https://doi.org/10.26552/com.C.2004.4.110-112

Igor Mihalik *

SPEECH COMPRESSION ALGORITHM BASED ON NON-EQUIDISTANT SAMPLING

The method presented here is one of the methods of time domain compression. The technique uses non-equidistant sampling. Human voice and voice conversation are characterized by transitions. During the speech frequency charactericstics change so that viariable sampling can be applied. The nonuniform sampling method which is implemented is described in this paper; its usage in speech synthesis software which is being developed at the Department of Info-Com Networks is shown.

1. Algorithm

There are several approaches to speech compression. One can divide them into 3 basic groups as shown in Fig. 1: *waveform coders* (PCM, ADPCM, ADM, DPCM), *vo-coders* (ITU standardized: G.728, G.729, G.723.1) and *hybrid coders*.

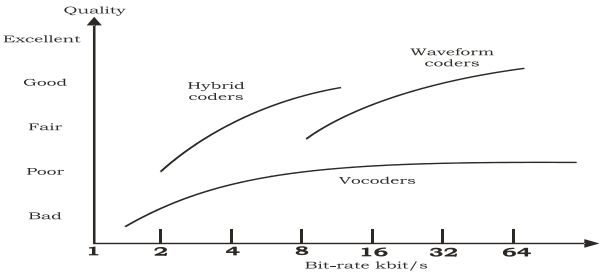


Fig. 1. Code types and quality/bit-rate dependency

The method presented in the article belongs to the group of waveform coders. Nyquist Theorem is the primary theory to keep in mind. Suppose the highest frequency component, in hertz, for a given analogue signal is f_{max}. According to the Nyquist Theorem, the sampling rate must be at least $2f_{max}$; or twice the highest analogue frequency component. The common techniques use constant sampling frequency. PCM [2], for example, uses 8 kHz sampling; this method is used in PSTN. Human speech is characterized by transitions of silence and voice. During the voice period the frequency characteristics change as show in Fig. 2. Some parts of speech contain higher frequencies; so that higher sampling frequency is required, some contain low frequencies; in that case lower sampling rate is required. This observation suggests dividing the speech signal into intervals and applying a different sampling rate to each. To effectively use transmission channels the suitable sampling frequency should be selected. The optimum is to divide the speech signal into intervals and determine the sampling frequency for each using frequency analysis to minimize the amount of transmitted data and keep the quality at a certain level.

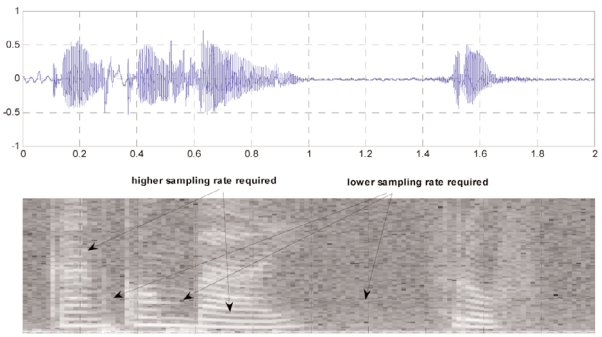


Fig. 2. Human speech: Frequency characteristics change over time. Parts of speech containing higher frequencies need to be sampled using higher sampling rates, parts containing low frequencies need to be sampled using lower sampling rates.

In the next part a method that changes sampling frequency adaptively using a mask is presented. The method works strictly in time domain and a new approach is examined.

2. Mask

The algorithm is based on a mask creation; the Mask *M* is characterized by several properties. The illustration of this mask is shown in Fig. 3. The Mask consists of a set of points, each characterized by its ID and [X, Y] coordinates:

```
typedef T_MASKPOINT struct {
        int ID;
        int X;
        int Y;
};
```

* **Igor Mihalik**
Department of Information Networks, FRI, University of Žilina, Veľký diel, 010 26 Žilina, Slovakia

$$M = \{b_1, b_2, b_3, ..., b_n\} \quad (1)$$

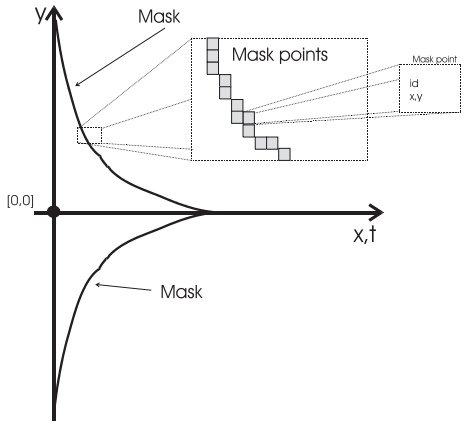


Fig. 3. The Mask consists of a set of points each characterized by its unique ID and coordinates.

