SOME INVESTIGATIONS ON THE VOCODED SPEECH PERCEPTION AND ENCODING

K. Ratkevičius and A. Rudžionis

Speech Research Laboratory
Kaunas Technical University

ABSTRACT - The relation between the quality of vocoded speech and its compression ratio and improvements in parametric coding techniques leading to data rates as low as 1200 bits/s were investigated.

INTRODUCTION

We do not know yet how the quality of vocoded speech is constrained by resolution of the information contained in its resynthesis parameters. The results presented in the paper (Clark, Mannell & Ostry, 1987) are very interesting, but they deal only with vowel and consonant intelligibility. This investigation reports on the effect of data rate on vocoded speech syllable intelligibility, speaker recognizability and the perception accuracy of some sound clusters. Simple differential pulse code modulation method (in the frequency domain) and one modification of variable frame rate encoding in time domain are proposed for producing acceptable speech at data rates as low as 1200 bits/s.

METHODOLOGY

A computer-aided hardware implemented 24 channel vocoder was used in our investigations. Its frequency range is from 100 Hz to 8 kHz. Major spectral variables in channel vocoder are: the number of spectral parameters L, bit number B (the number of quantization bits per one spectral parameter) and the sampling period T. Basic values of these variables were: L=24, B=8 bits, T=10 ms. Samples of spectrum from the output of the analyzer were loaded to the computer, where the reducing of L by averaging a certain number of neighbouring spectrum samples, the reducing of B or the increasing of T by repeating spectrum frames, were performed. Samples of compressed spectrum were fed in to the synthesizer and recorded on a magnetic tape.

INTELLIGIBILITY OF VOCODED SPEECH

Five tables of phonetically balanced syllables of Russian words (the total number was 250 syllables) of one male speaker were processed in the vocoder for each measuring of intelligibility. Synthesized syllables were played before three skilled listeners. The listeners were asked to exactly transcribe what they heard. The overall score of syllable intelligibility S was calculated by dividing the number of correct responses by the total number of test syllables. The determined relations are listed in Table 1.For the same set of spectral parameters L, the relations between syllable intelligibility and sampling period T were measured and are shown in Table 2.

L	B			
	2	3	4	8
6	36.8	46.7	50.2	52.1
8	52.3	59.9	62.1	65.6
12	70.4	75.9	78.0	80.2
24	79.1	84.7	86.3	88.0

Table 1. Syllable intelligibility as a function of the number of quantization bits B for a set of spectral parameters L, %

L	T, ms		
	10	20	40
6	58.9	52.1	39.3
8	70.7	65.6	53.0
12	82.5	80.2	67.5
24	90.1	88.0	78.3

Table 2. Syllable intelligibility as a function of sampling period T for a set of spectral parameters L, %

The results we report here allow to evaluate the individual contribution of variables L, B, T to the intelligibility of vocoded speech. The reducing of L and B or increasing of T cause the loss of intelligibility, but there are ranges of variables L=12-24, B=15-30 ms, inside which the decreasing of representation accuracy of one variable can be compensated by proportional increasing of representation accuracy of another variable.

PERCEPTION OF SOME SOUND CLUSTERS

An investigation was carried out to determine the influence of the number of spectral parameters L on the perception accuracy of consonants /m/, /n/, /v/, /l/, which all are rather similar in acoustic features. Test stimuli were close /CVl/ syllables, where C={m, n, v, l}, l-consonant /l/. Vocoded syllable samples were collected of two speakers and a total number of 384 syllables was heard by 5 listeners in each listening condition.

Percentages P of incorrectly recognized consonants for a set of spectral parameters L were as follows: P=28.3% (L=6), P=12.3% (L=8), P=6.0% (L=12), P=2.1% (L=24).

The reducing of the number of spectral parameters, while other spectral variables are constant (B=8, T=10 ms), causes the appreciable loss of the perception accuracy of consonants /m/, /n/, /l/, /v/.

Another investigation was made to determine the influence of the sampling period T on the perception accuracy of plosive consonants, which are most sensitive to speech sampling period restrictions. Test stimuli were again /CV1/ syllables, but in this case plosive consonants C={-, b, d, g, p, t, k} were used. The dash indicates the absence of a plosive consonant. A total number of responses was 1470. The results of listening experiments are summarized in Table 3.

