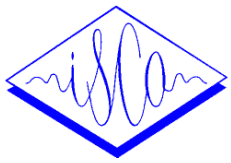

3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

# TIME-SPECTRAL APPROACH TO COMPILING SPEECH RECONSTRUCTION

Alexander Osipov and Vladimir Zentsov

St.Petersburg Air- and Spacecraft Equipment Academy,
St.Petersburg , Russia

## ABSTRACT

The method of compiling speech reconstruction with its modifications which differs in favour way in its good articulation, natural properties of speech synthesized as well hardware implementation simplicity is proposed. The method is a combination of direct and parametric speech encoding and based on spectral-time approach when time intervals between signal zero crossings are used as spectral expansion ones and spectral expansion coefficients - as spectral parameters.

Keywords: compiling speech reconstruction, spectral expansion in functional series, speech articulation and quality estimates.

Compiling speech reconstruction implementation is based on mostly either time or spectral approach /1/. Moreover, some direct speech encoding methods (like clipping ) produce considerable articulation and quality reduction of the speech synthesized. The time - spectral technique for compiling speech reconstruction which combines elements of direct and parametric speech encoding is being proposed. Here, a sequence of time intervals T(i) between speech

Signal zero crossings is used as spectral expansion intervals and spectral expansion coefficients with approximation sum - as spectral parameters. The number of basis functions is defined with the accuracy of an approximation and has been chosen with respect to the properties of human hearing. A speech signal on T(i) can be expressed with the approximation sum S(k) as follows ( in discrete way) /2/:

$$f(k) = S(k) \cong \sum_{j=0}^{N-1} [c(j) * g(j,k)] \qquad (1)$$

where:
- g(j,k) - the basis function set;
- c(j) - the spectrum of f(k) with the basis g;
- k - a time parameter.

The speech signal approximation with harmonic, piecewise-linear (Schauder) and piecewise--quadratic basis functions is shown accordingly on fig.1,2 and 3. The speech generators based on piece-wise linear expansion as well as ones based on delta modulation are the linear, therefore weak noises at the speech synthesized are discovered in both cases for a lack of smoothness in the signal while piece-wise polynomial

