Proceedings of the 2^{nd} European DSP Educational and Research Conference, September 1998, Paris, pp. 83-87.

Internet Telephony Vocoders

Vladimir Babkin, Vladimir Ivanov, Arthur Lanne, Ilya Pozdnov
DSP Center
St. Petersburg University of Telecommunications
E-mail ivanov@internettelecom.com,
pozdnov@internettelecom.com

ABSTRACT:

In work reported we present the results of research on speech compression algorithms for Internet telephony systems. The main features and key requirements to speech compression algorithms for systems with packet data transmission are formulated. On basis of these requirements two CELP based algorithms adapted to work in Internet telephony systems are developed and discussed. The results of preliminary researches on development of PWI variable rate algorithms are reported.

1. INTRODUCTION

Development of communication systems with packet data transmission and methods of packetized voice transfer for realization of telephone talk over Internet has led to necessity to develop speech compression algorithms adapted to condition of Internet. The purpose of the given work is to determine the requirements to speech compression algorithms, intended to work in Internet telephony systems and present results on research and development of CELP and PWI based algorithms.

2. FEATURES OF SPEECH COMPESSION ALGORITHMS FOR INTERNET TELEPHONY SYSTEMS

Modern state of Internet can be characterized by

- The dynamically varied traffic with non-negligible probability of overloads;
- Large time of packet delivery from end to end frequently exceeding delay acceptable for comfortable duplex connection;
- Rapid dynamics of this delay during one session of connection;
- Packet losses increasing in conditions of traffic growth;

A growth of network traffic leads to increase of network overload probability that are resulted in increase of packet delivery time and even to packet losses during transfer. Experimental distributions of delivery delay of packets (a) and number of successive lost packets (b) for various network load are shown in Fig.1.

The analysis of network functioning shows that data transmission over Internet is connected with essential delay of packet delivery. This delay exceeds in several times 10 - 30 ms delay introduced by usual speech compression algorithms. In this case application of low delay algorithms is not acute necessity. One of the most important requirements to speech compression algorithms for Internet telephony systems is the reduction of coding bit rate at preservation of toll or near to toll quality of synthesized speech. The asynchronous mode of the data transfer with an inherent possibility to change the size of a packet makes application of speech compression algorithms with variable bit rate natural in Internet telephony systems. Introduction of algorithms with variable bit rate partially solves the problem of network load reduction. However, the probability of packet losses is not negligible even in the case of network load reduction.

Our investigations show that distortion and reduced intelligibility of synthesized speech in Internet telephony systems are basically caused by interruption of speech data flow forwarded to decoder due to packet loses during transmission or when delay of packet delivery exceeds maximal allowed time. Data transfer protocols usually used in Internet telephony systems such as User Datagram Protocol (UDP) provide error free delivery of data. In these conditions the error correction coding of the speech data is not expedient. Moreover the redundancy introduced by error correction coding increases volume of data transmitted via network and the probability of new packet losses. The distribution of the number of successive lost packets given in Fig.1b shows that the probability of single losses is higher than probability of multiple losses. Apparently, with further development of Internet and growth of its throughput the single losses will play more prevailing role.

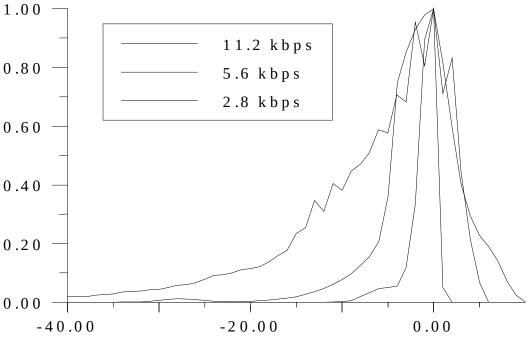

Fig 1a. Packet delay

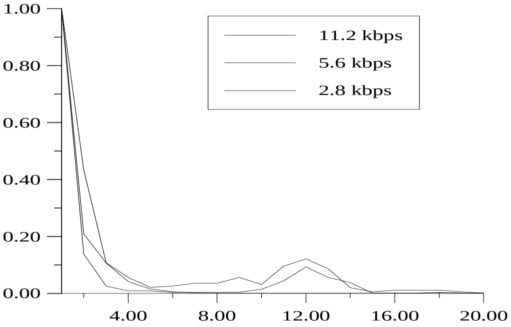

Fig 1b. Number of successive lost packets

Thus, above-mentioned analysis allow us to formulate the optimum strategy to develop speech compression algorithms for the systems with packet data transmission such as Internet telephony:

- Bit rate of the transmitted data should be as little as possible maintaining toll quality of the synthesized speech;
- The frames of speech having different information measure should be coded with various bit rate;
- The algorithm should be tolerant to losses of packets containing speech data;
- The algorithm should include a procedure responsible for synthesis of speech-like signal on the frames corresponding to lost packets. Single losses should not considerably reduce intelligibility of synthesized speech.