Vowel	T, ms		
	40	20	10
/a/	6.5	1.4	0
/u/	10.2	5.8	0.7
/i/	20.1	10.5	5.8
Average	12.3	5.9	2.2

Table 3. Percentages of incorrectly recognized plosive consonants as a function of sample period T taking into account the vowel context

The results show that the perception accuracy of plosive consonants strongly depends on vowel context. Subsequent investigations must be done to determine the sufficient sampling period T, which could enable to ensure the more exact perception of palatalized plosives.

SPEAKER IDENTIFICATION

Tests to produce speaker identification ratings have been carried out to supplement the results of intelligibility and perception tests.

First an attempt was made to find a relation between the speaker identification accuracy and the duration of speech sample from unprocessed speech and vocoded speech. Speech samples were collected from 11 familiar speakers and 4 unknown speakers. 5 short words with average duration of 0.5 s, 5 long words with average duration of 1.5 s and 5 sentences, each of 4 s average duration were used. Each sequence of samples was

chosen at random, recorded on a magnetic tape and played to 11 listeners. A warning concerning the unknown speakers was made. Unprocessed speech (US), vocoded speech (VS) and monotonous vocoded speech with constant pitch (MVS)were tested.

The relations between the speaker identification accuracy and the duration of speech sample are shown in Table 4.

Speech type	t, s		
	0.5	1.5	4.0
US	71.6	94.5	98.8
VS	57.2	86.1	92.3
MVS	42.6	62.0	72.2

Table 4. Speaker identification accuracy as a function of the duration of speech sample t for unprocessed speech US, vocoded speech VS and monotonous vocoded speech MVS, %

The second test was performed using 11 familiar speakers and sentences only. Vocoded speech with different numbers of spectral parameters L was evaluated.

Percentages of speaker identification accuracy I for a set of spectral parameters L were the following: I=65.7% (L=6), I=74.7% (L=8), I=90.9% (L=12), I=97.0% (L=24).

These results lead us to a conclusion that the prosodic information of speech, i.e. pitch and the duration of the speech sample, have the main influence on the accuracy of speaker identification. On the other hand spectral features are important for speaker identification too.

VARIABLE FRAME RATE ENCODING OF SPEECH PARAMETERS

Two modifications of variable frame rate (VFR) encoding of speech parameters were used: (1)VFR with averaged frames repeating; (2)VFR with linear interpolation. In a VFR encoding the frame rate is adapted to the speed of articulatory movement; i.e., frames are selected in rather large time intervals during stationary segments such us vowels, whereas the frame rate during rapid transitions is rather high. First modification of VFR encoding use simple repeating of the averaged frames to restore missing frames in the synthesizer.The linear interpolation procedure for the frames which are not selected is employed in the second modification of VFR encoding.

Tests, based on paired comparisons of synthesized words, were carried out. The results of informal listening experiments showed that, compared to a vocoder with constant frame rate, first VFR encoding principle permits increasing of the sampling period T to 30-40 ms and the second VFR encoding principle - to 50 ms without a perceptible loss of quality.

Frequency domain rate reduction was accomplished by differential pulse code modulation (DPCM), in which the reference parameter is coded with a 4-bit logarithmic code, other parameters - by 2-bit DPCM.

A combination of DPCM and VFR encoding with linear interpolation was used. The bit allocation was the following: reference channel - 4 bits, 23 channels - 46 bits, pitch and voicing - 6 bits, interpolation step - 4 bits. A total number of bits per frame was 60 and resulting bit rate was 1200 bits/s. The results of listening test showed that the loss of synthesized words quality due to the encoding procedure was almost negligible.

CONCLUSIONS

The presented relations between vocoded speech quality measures (syllable intelligibility and speaker recognizability) and major spectral variables of the channel vocoder permit to choose the essential parameters of channel vocoders in order to guarantee the required vocoded speech quality at pre-given data rate.

From the results it can be concluded, that the reducing of the number of spectral parameters below 20 causes the appreciable loss of the perception accuracy of consonants /m/, /n/, /l/, /v/. The perception accuracy of plosive consonants strongly depends on vowel context. Palatalized plosives are most sensitive to speech sampling period restrictions. They are confused among themselves when sampling period T is greater than or equal to 10 ms.

Time and frequency domain data rate reduction is accomplished leading to data rate 1200 bits/s without a perceptible loss of vocoded speech quality.

REFERENCES

Clark, J.E., Mannell, R.H., Ostry, D. (1987) *Time and frequency resolution constraints on synthetic speech intelligibility*, Proc. XIth ICPhS, Tallinn, 215-218.