OBJECTIVE SPEECH MEASURE FOR CHINESE IN WIRELESS ENVIRONMENT

K.H. Lam, O.C. Au, C.C. Chan, K.F. Hui, S.F. Lau
Hong Kong University of Science and Technology

ABSTRACT

Nowadays, cellular phone is becoming an important mobile wireless communication means, especially in metropolitan areas. One of the important operating considerations of cellular phone service providers is to maintain good speech quality of the cellular phone network. Subjective evaluation by repeated listening tests at various sites within the coverage area is impractical due to its intrinsic laborious and expensive nature. As a result, it would be much desirable to have an automatic objective evaluation system which applies a good objective speech measure to estimate the statistical average of subjective opinions of the typical conversational speech sentences sent through the cellular network. While extensive work was done for objective speech measures for languages such as English, Japanese, French, and other western languages, little has been done for Chinese. In addition, little has been done to quantify speech quality in the wireless environment.

1 Introduction

In this paper, we study experimentally the effectiveness of the class of spectral distance(SD) based speech measures[1] in estimating the subjective quality of Cantonese speech in an analog AMPS (Advanced Mobile Phone System) system. Cantonese is one of the most popular Chinese dialects. We design a set of phonetically-balanced Cantonese conversational sentences, record the sentences with various distortions from an AMPS cellular phone, obtain subjective mean opinion scores(MOS) through surveying, and find the best objective measure that is most statistically correlated with the MOS.

Owing to the asynchronous nature of the received cellular phone speech signal as compared with the original, we make the novel observation that synchronization is an important pre-processing step for objective speech evaluation because objective speech measures are usually sensitive to the accuracy of synchronization.

2 Spectral Distance Based Measures

To measure the speech quality of a distorted sentence as compared with the original one, we assume that the two sentences are synchronized. We divide the two sentences into frames of 15ms in duration. For each measure f, a local function $f(n)$ is computed for each frame n and the speech measure is defined as the time average of the local function over all the frames.

$$f = \frac{1}{N}\sum_{n=1}^{N} f(n)$$

Within each frame, an M^{th} order linear predictive coding(LPC) analysis is performed using Durbin's recursion. Let $V_d(n,\theta)$ be the magnitude spectrum for the n^{th} frame of the distorted sentence at frequency θ.

$$V_d(n,\theta) = \left| \frac{1}{1-\sum_{k=1}^{M} a_d(k)e^{-jk\theta}} \right|$$

where $a_d(k)$ is the k^{th} LPC coefficient in the frame. Similarly, let $V_o(n,\theta)$ be the magnitude spectrum for the n^{th} frame of the original sentence at frequency θ.

let L be the number of samples within each frame. The spectral distance(SD), frequency weighted spectral distance($FWSD$), energy weighted spectral distance($EWSD$), and auditory frequency weighted spectral distance($AFWSD$) are defined as

$$SD(n) = \left\{ \frac{1}{L}\sum_{i=1}^{L} |V_o(n,\theta_i) - V_d(n,\theta_i)|^p \right\}^{\frac{1}{p}} \qquad (1)$$

$$FWSD(n) =$$

$$\left\{ \frac{\sum_{i=1}^{L} |V_o(n,\theta_i)||V_o(n,\theta_i) - V_d(n,\theta_i)|^p}{\sum_{i=1}^{L} |V_o(n,\theta_i)|} \right\}^{\frac{1}{p}} \qquad (2)$$

$$EWSD(n) = \left\{ \frac{1}{L}G(n)\sum_{i=1}^{L} |V_o(n,\theta_i) - V_d(n,\theta_i)|^p \right\}^{\frac{1}{p}} \qquad (3)$$

