A RELATIONSHIP BETWEEN THE QUALITY OF VOCODED SPEECH AND ITS COMPRESSION RATIO

KASTYTIS RATKEVIČIUS ALGIMANTAS RUDŽIONIS

Speech Research Laboratory
Kaunas Polytechnical Institute
Kaunas, Lithuania, USSR 233028

ABSTRACT

A 24-channel vocoder was used to study the quality of vocoded speech under the effects of its compression variables - the number of spectral parameters n, sampling period T and bit number m in vocoder spectral parameters. Syllable intelligibility S and speaker recognition $\mathcal{J}$ (identification) were used as measures of the quality. To reduce the number of spectral parameters the method of averaging over subsequent amplitude spectrum samples (AS) is suggested.

INTRODUCTION

Major variables of the compression ratio in a channel vocoder are: the number of spectral parameters n (the number of channel signals, which represents the envelope of the short-time spectrum of the speech signals), sampling period T (the sampling interval of any spectral parameter), bit number m (the number of quantization bits per one spectral parameter). Design of vocoders with pre-given properties demands a proper knowledge of relations between the quality of vocoded speech and the above variables. We are not aware of any efficient and reliable method of evaluating vocoded speech quality. Most reseachers rely on intelligibility and speaker recognition in their judgements over processed speech. There is a number of publications on intelligibility, but none of them reflects properly relations between vocoded speech intelligibility and variables n, m, T. As to speaker recognition, we can mention just one study dealing with the evaluation of recognition accuracy of LPC speech /1/. Tape recordings of 24 speakers conversing over an unprocessed channel and over an LPC voice processing system with the rate 2400 bit/s were subjected to listening tests. The listeners were 24 co-workers who attempted to identify each speaker from a group of about 40 people working in the same branch. The average duration of the speech samples was 29,8 s. Recognition accuracy was 88% for unprocessed speech, and 69% for LPC speech. No evaluation of the effect of compression ratio on the speaker recognition accuracy was made.

Note, that most of industrial vocoders use differential pulse code modulation (DPCM), in which the reference parametric signal is coded with a 3-bit logarithmic code, other signals-by 2-bit DPCM /2, 3/. This type of coding is very popular, yet, the relation between intelligibility of vocoded speech, as its main quality measure, and the number of quantization bits per one spectral parameter is interesting from the point of view of the relative information of variable m in comparison with n and T.

We also attempted to find a simple and reliable method of reducing the number of spectral parameters. This may be done by averaging subsequent samples of the amplitude spectrum (AS).

ACCURACY OF SPECTRUM REPRODUCTION

To evaluate the efficiency of the suggested averaging approach, we performed a comparative analysis of spectrum reproduction by the two methods of reducing the number of spectral parameters. A samples of amplitude spectrum of the vocoder analyser output signal were subjected to harmonic approximation (approximate representation of the spectral envelope by means of a Fourier series) by the first method, and a certain number of subsequent components was averaged by the second method, that is several subsequent samples of the amplitude spectrum from the vocoder synthesizer output signal were replaced by the average value of the analyser output samples. The test was performed on a micro-computer-aided 24-channel vocoder with a high syllable intelligibility of the vocoded speech (average score 94,3% at data rate 4800 bit/s). Its freguency range was from 100 Hz to 8 kHz. The analyser was equiped with 6-th order Bessel band-pass filters having 3 dB attenuation at 25 Hz. In the synthesizer 20 narrow-band 2-nd order filters with outputs combined in an antiphase summation were used to cover the speech bend to 5 kHz and 4 wideband filters were used in the upper freguency range. Modulation was done by digital-analogous converters.From identical speech samples, the absolute average error of one spectral component-error of spectrum reproduction - was found by

$$\delta = \frac{1}{K}\frac{1}{n_{max}}\sum_{k=1}^{K}\sum_{i=1}^{n_{max}} |F_{Ak}(i) - F_{Sk}(i)|$$

where $F_{Ak}(i)$, $F_{Sk}(i)$ - the i-th sample of the k-th spectrum frame on the analyser output, and on the synthesizer input, respectively. Fig.1 presents the dependence of δ on the number of spectral parameters n for the two methods of reducing this number.

