БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ . BULGARIAN ACADEMY OF SCIENCES

ПРОБЛЕМИ НА ТЕХНИЧЕСКАТА КИБЕРНЕТИКА И РОБОТИКАТА, 50
PROBLEMS OF ENGINEERING CYBERNETICS AND ROBOTICS, 50

София . 2000 . Sofia

Speech Coding with Wavelet Packet Excitation Signal Compression*

Atanas Gotchev, Elena Rangelova, Zdravko Nikolov

Institute of Information Technologies, 1113 Sofia

I. Introduction

Digital speech coding is important problem in the area of network speech applications where there is a necessity of transmission of big amount of data through limited bandwidth channels. Investigations in this field aim at finding new methods for signal *compression* and compatibility with the *network transfer protocols*.

Digital speech coding involves sampling and amplitude quantization of the speech signal using minimum number of bits, while preserving the quality of reconstructed speech.

For achievement of good compression rates three basic manners in speech coders are used:

- ♦ representation of frequency-domain characteristic of speech signal
- ♦ providing waveform coincidence
- ♦ coder's optimization according to perceptual properties of the ear

According to the coding mechanism there are two types of systems: *waveform coders* and *vocoders*. The latter achive better compression rates and work with speech models the most famous type being linear predictive vocoders.

II. Linear-predictive vocoders with differentiated glottal wave excitation

Linear prediction (LP) is the most widely used method of speech processing during the last 20 years. It is based on Fant's linear voice production model [1]. According to this model sounds are generated by vocal tract excitation from the source signal. The latter is a periodic sequence of impulses for voiced sounds and a random noise for unvoiced sounds. The *vocal tract* is modeled as an all-pole system; *glottal model* is represented as two-pole low-pass filter and the *lip radiation* as differential unit (Fig. 1a). The system

* This research was supported by the National Science Fund under grant No I 625/96.