In one step one point of the mask is chosen. This point relatively characterizes the signal. The effort is to select such a sequence of points so that by using them the original signal can be accurately reconstructed. The technique is demonstrated in Fig. 4. It shows input signal and its intersection with the mask at points 1, 2 and 3. Each intersection has a point with a specific ID. The sequence of IDs is then transmitted. The decompression part uses the same mask and using the IDs tries to reconstruct the original signal.

3. The Shape of Mask

It is possible to determine the shape of mask experimentally, so that an acceptable compromise between the quality and the compression ratio is achieved. Signal to Noise Ratio is used to measure the quality; for continuous signals it's defined as (2); $f(t)$ stands for an original signal, $f'(t)$ stands for a reconstructed signal. SNR represents the ratio between the energy of the signal and the energy of the noise.

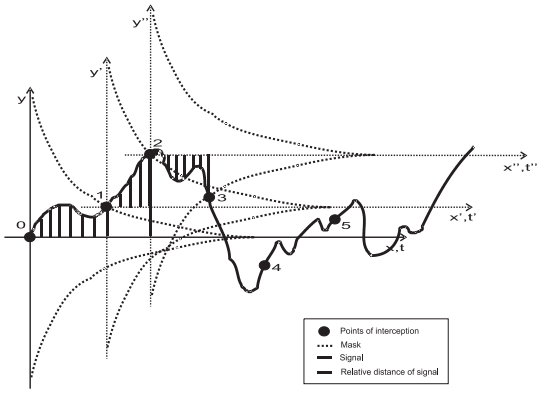
$$SNR = 10 * \log \frac{\int_0^T f^2(t)dt}{\int_0^T (f(t) - f'(t))^2 \, dt} \quad (2)$$

Compression ratio R (2) measures the ratio between the size of an original data and the size of a compressed data.

$$R = \frac{|audio\_data|}{|compressed\_data|} \quad (3)$$

4. The Results

The principle of determining the shape of the mask is based on random generation. For each generated shape (mask instance) the compression ratio and the SNR is measured. The resulting graph is shown in Fig. 5. It shows the dependency between quality and compression rate. The masks with IDs of bit-size 4, 5, 6, 7 and 8 were generated. It was not possible to generate all the possible shapes due to the time complexity, therefore an adaptive algorithm was developed [8]. The number approximates $\approx 256^{256}$ for 8-bit



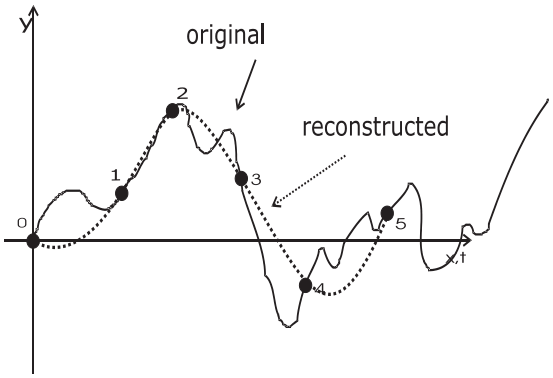


Fig. 4. A compression and a signal reconstruction

There are some elementary mask properties. In the first place, it has to be guaranteed the mask covers the signal so that the signal never moves beyond the mask boundary. The next property is the size of the mask; i.e. the number of points. The more points there are, the more IDs we need; so that we need more bits to encode the IDs for transmission. The shape of the mask is the highest factor influencing the resulting quality and the compression ratio. The method used during the signal reconstruction has a high influence, too.