0-7803-2431-5/95 $4.00 © 1995 IEEE

$AFWSD(n) =$

$$\left\{ \frac{\sum_{i=1}^{L} AU(\theta_i)|V_o(n,\theta_i) - V_d(n,\theta_i)|^p}{\sum_{i=1}^{L} AU(\theta_i)} \right\}^{\frac{1}{p}} \quad (4)$$

where $G(n)$ is the gain term of the n^{th} frame of the original sentence

$$G(n) = [R(0) - \sum_{all} a(k)R(k)]^{\frac{1}{2}}$$

with $R(i)$ being the correlation function of the original signal, and $AU(\theta_i)$ is the minimum human auditory frequency response of equal-loudness at frequency θ_i. The log spectral distance(LSD), frequency weighted log spectral distance($FWLSD$), energy weighted log spectral distance($EWLSD$) and auditory frequency weighted log spectral distance($AFWLSD$) are defined similarly by using 20 log $V_o(n,\theta_i)$ - 20 log $V_d(n,\theta_i)$ instead of $V_o(n,\theta_i)$ - $V_d(n,\theta_i)$ in Eqns. 1, 2, 3 and 4 respectively.

The inverse linear spectral distance(ISD) and the inverse log spectral distance($ILSD$) are defined in terms of SD and LSD as

$$ISD = \frac{1}{b+SD}, \quad ILSD = \frac{1}{b+LSD}$$

where b is a constant. We pick $b = 1$.

While these measures can be found in [1], their application to Chinese language in wireless environment is certainly new.

3 Design of Speech Database

Cantonese[2] is a monosyllabic and tonal language, unlike most western languages. Each syllable in Cantonese is made up of an Initial and a Final. The Initial is an optional consonant whilst the Final consists of a nucleus upon which vocalic and consonantal qualities and tonal changes are perceived. The four basic Cantonese syllable types are V, C_1V, C_1VC_2 and VC_2, where C is a consonant and V is a vowel or diphthong. The C_2 must be either a nasal or a stop consonant. In Cantonese, there are nine lexical tones. A change in tone may produce a difference in lexical meaning.

The speech database consists of 10 Chinese sentences. Each sentence is spoken by two male and female native speakers with no strong ascent. Each sentence is about 20 syllables in length and is constrained to have at least one vowel from each of the three categories: front, mid, back. Each sentence also encompasses the consonants of Cantonese as well as all the nine tones. In this way, the speech material are made to be phonetically balanced, including the main characteristics or features of Cantonese.

4 The Experiment

As shown in Fig. 1, the original undistorted speech signal is played from a DAT walkman to a cellular phone via a car kit interface box. The whole setup is installed in a van which moves along some designated route. Simultaneously, in the laboratory, the distorted speech signal is recorded from a normal telephone to a DAT deck via an interface circuit. Four routes are selected as shown in Table 1 to yield various signal distortions due to the wireless channel. Database 1, 2, and 3 contain distortion due to low power, multipath fading and slow/fast fading. Database 4 serves as a control situation with little distortion due to high power.

After the recording, the distorted speech signal is transferred to computer through an A/D convertor with 16 bit resolution and a sampling frequency of 48kHz. The oversampling is needed to give good resolution of the signal for synchronization purpose. Sampling at Nyquist frequency can retain enough information for perfect reconstruction of the signal but insufficient information for the correct synchronization of the two signals due to phase uncertainty.

For detailed study, 160 distorted sentences were selected with a mixture of good and bad quality. A survey was conducted to collect the subjective MOS on the 160 sentences.

Time alignment between the original and distorted signals is indeed vital to the objective measures because many objective measures are very sensitive to the accuracy of the synchronization. Direct application of any objective measures to unsynchronized signals can be totally meaningless.

We perform the synchronization in two stages. Firstly, sub-sampled synchronization is done to find the approximate position of the optimal point in an efficient manner. The signals are decimated by a factor of 6 and the position p_o of minimum mean square error is computed. Synchronization without decimation is then carried out around p_o to accurately locate the optimal point. The synchronization is visually verified. When two signals are synchronized, they are indeed time aligned with similar zero crossings and peak locations in the time domain.

5 Results

Shown in Fig.2 are two typical plots of the observed MOS values for the 160 sentences against the measures: LSD and $AULSD$. Shown also is a second order polynomial predictor fitted by least square regression to the scatter plot:

$$\widehat{MOS} = af^2 + bf + c$$

where f is the objective measure being studied. For a measure to be good, the polynomial predictor should follow the scatter plot closely. To measure the relative performance of the objective speech measures, the correlation coefficient[3]

$$r = \sqrt{\frac{\sum(\widehat{MOS}_i - m_y)^2}{\sum(MOS_i - m_y)^2}}$$

is used. A perfect prediction would yield $r = 1$. The variable s is the standard deviation of the prediction error, and in ideal conditions, it would be zero.