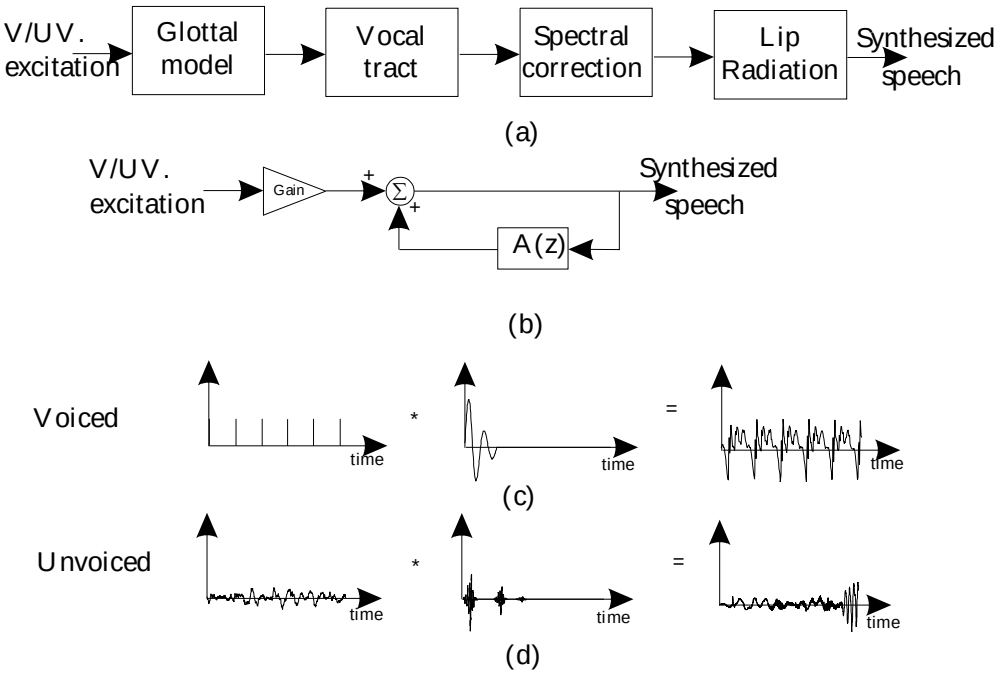

Fig. 1

can be reduced to an all-pole model by cancellation of one pole by one zero from lip radiation model (Fig. 1b). Figures 1c and 1d shows generation mechanism for voiced and unvoiced speech according to the described model. Classical linear prediction method estimates vocal tract parameters (LP coefficients). According to shape of *A(z)* filter and length of data segment there are different sets of coefficients obtained using different methods. For long segments there are *autocorellation* and related to them *reflection coefficients* [2]. For short speech segments there are algorithms for generation of *covariance coefficients* [2]. There are different representations of these coefficients that are less sensitive to quantization errors, like Log Area Ratios (LAR) and Line Spectrum Pairs (LSP). Estimation of 8-14 LP parameters is usually enough for good representation of vocal tract.

Systems with different grades of complexity according to coding of *excitation source* have been proposed. Initially the source was represented by two-state scheme: pulse sequences or white noise. For example according to the American Federal Standard FS1015 [3] 10 LP vocal tract parameters are excited by a source represented by gain parameters, pitch and flag for segment type (voiced/unvoiced). Main drawback of this method is wrong classification for voiced/unvoiced segment. Perfect generator is *residual signal* obtained as a difference between speech signal and its LP model. Residual signal carries all information that has not been captured by LP analysis: phase, pitch, zeros due to nasal sounds, etc. Vocoders with similar source are called Residual Excited Linear Prediction (RELP) which operate among 6 and 9.6 KB per s. Decreasing of bit rate is achieved by down-sampling of the residual signal and it's bandwidth is restricted to 800 Hz. However this decreases the speech quality. Excitation models derived by feedback loops known as analysis-by-synthesis scheme are proposed to avoid this problem. Two LPs with long-term and short-term periods, represent the pitch and the formant structure respectively. Weight filter $W(z)$ "distort" error so that the quantization noise be masked by the high energy formants. Excitation source forms or selects from dictionary excitation sequence so that the Mean Squared

Error (MSE) be minimized. This scheme has been proposed first by [5] and it is a milestone in speech coding and provided a big quality improvement. According to excitation models there are: *Multipulse excitation* (MPE), *regular pulse excitation* (RPE) [6] and *code excitation* (CELP) [7].

An alternative approach for speech coding aims at exact modeling of vocal tract features and using them in inverse filtering (IF) for estimation of glottal waveform (GW) which represents the function of the voice source [8]. It gives information for phonation type, emotional status and other individual speaker's characteristics [9]. Restoration of coded speech signal using excitation signal that is close to glottal waveform leads to more naturally sounding. This approach allows separation of phonation and speech quality of different speakers [10]. Transmission (or storage) of exact GW requires additional resources. A Differentiated Glottal Wave (DGW) is used in speech encoding tasks. It represents voice source function and lip radiation. Its typical shape is shown on Fig. 2.

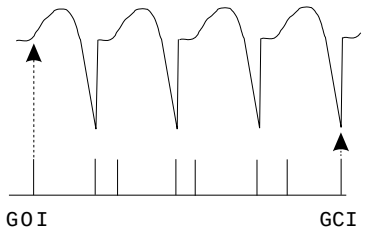

Fig. 2

The most important instants are glottal closure instant (GCI) and glottal opening instant (GOI). Between these instants the glottis is closed and vocal tract system is in free oscillating state. The exact determination of these instants is very important for adequate vocal tract modeling [11].

III. Compression of DGW with wavelet packets

Wavelet packets and wavelet packet transform (WPT)

Wavelet packet (WP) w is an integrable function with finite energy, zero mean and well localized in both space and frequency. It may be assigned three parameters: scale (time uncertainty), frequency and time position.