IDs. For 8-bit IDs we get 256 points. Each point can be anywhere within the coordinates. The case with more than one point placed on the same coordinate eliminates the final number of possibilities. But it does not reduce the number as radically as to be able to examine all of them. For the generation exactly two points to be generated having the same X coordinate (above the X axis, below the X axis) are supposed. The mask has to be continuous and at least two points are generated with X coordinate equal to zero.

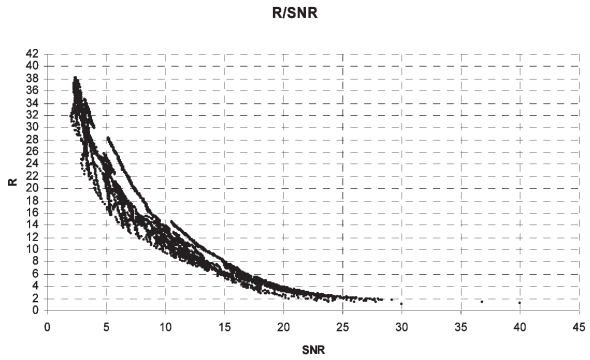


Fig. 5. Each point in graph represents one instance of mask. Each instance is characterized by achieved SNR and compression ratio (R). R/SNR dependency diagram gives an overview of possibilities of presented approach

5. Speech synthesis (TTS)

One of the methods of speech synthesis is based on usage of speech units called diphones. Using the diphones the resulting speech is created, units are concatenated together. This method is known as the concatenate method. The text-to-speech system developed at the Department of Info-Com networks uses this concatenate technique. The memory requirements are the problem bearing in mind the number of used diphones. There are about 2000 diphones used and the size of each is 10 Kbytes on average. It is necessary to minimise the memory requirements for the mobile devices such as PDAs, cellular phones and others. The compression contributes to this goal noticeably. Fig. 6 shows the structure of the diphone used in the TTS system. The diphones

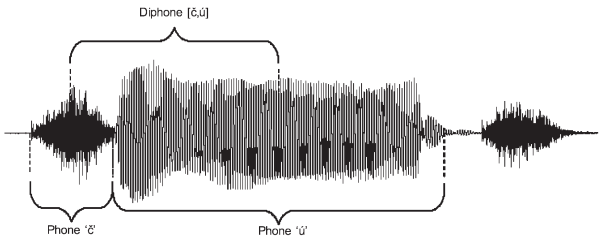


Fig. 6. A diphone structure

are stored compressed; for each experimentally determined the suitable mask shape to meet the requirements for quality (SNR > 25dB). Using compression the total amount of data was reduced to 40% of its original size.

6. Conclusion

The presented technique gives acceptable results and reduces the amount of data needed for diphone storage. More experiments with the shape of the mask and the reconstruction method are required to cover a larger area of possibilities. During the experiments linear, cubic and spline approximation was used. Other methods of interpolation would also give a more complete view of the technique although noticeable improvements are not expected. The results show that presented algorithm compared to other techniques (i.e. ADPCM, that is loss-less and reduces data up to 25% of original size) does not excel and achieved compression ratio and SNR are not satisfactory. However there are not any known sources or publications that would use presented approach so this paper may contribute to future research in the area of non-equidistant sampling rate.

References

[1] MINOLI, D., MINOLI, E.: *Delivering Voice over IP networks,* ISBN 0-471-25482-7, (1998)
[2] PULSE CODE MODULATION (PCM) OF VOICE FREQUENCIES *ITU-T Recommendation G.711*
[3] 40, 32, 24, 16 kbit/s ADAPTIVE DIFFERENTIAL PULSE CODE MODULATION (ADPCM) *ITU-T Recommendation G.726*
[4] 7 kHz AUDIO - CODING WITHIN 64 KBIT/S *ITU-T Recommendation G.722*
[5] Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3kBit/s ITU-T Recommendation G.723.1
[6] Engineering Fundamentals, Sampling Theorem and Nyquist Rate, *http://www.efunda.com/designstandards/sensors/methods/DSP_nyquist.cfm*
[7] Black, A. W., Taylor, P., Caley, R.: *"Festival Speech Synthesis System",* http://www.cstr.ed.ac.uk/projects/festival/festival-toc.html
[8] Igor Mihálik: *Návrh parametrov kompresného algoritmu,* DP reg.číslo 275/2000, (2001).