In Table 2, the r is shown with various norm values of p. For objective speech measurement of Chinese in the wireless environment, it is found that for the linear spectral distance based measures, the 1-norm (p=1) yields the best r. The r appears to decrease monotonically with increasing p. On the other hand, the 5-norm is better for the LSD, $AULSD$ and $ILSD$. The 2-norm is better for $FWLSD$ and 1-norm is better for $EWLSD$. In general, the log-measures have remarkably higher correlation coefficients than the corresponding linear counterparts. The objective measures tend to be sensitve to the synchronization accuracy. Among the ten measures considered, the $AULSD$(0.7853), LSD(0.7627) and $ILSD$(0.7536) are superior.

References

[1] S.R. Quackenbush, T.P. Barnwell, M.A. Clements, Objective Measures of Speech Quality, Prentice Hall, 1988.

[2] P.C. Ching, et al, "From Phonology and Acoustic Properties to Automatic Recognition of Cantonese", ISSIPNN, pp.127-132, Apr. 1994.

[3] N.R. Draper, H. Smith, Applied Regression Analysis, John Wiley & Sons, New York, 1981.

Acknowledgement

This work supported in part by Hutchison Telephone Company Limited under project # HT92/93.EG01

Database	Power	Situation	Environment	Distortion
1	Low	Standing	Surrounded by buildings	Multi-path fading & slow fading
2	Low	Moving (45 km/hr)	conducted in high-way	fast fading
3	Low	Standing	conducted in open area	slow fading
4	High	Moving (50 km/hr)	conducted in high-way	fast fading

Table 1: Characteristics of the 4 databases

	1-norm	2-norm	5-norm	6-norm	8-norm	10-norm	20-norm	50-norm	Infinity -norm
Spectral Distance **SD**	0.5114	0.3105	0.0804	0.0612	0.0408	0.0306	0.0156	0.0106	0.0096
Frequency Weighted SD **FWSD**	0.4074	0.2949	0.1545	0.1437	0.1325	0.1272	0.1203	0.1195	0.0096
Energy Weighted SD **EWSD**	0.2663	0.1763	0.0537	0.0406	0.0264	0.0196	0.0106	0.0086	0.0096
Auditory Frequency Weighted SD **AUSD**	0.4785	0.2880	0.0679	0.0527	0.0365	0.0280	0.0136	0.0082	0.0096
Log SD **LSD**	0.6829	0.7386	0.7627	0.7299	0.6728	0.6331	0.5542	0.5161	0.5002
Frequency Weighted Log SD **FWLSD**	0.4108	0.5314	0.4555	0.4378	0.4115	0.3932	0.3520	0.3282	0.5002
Energy Weighted Log SD **EWLSD**	0.5508	0.5091	0.4935	0.4964	0.5033	0.5096	0.5238	0.5181	0.5002
Auditory Frequency Weighted Log SD **AULSD**	0.6623	0.7127	0.7835	0.7594	0.7032	0.6584	0.5648	0.5197	0.5002
Inverse SD **ISD**	0.5100	0.3058	0.0795	0.0630	0.0479	0.0421	0.0374	0.0374	0.0376
Inverse Log SD **ILSD**	0.6827	0.7339	0.7536	0.7208	0.6682	0.6337	0.5655	0.5292	0.5128