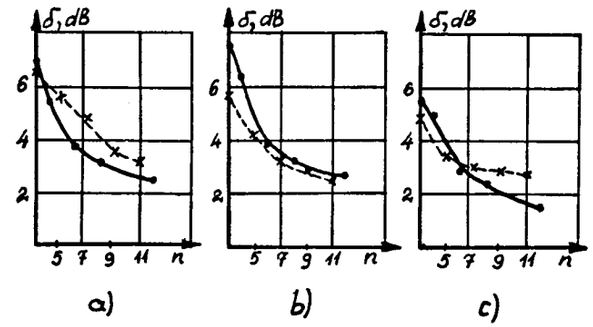

Fig.1. Accuracy of spectrum reproduction for long utterance (a), long vowels /a,i/ (b) and fricative consonant /ʃ/ (c).
• - averaging
× - harmonic approximation

Approximation by harmonic functions is only preferable for vowel phonemes, and long utterances (sentences) are more accurately reproduced by averaging subsequent spectrum samples. Human perception tests on the two methods suggest their comparative effects of spectrum reproduction. Further we restrict ourselves by the more simple method of averaging.

INTELLIGIBILITY OF VOCODED SPEECH

To evaluate the quality of vocoded speech, its syllable intelligibility S was evaluated as a more objective factor, as compared to intelligibility of phrases and words. For each test, five tables of phonetically balanced syllables of Russian words (total 250 syllables) were recorded by one male reader and processed in the vocoder. Samples of spectrum cutoffs on the output of the analyzer were

microcomputer-processed to reduce the number of quantization bits per one spectral parameter and the number of spectral parameters. The necessary sampling period was controlled manually by the switch. Samples of compressed spectrum cut-offs were fed to the synthesizer, as samples of the following spectrum cut-offs from the output of the analyser were fed to the computer. Processed syllables were recorded on a magnetic tape and played before three listeners. Syllable intelligibility for separate listeners and average intelligibility of vocoded speech were then determined. Two more simple ways of reducing bit number m were tested: a) transfer of amplitude spectrum samples (TS), when several subsequent amplitude spectrum samples on the synthesizer input $F_S(i)$ are replaced by a single amplitude spectrum sample on the analyser output, $F_A(i)$ and b) deletion of samples (DS), when separate values $F_S=0$. The determined relations between syllable intelligibility S and the number of spectral parameters n for TS, DS and AS methods are shown in Fig.2.

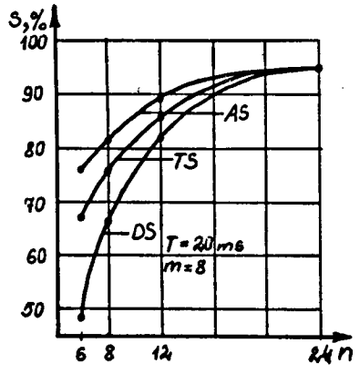

Fig.2. Dependence of syllable intelligibility S on the number of spectral parameters n

Deviation from the natural speech was observed at n<12. The undoubtful advantage of the averaging method was stated and further it was used in the evaluations of intelligibility.

A relations between syllable intelligibi- S of vocoded speech and the number of quantization bits per one spectral parameter m for given numbers of spectral parameters n are shown in Fig.3a. A distinct deviation from natural speech occurs at m< 3.

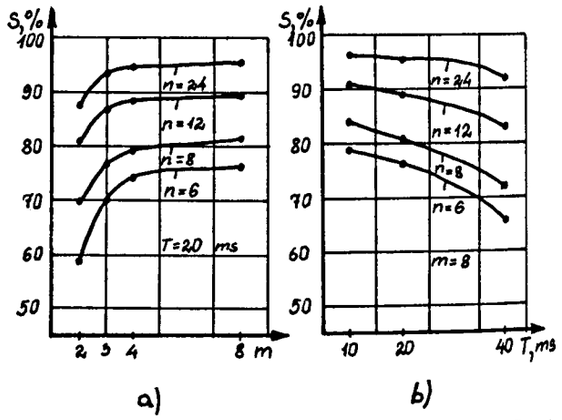

Fig.3. Dependences of syllable intelligibility S on variables m(a), T(b).