Fast wavelet packets can be defined by a pair of quadrature mirror filters (QMF) [12].

Let $h=\{h_j\}$ is a low-pass filter with the following properties:

(1)
(a) for $\varepsilon>0$ $\sum_j |h_j||j|^{\varepsilon}<\infty$;
(b) $\sum h_{2j+i}=1/\sqrt{2}$ for $i=0$ and 1;
(c) $\sum h_j h_{j+2k}=\sigma_k$, where σ is the Kronecker symbol.

The (a) property is related to filter coefficients decay while (b) and (c) concern its orthogonality.

Let $g=\{g_j\}$ is defined by:

(2) $$g_j=(-1)^{1-j}h_{1-j}.$$

Two sequences constitute QMF pair. These filters avoid definition of two convolution-decimation operators:

(3) $$\boldsymbol{H}x(t)=\sum_j h_j x(2t-j) \text{ and } \boldsymbol{G}x(t)=\sum_j g_j x(2t-j)$$

and their adjoints:

(4) $$\boldsymbol{H}^*x(t) = \frac{1}{2}\sum_j h_j\, x\left(\frac{t}{2} + \frac{j}{2}\right) \text{ and } \boldsymbol{G}^*x(t) = \frac{1}{2}\sum_j g_j\, x\left(\frac{t}{2} + \frac{j}{2}\right)$$

Let suppose that the sequences h and g are finite and define:

$$\Phi = \lim_{n\to\infty} \boldsymbol{H}^n S$$

where S is indicator function of $[-1/2, 1/2]$. This is the only fixed point of the equation $\Phi = \boldsymbol{H}\Phi$. A fast wavelet packet is the image of Φ under any finite composition of $\boldsymbol{H}$ and $\boldsymbol{G}$, possibly translated by an integer and unitary dilated by an integer power of 2. All wavelet packets are orthogonal to their dilated and translated versions. Order the frequency, scale and position parameters of wavelet packets $w_{f,s,p}$ we can write $w_{0,0,0}(t) = f(t)$; $w_{2f,0,0}(t) = \boldsymbol{H}w_{f,0,0}(t)$; $w_{2f+1,0,0}(t) = \boldsymbol{G}w_{f,0,0}(t)$, etc.

Wavelet packets allow approximation of continuous function $x \in L^2(\boldsymbol{R})$ to accuracy $O(2^{-L})$ by the l^2 sequence of inner products. We can recursively compute from these the other WP coefficients as follows:

(5)
$$\langle x, w_{2f,s+1,p}\rangle = \sum_j h_j \langle x, w_{f,s,2p+j}\rangle,$$
$$\langle x, w_{2f+1,s+1,p}\rangle = \sum_j g_j \langle x, w_{f,s,2p+j}\rangle.$$

Operators $\boldsymbol{H}$ and $\boldsymbol{G}$ and their adjoints refer to the discrete sequences (signals) too:

(6)
$$\boldsymbol{H}:\ l^2 \to l^2,\ \boldsymbol{H}x_n = \sum_j h_j x_{2n+j},$$
$$\boldsymbol{G}:\ l^2 \to l^2,\ \boldsymbol{G}x_n = \sum_j g_j x_{2n+j}.$$

Wavelet packets form a dictionary of basis functions. Their approximations by 2^N vectors in R^N form a set of $N \lg N$ vectors. Vectors and their coefficients are disposed in a binary tree nodes. Nodes of one level correspond to one scale and differ by frequency localization, and coefficients in a single node differ by time position (Fig. 3).

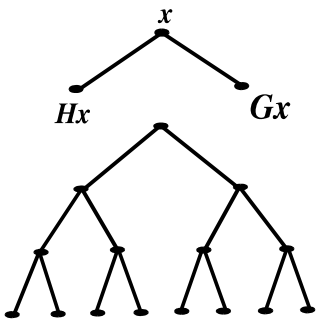

Fig. 3

Each node is an orthogonal sum of its sons. We can obtain a basis from dictionary by connection of tree branches. Different bases dictionaries are derived from different QMF pairs, which form a WP basis *library*.

