP98, Singapore, Nov 1998.

[11] Adam, B. Fineberg and Kevin C. Yu, "Time-frequency Representation Based Cepstral Processing for Speech Recognition", In Proc ICASSP96, Atlanta, Georgia, USA, pp. 25-28. May 1996.

[12] C. L. Nikias and M. R. Raghueer, "Bispectrum Estimation: A Digital Signal Processing Framework", Proceedings of the IEEE 76, pp.869-891,1987.

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

[14] CHATR Speech Synthesis (see references and speech samples under www.itl.atr.co.jp/chatr)

[15] Chen, J. D., & Campbell, W. N., "Speech Synthesis Evaluation by Objective Distance Measures", in SP-99-xxx, Tech Rept of the IEICE, May 1999.