For the same number of spectral parameters n, relations between syllable intelligibility S and sampling period T were measured (Fig.3b). Deviation from natural speech at T=40 s is mainly due to the failure in the synthesis of short sounds.

SPEAKER RECOGNITION

The evaluation of speaker recognition $\mathcal{J}$ comes from two tests. First an attempt was made to find a relation between speaker recognition from unprocessed speech and vocoder-processed speech and the duration of speech sample. Speech samples were collected from 11 known speakers and 4 unknown speakers. 5 monosyllabic words each of 0.5 s average duration, 5 polysyllabic words - 1.5 s and 5 phrases - 4 s were used. Each sequence of samples was chosen at random, recorded on a magnetic tape and played before 11 listeners.

A warning concerning the unknown speakers was made. The listeners were carefully instructed not to chek off names or to use any process of elimination because some speakers were sampled more than once. Unprocessed speech, vocoded speech and monotonous vocoded speech were tested. The values of compression variables were: n= =24, m=8, T=40 ms. The relations of speaker recognition $\mathcal{J}$ and the duration of speech sample t are shown in Fig.4a.

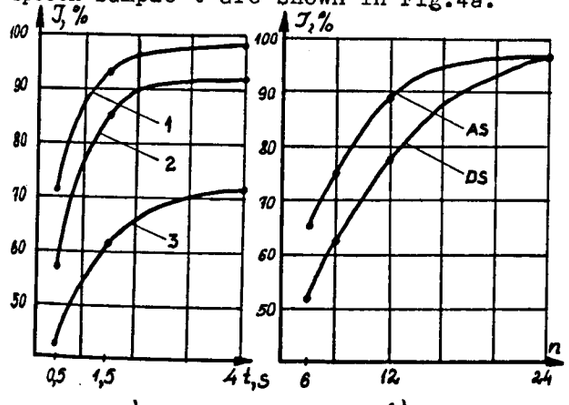

Fig.4. Speaker recognition $\mathcal{J}$ in relation to the duration of speech sample t (a) (1-unprocessed, 2-vocoded, 3- monotonous vocoded speech) and to the number of spectral parameters n (b).

The absence of fundamental frequency of the voice results in a 20% decrease in the recognition accuracy. For a successful identification and verification of a speaker from vocoded speech, just a phrase or a polysyllabic word may be used without appreciable loss of accuracy.

The second test was done by 11 known speakers. Only phrases were read, and only vocoded speech with different numbers of spectral parameters n, was evaluated. The resultant relations of speaker recognition $\mathcal{J}$ and the number of spectral parameters n for two methods of reducing this number are shown in Fig.4b. The recognition accuracy becomes significantly lower at n<12. The comparison supports the advantages of the averaging method.

CONCLUSIONS

The presented relations of vocoded speech syllable intelligibility S to the number of spectral parameters n, sampling period T and the number of quantization bits per one spectral parameter m, as well as the relations of speaker recognition $\mathcal{J}$ to the number of spectral parameters n open the ways towards designing vocoders for pregiven properties of the processed speech, which can be achieved by choosing proper design parameters. Test evaluations of the quality of vocoded speech may be described by the limiting values of variables n, m, T, which are: n≥12, m≥3, T≤40 ms. Vocoded speech becomes significantly unnatural whenever one of the variables exceeds it limiting value. We underline the advantages of the method of reducing the number of spectral parameters by averaging subsequent samples of speech amplitude spectrum. Problems of speaker recognition and verification from vocoded speech may be solved on single utterunce from 2 to 4 s.

REFERENCES

/1/ A. Schmidt-Nielsen, K. Stern, "Identification of known voices as a function of familiarity and narrow-band coding", JASA, Vol.77, pp.658-663, 1985.

/2/ B. Gold, P.E. Blankenship, R.J. McAulay," New applications of channel vocoders", IEEE Trans. ASSP, Vol.29, No.1, pp.13-32, 1981.

/3/ J.N. Holmes, "The JSRU channel vocoder", IEE Proc. Communications, Radar and Signal Processing, 127, Pt.F, pp.53-60, 1980.