The best bases selection is carried out by minimization of additive information measure for all bases of a dictionary in R^N. Usually the measure is of entropy type [12]. Such a procedure can be repeated for the bases library.

Wavelet packets are widely used in signal compression because of good localization and possibility of optimal decomposition choice. Compression is achieved by signal reconstruction using the k biggest in absolute value WP coefficients. This approxima-

tion is optimal in mean square sense. If the analyzed signal has marked peculiarity then more coefficients are needed and the opposite is valid for flat signals. In this way WP packets focus to significant parts of the signal in information sense. A threshold selection to eliminate non significant WP coefficients is very important.

The minimum description length principle

Differentiated glottal waves are characterized by abrupt transitions round to closure instances and comparatively slanting sections in closed phase. Hence we can expect that wavelet transforms can represent DGW by few coefficients due to their capability of singularity detection.

Let us consider DGW as a discrete model of *signal-noise* mixture:

$$y = x + \varepsilon, \text{ where } y, x, \varepsilon \in R^N, N = 2^{n_0}. \tag{7}$$

The vector y represents the noisy observed signal, x is information signal, ε is white Gaussian noise with unknown variance σ^2:

$$\varepsilon \sim N(0, \sigma^2 \boldsymbol{I}). \tag{8}$$

Noise component is generated by inadequacy of vocal tract model or rounding errors.

We can generate a library of m orthogonal WP bases: $\alpha = \{A_1, A_2, A_3, \ldots, A_m\}$ where $A_1, A_2, A_3, \ldots, A_m$ differ by type of QMF's and comprise the best basis from dictionary m.

We suppose the signal can be completely represented by k coefficients of a basis A_m.

$$x = \boldsymbol{W}_m \alpha_m^{(k)}, \tag{9}$$

where $\boldsymbol{W}_m \in \boldsymbol{R}^{N \times N}$ is an orthogonal matrix whose columns are the basis vectors of A_m, and $\alpha_m^{(k)} \in \boldsymbol{R}^{N)}$ is the vector of expansion coefficients with only k non-zero elements. In the expression (9) the actual values of k and m are not known.

The idea for determination of k and m by simultaneous noise suppression and signal compression algorithm is developed by Saito in [13]. One of the most suitable criteria for this purpose is the so-called *Minimum Description Length Principle* (MDLP) [14]. According to the latter, minimal length of description of numbers or vectors, i.e. code length in bits is found. In the Saito algorithm code lengths for representation of the all components of model (7) are estimated.

Let assume L as the operator for determination of code length. Total code length is composed of the following terms.

1. Code length of the integers k and m: $L(k,m)$;
2. Code length of a k number real coefficients of the best basis: $L(\hat{a}_m^{(k)} | k, m)$;
3. Code length of the noise variance estimation: $L(\hat{\sigma}^2 | k, m)$;;
4. Code length of the deviation of the observed signal y from the estimated signal x (9): $L(y | \hat{\sigma}^2, k, m)$.

The total code length to minimize is:

$$L(y, \hat{a}_m^{(k)}, \hat{\sigma}^2, k, m) = L(k, m) + L(\hat{a}_m^{(k)}, \hat{\sigma}^2 | k, m) + L(y | \hat{a}_m^{(k)}, \hat{\sigma}^2, k, m). \tag{10}$$

By assumption of white Gaussian noise it can be seen that maximal likelihood estimation of variance is obtained by sum of the $N-k$ squared least coefficients [13]:

$$\hat{\sigma}^2 = (1/N) \| \alpha_m^{(N)} - \alpha_m^{(k)} \|^2. \tag{11}$$

Terms analysis by MDLP lead to the following expression :