3. CELP BASED ALGORITMS FOR INTERNET TELEPHONY SYSTEMS

In Internet telephony systems [1] two speech compression algorithms based on CELP model are used: low rate near to toll quality 4.6 kbps and toll quality 7.6 kbps. In ideal conditions (no losses) the quality of the synthesized speech is comparable with G.723 and G.729 recommendations of ITU correspondingly [2,3]. Both algorithms have similar design. The simplified signal flow and block diagram of coder are described in Fig.2.

To provide steady work of the decoder when speech data are lost we decline to exploit any quantization schemes adaptive to the previous values of speech parameters.

The second peculiarity of speech compression algorithms presented in this section is the design of long term predictor (LTP) [4], tolerant to energy variations caused by speech packet losses. The scheme of LTP can be explained as follows. The output signal $s(n)$ of typical LTP [4] of fifth order used for example in [2] is given by expression

$$s(n)=\sum_{i=-2}^{i=2}\beta_i\, p\ (n+i-L)\ ,\quad (1)$$

where β_i - gains of LTP, $p(n)$ - excitation signal generated on the last frame of the analysis, L - pitch lag. If speech packet loss occurred on previous frame of the synthesis excitation restored in the decoder does not correspond on level and waveform to excitation in the coder. Due to LTP gains determined by Eq.(1) this distortion is extended over several successive frames.

To avoid this undesired phenomenon we modified LTP scheme in such way that LTP gains β_i do not depend on energy of previous frame. The output signal $s(n)$ of proposed LTP can be described by relation:

$$s(n)=\frac{E_r}{\sqrt{\frac{1}{N+5}\sum_{i=-2}^{i=N+2}p^2(i-L)}}\sum_{i=-2}^{i=2}\beta_i\, p\ (n+i-L)\,,\ (2)$$

where N - length of a frame, E_r - energy of residual signal determined as [5,6]:

$$E_r=\sqrt{E\prod_{i=0}^{9}\left(1-r_i^2\right)}\,,$$

here E - energy of a speech frame, r_i - reflection coefficients corresponding to quantized spectral parameters.

Fulfilled experiments have shown that in condition of packet losses the purposed scheme of LTP with gain β_i determined by Eq.(2) does not result in noticeable change of a synthesized speech level and outperforms LTP used in G.723 and G.729.

The third peculiarity of described algorithms is the novel structure of the algebraic codebook based on spherical tetra-code [7] used to form stochastic part of excitation signal in CELP model.

The reduction of average bit rate of algorithms is carried out by application of Voice Activity Detector at the coder. When pause is detected the parameters describing environmental noises are transmitted to the decoder to produce comfort noise. Assuming ratio of active speech and pauses during human dialogue to be equal the average bit rate for presented algorithms is comparable to 2.4 kbps and 4 kbps.

The important component of speech compression algorithms for Internet telephony systems is the procedure of speech restoration on frames corresponding to lost packets. The conventional strategy to restore excitation signal in CELP model concludes in extrapolation of speech parameters from previous frames with exponential decay of signal level. Similar technique is used in G.729 [3] and proprietary algorithm SX7300 of Lucent Technologies targeted to application in systems with packet data transmission.

The dynamically varied delay of packet delivery results in inevitable necessity to use a dynamic buffer containing speech data to smooth irregular data flow incoming from network. In the case of single packet loss and not empty buffer it is possible to obtain the information not only about past, but also about future speech parameters and to perform their interpolation. Carried out tests have shown that quality of the synthesized speech on frames corresponding to single losses can be considerably improved interpolating the following speech parameters: line spectral frequencies, energy of a frame E, pitch L and gain of algebraic codebook. Interpolation of LTP gains β_i does not result in noticeable improvements of synthesized speech. Described scheme has been implemented in Internet telephony system [1] and has shown good intelligibility of the synthesized speech in conditions of 30% of lost packets.

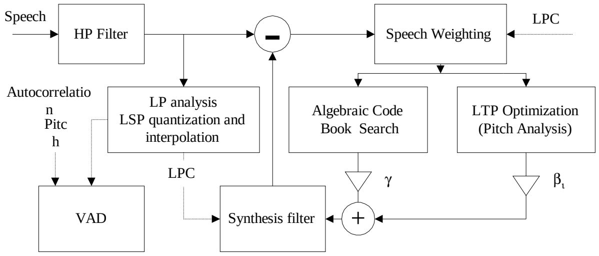

Fig. 2 Signal flow and block diagram of CELP algoritms

Both algorithms are implemented on TMSC320 line of Texas Instruments DSP. Computational load of TMS320C31 run at 60 MHz is equal to 74% for low rate near to toll quality 4.6 kbps algorithm and 81% for toll quality 7.6 kbps algorithm. We intend to complete implementation of discussed algorithms on TMS320C548 in near future.