Table 2: Overall correlation results for different norm values

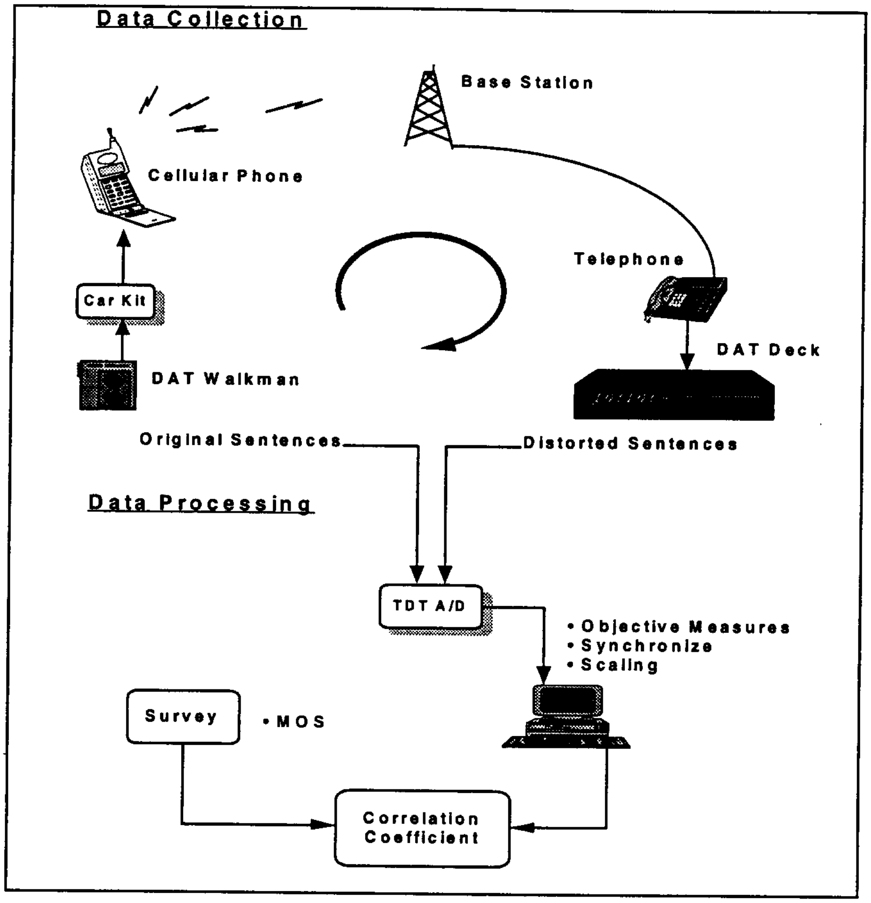

Fig. 1 Block Diagram for the Experiment

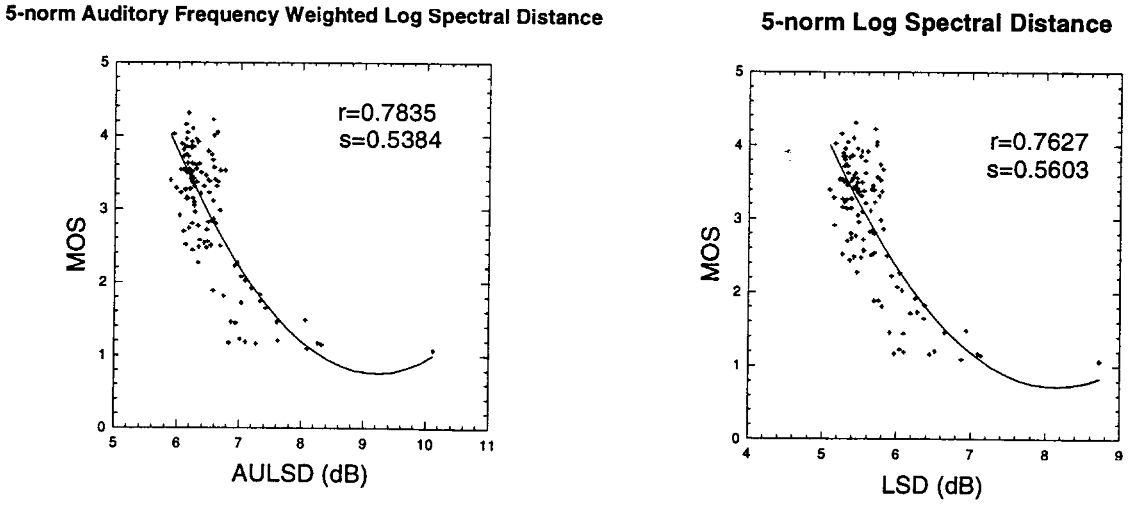

Fig.2 Observed MOS Vs Objectively measured distortion