(12) $$L(k^*, m^*) = \min_{\substack{0 \le k < N \\ 0 \le m \le M}} \left((3/2)\, k \lg N + (N/2) \lg \| \alpha_m^{(N)} - \alpha_m^{(k)} \|^2 \right).$$

Minimizing of the latter by finding the best k^* and m^* simultaneously. Reconstructed signal is obtained by:

(13) $$\hat{x}^{(k)} = \mathbf{W}_{m^*} \alpha_{m^*}^{(k^*)}.$$

IV. Experimentional results

For present investigations compactly supported wavelets, which are represented by finite length filters, are used. Basis library consists of Daubechies wavelet family, less asymmetric wavelets and coiflets [15, 16]

We apply the method described in section III (equations (10)–(13). The entropy minimum is used as the best basis criteria. Each QMF pair from the library leads to decomposition upon the bases dictionary. From the current dictionary (with number m) the minimum entropy basis is selected. The obtained basis determines the value of k that minimizes expression (12). Passing thought all the bases from the library we can obtain the (k^*,m^*) pair, where k^* is the number of essential coefficients and m^* is the number of bases.

Synthesized signals

There are a set of 16 DGW (each of length 512 samples) obtained after IF of synthesized vowels /a/, /e/, /i/ and /u/ [8].

Results of the processing of synthesized signals by the method based on MDLP are displayed in Table 1. The number of the essential coefficients k^* is shown.

Table 1

Signal x10-3	Wavelet	k*	R M S E
mei1	S20	30	8.5
mei2	S10	24	8.7
bab1	D16	11	14.8
bab2	D14	12	13.3
babi3	S10	28	10.3
babi4	S10	29	6.2

Table 2

Signal	Wavelet	k*	R M S E
/a/	D20	94	0.1
/e/	S6	78	0.2
/i/	S5	97	0.19
/u/	S5	159	0.1
/aa/	C4	96	0.2
/ee/	D14	79	0.2
/ii/	S8	105	0.4
/uu/	S4	156	0.6
/uu1/	D 8	118	0.9
/ii1/	S4	78	1.9

The comparison reveals that wavelet packets representation of DGW uses few coefficients. Due to the higher frequency resolution in scales more efficient grouping of information contents in basis vectors is achieved. Entropy as an information measure leads to best basis finding too. Minimum description length principle combines coding with noise suppression without the necessity of separate noise estimation.

Natural signals

The database consists of six DGW (lenght of the analyzed segments is 512 samples), obtained via IF of voiced sounds of two speakers (male and female). Signals are shown in Table 2 together with root mean squared error (RMSE) between DGW after IF and DGW after wavelet packet reconstruction from reduced set of coefficients. The latter is vastly less than number of coefficients obtained by synthesized signals analysis. Mean squared errors are comparable in the two cases because of the existence of more noise components in natural sounds which influence on processing.

Fig. 4 shows inverse filtered DGW, reconstructed DGW after WPT compression and the difference between them. Recovered signal is very close to the original one and is achieved by a low number of coefficients. The corresponding glottal waves are almost of the same shape according to possibility of WP transform for detection of local features with good time-frequency resolution. The best WP basis "finds" essential high-frequency components too.

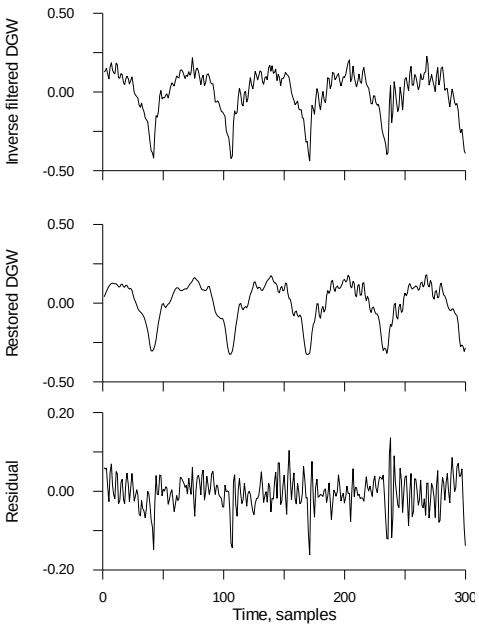

Fig. 4

V. Conclusions

The potential of WPT to compress effectively DGW is reported in the present paper. This transform has been chosen having in mind the possibilities of preserving points in DGW which enables the natural sounding of the reconstructed signal.