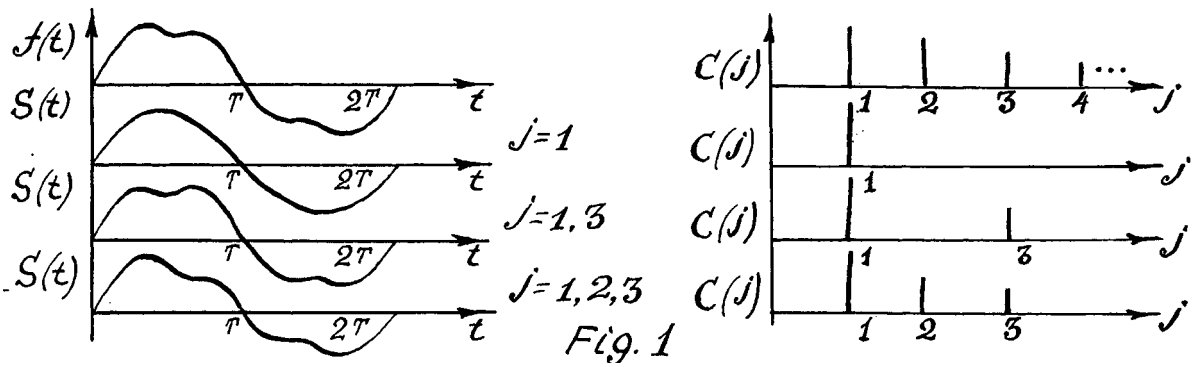


Fig. 1

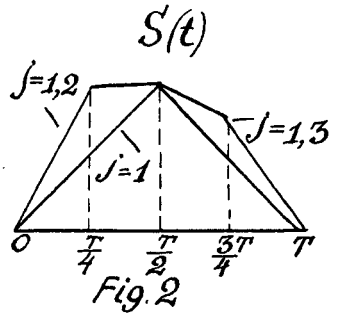

Fig. 2

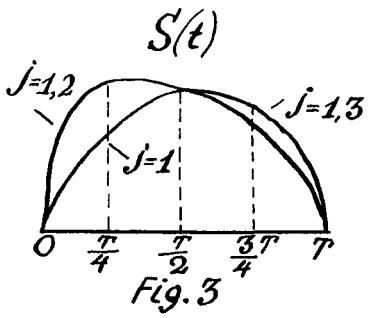

Fig. 3

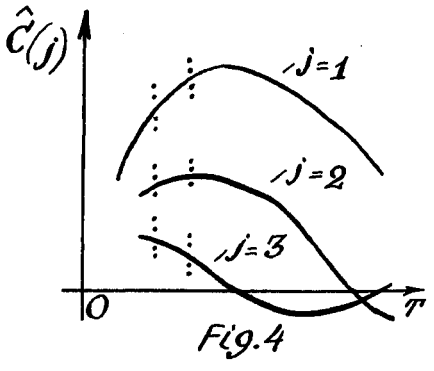

Fig. 4

function expansion enables to avoid the drawback. In order to encode and reconstruct speech signals it is proposed to take into consideration word spectrum statistics. The statistics is defined on equal intervals of some samples of a word pronounced by the speaker assigned and determined as an average of c(j) spectral coefficient values. The statistics of the first three Schauder spectral coefficients for a word "mead" is shown on fig. 4. Generally speaking, a word spectrum statistics ( as well as statistics for a number of words or a phrase) can be represented with either a table or functional approximation. To obtain articulation and quality estimates of the speech reconstructed there are objective and subjective - stochastic methods. Experimental obtaining subjective estimates is a very troubling procedure, but these estimates are reliable provided to large amount of data processed only. Objective estimates enable to judge of speech articulation and quality as well to be reliable enough for taking decision on actions during the experiment. The new objective estimate r(i) for phonetic balancing of word samples compared has been proposed:

$$r(i)=(1\backslash N)*\sum_{k=1}^{M}[\mathrm{Log}(|f(k)|\backslash(|\tilde{f}(k)|*a))]^2 \qquad (2)$$

where

M - the number of samples;

f(k) - the initial speech signal;

$\tilde{f}(k)$ - the reconstructed speech signal;

a - a coefficient for normalizing the both signal powers.

The estimate is a measure unlike other objective ones (e.g. correlation criteria) and based on biological ideas about speech perception processes. So, a lack of distortion could be interpreted as zero loudness as well the measure itself - as a monotonously decreasing function of distortion loudness , which has a zero minimum. The experiments for comparing r(i) values with real perception parameters have been carried out and the classification of how reconstructed speech perception levels depends on the criterion considered has been obtained. Computer simulation of near 20 encoding and reconstruction models , among of them, the speech models whose parameters do not coordinate with the initial speech spectrum as well ones of males and females has been carried out. The best was the speech reconstructed on the base of piecewise-quadratic basis functions. The digital speech synthesizer on the base of the methods considered consisting of spectral and time encoders, a memory, a basis function generator has been proposed. Experiments show that the speech reconstruction method considered competes against well-known parametric one concerning a data stream rate (up to 6000 bit/sec) but has better articulation properties and natural properties of the speech synthesized as well its hardware implementation is simpler. Such low data stream rate enables users to manipulate with vocabularies of larger size than usually.

## REFERENCES

/1/: Flanagan J.L.: Speech Analysis,Synthesis and Perception. 2 nd edition, Springer Verlag, New York, 1972.

/2/: Zentsov V.A.: Feature Extraction with Piece-Wise Polynomial Function Sets. Proc. Mustererkennung 1992, 14 DAGM-Symposium, pp. 437-442, 1992.