4. PERSPECTIVE RESEARCHES

Nowadays at DSP center research and development of new generation of variable bit rate speech compression algorithms is actively carried out. These algorithms are pointed to Internet telephony systems and expected to have maximal bit rate comparable to 3-3.5 kbps. These algorithms are based on classification principle of input signal. For the purpose of compression input signal is assumed to be several types: Voiced, Unvoiced, Transition, Pause. Depending on classification each type is compressed with different bit rate. Frames classified as Voiced, are coded based on Prototype Wave Interpolation (PWI) technique with bit rate 3-3.5 kbps. Frames classified as Unvoiced, are described by spectral parameters and energy level and coded at 1 kbps. Transition frames are coded on the basis of CELP model with expected bit rate 3-3.5kbps. Pauses are detected by Voice Activity Detector. The parameters of environmental noise for comfort noise generation at the decoder are transmitted at the bit rate of 0.2-0.3 kbps.

To compress excitation signal $x(n), \quad n = 0,..,N$ for speech frame classified as Voiced the problem on optimum bases selection to represent signal was investigated. According to PWI technique one period of excitation signal $x(n), \quad n = 0,..,L-1$ length of pitch L is to be approximated and compressed [8]. For the further analysis this signal is transformed to fixed length vector $x(n), \quad n = 0,..,255$ dimensions of 256 and approximated by the functional series $\sum_k a_k \varphi_k(n)$ where $\{\varphi_k\}$ - system of functions. To determine optimal system $\{\varphi_k\}$ we supposed these functions to satisfy the following conditions:

- To represent signals $x(n) \in X$ as precise as possible, X - set of excitation signals;
- To allow simple calculation of factors on the given samples;

- Not to be sensitive to variations of factors a_k.

Thus, for the signal compression problem it was necessary:

- To find subspace, on which the set of signals X is to be projected;
- To select basis in this subspace;
- To find an algorithm of fast computation of factors a_k.

The appropriate choice of subspace and the way of calculation is responsible for accuracy of approximation. The sensitivity and complexity of calculation (computational cost) depends on choice of basis. The task was solved on the basis of the width theory [9]. To select optimal subspace the following problem was considered

$$\max_{x\in X}\min_{a_1}\left\| x(n)-\sum_{k=0}^{N} a_k\,\varphi_k(n)\right\| = \min_{G\varphi},$$

where $\varphi_k(n)\in G\varphi$.

The obtained results of our study can be formulated as follows.

- Extreme subspace consists of trigonometric polynomial, where the search of bases was carried out;
- Good result on accuracy of approximation and most simple way of calculation of factors of approximating polynomial provides interpolation Lagrange polynomial.
- The interpolation with following decimation on a basis of DFT (FFT) slightly outperforms Lagrange polynomial in the meaning of accuracy, but loses in complexity of computations.
- The optimum result (realization of a width) provides DFT of vector $x(n),\quad n=0,..,255$ with subsequent rejection of the high frequency terms.
- The sensitivity of Fourier and Lagrange basis to variations of factors is almost identical.

5. CONCLUSION

Most important requirements to speech compression algorithms intended to work in Internet telephony systems have been formulated on the basis of analysis of Internet network functioning. Two CELP based algorithms of speech compression met this requirements have been implemented. In these algorithms we have proposed and implemented long term predictor which is steady to packet loss. The results of development of variable rate speech compression algorithm with overage bit rate 1.5-2.5 kbps have been discussed. Preliminary result of investigations on classification of excitation signals, selection of optimum bases and approximation to of these signals have been presented and briefly discussed.

ACKNOWLEDGEMENTS

The authors would like to thank V.Kolesov for the statistical data.

REFERENCES

[1] Ivanov V., Bogushevich V., Kolesov V., Semenov O., Lanne A. *The Gateway for Internet Telephony*, Proc. The Second European DSP Education & Research Conference, Paris, 1998.

[2] ITU-T Recommendation G.723.1 *Dual Rate Speech coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s*, International Telecommunication Union, 1996.

[3] ITU-T Recommendation G.729 *Coding of Speech at 8 kbit/s Using Conjugate Structure Algebraic Code Excited Linear Prediction (CS-ACELP)*, International Telecommunication Union 1996.

[4] Kondoz A.M. *Digital Speech (Coding for Low Bit Rate Communication Systems),* Chichester, John Wiley and Sons, 1994.

[5] Markel J.D., Gray A.H., Jr. *Linear Prediction of Speech,* New York, Springer-Verlag, 1976.

[6] Rabiner L.R., Schafer R.W. *Digital Processing of Speech Signals*, New Jersey, Prentice-Hall, 1978.

[7] Conway J.N., Sloane N.J. *Sphere Packings, Lattices and Groups*, New York, Springer-Verlag, 1988.

[8] Kleijn W.B., Haagen J. *Waveform interpolation for speech coding and synthesis, in Speech Coding and Synthesis*, (Kleijn W.B. and Paliwal K.K., eds.) pp. 175-208, Elsevier Science Publishers, 1995.

[9] Tichomirov V.M. *Some Problems of Approximation Theory*, Moscow, MSU, 1976 (in Russian).