The results of DGW compression make possible the construction of low and medium bit-rate speech coders with equivalent or higher quality in comparision to the present CELP coders in similar transmission rate.

References

1. Fant, G. Acoustic Theory of Speech Production. Gravenhage. The Netherlands: Mounton and Co., 1960.
2. Markel, J., A. Gray. Linear Prediction of Speech. New York, Springer-Verlag, 1976.
3. Federal Standard 1015. Telecommunications: Analog to digital conversion of radio voice by 4800 bit/second code excited linear predictive coding, national communication system. National Communication System-Office of Technology and Standards, Nov. 1984.
4. Atal, B., J. Remde. A new model for LPC excitation for producing natural sounding speech at low bit rates. – In: Proc. ICASSP-82, Apr. 1982, 614-617.
5. Singhal, S., B. Atal. Improving the performance of multipulse coders at low bit rates. – In: Proc. ICASSP-84, p.1.3.1, 1984.
6. Kroon, P., E. Deperette, R. J. Sluyeter. Regular-pulse excitation - a novel approach to effective and efficient multipulse coding of speech. – IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-34, No 5, Oct. 1986.
7. Schroeder, M. R., B. Atal. Code-excited linear prediction (CELP): High quality speech at very low bit rates. – In: Proc. ICASSP-85 (Tampa, FL, Apr. 1985), p. 937.
8. Fujisaki, H., M. Ljungqvist. Proposal and evaluation of models for the glottal source waveform. – In: Proc. ICASSP, 1986, 1605-1608.
9. Rosenberg, A. Effect of glottal pulse shape on the quality of natural vowels. – J. Acoust. Soc. Am., **49**, 1971, 583-590.
10. Riegelsberger, E. L., A. K. Krishnamurthy. Glottal source estimation: Methods of applying the LF. – In: Model to Inverse Filtering. Proc. ICASSP 1993, 542-545.
11. Gothcev, A. Determination of closure glottal instance via wavelet transform. – Techn. Ideas, 1995, No 2, 28-42, (in Bulgarian).
12. Coifman, R. R., M. V. Wickerhauser. Entropy-based algorithms for best-basis selection. – IEEE Trans. Info. Theory, vol. 38, 1992, 713-718.
13. Saito, N. Simultaneous noise suppression and signal compression using a library of orthonormal bases and the minimum description length criterion. – In: Wavelets in Geophysics (E. Foufoula-Georgiou and P. Kumar, eds.), San Diego, CA, Academic Press, 1994, 299-324.
14. Rissanen, J. Universal coding, information, prediction, and estimation. – IEEE Trans. Inf. Theory, **30**, 1984, , 629-636.
15. Daubechies, I. Orthonormal bases of compactly supported wavelets. – Comm. in Pure and Applied Math., **41**, 1988, 909-996.
16. Daubechies, I. Ten lectures of wavelets. Philadelphia, SIAM, 1992.

Кодирование речевых сигналов при помощи компрессии волновых пакетов

Атанас Гочев, Елена Рангелова, Здравко Николов

Институт информационных технологий, 1113 София

(Резюме)

В работе предложен метод компрессии дифференциальных глотисных волн, полученных после инверсной фильтрации речевых сигналов на основе декомпозиции волновых пакетов. Для определения коэффициентов применяется принцип описания с минимальной длиной. Результаты показывают применимость метода в кодировании речевых сигналов в компьютерных сетях.