Transient Detection for Speech Coding Applications

Grzegorz Szwoch, Maciej Kulesza, and Andrzej Czyżewski,

Gdansk University of Technology, Multimedia Systems Department, Gdansk, Poland

Summary
Signal quality in speech codecs may be improved by selecting transients from speech signal and encoding them using a suitable method. This paper presents an algorithm for transient detection in speech signal. This algorithm operates in several frequency bands. Transient detection functions are calculated from energy measured in short frames of the signal. The final selection of transient frames is based on results of detection in all frequency bands. Performance of the algorithm is evaluated and some enhancements are proposed. The algorithm described here allows for accurate transient detection in speech and is suitable for use in practical speech coding applications.

Key words:
speech coding, transient detection, VoIP applications, signal processing

1. Introduction

For many years speech coding applications utilized encoding algorithms that concentrated on ensuring low bit rate and low encoding delay, and provided 'intelligible' speech, without any requirements to subjective signal quality. However, recent development of telecommunication networks and increase in processing power of computer hardware decreased significance of low delay and low bit rate criteria. On the other hand, for end users, quality and naturalness of the decoded speech is important and influences 'quality of service' assessment. Therefore, a new trend in speech coding focuses on improving subjective signal quality of decoded speech, maintaining bit rate and processing delay at a reasonable level. An example of this is AMR-WB+ codec which improves speech quality by introducing wide-band signal encoding and other enhancements into traditional speech encoding scheme [1]. However, there are still some unexplored methods of enhancing speech quality that might be incorporated into speech coding architecture. One of these methods is efficient detection and encoding of speech transients, which is the main subject of this paper.

The authors aim to develop a novel speech coding architecture that would fulfill the requirements of modern Voice-over-IP (VoIP) applications, with special emphasis placed on improving speech quality, assessed by the end-user, listening to the decoded speech [2]. The proposed codec architecture diverts from the traditional approach to speech encoding, based on linear prediction. Instead, authors decided to utilize various signal encoding techniques (waveform, transform, perceptual and possibly also parametric coding) for encoding of speech components having different structure, e.g. voiced, unvoiced and transient parts of the signal [3].

Using a dedicated algorithm for encoding of speech transients seems to be potentially useful method of improving signal quality. None of practically used speech codecs select transient parts from the signal and encode them using a dedicated method. Transients in speech signal are characterized by rapid changes in signal energy and/or in spectral distribution of the energy. They occur at the beginning of speech segments and also when the character of signal components change, for example at the boundaries of voiced/unvoiced speech parts. Due to non-stationary character of transients, speech codecs based on linear prediction (CELP and its derivatives) are not capable of encoding transients accurately, which results in deterioration of signal quality. Therefore, using a dedicated algorithm for transient encoding is expected to enhance quality of the speech. However, this approach also needs an algorithm for selection of transients in speech that is both efficient and accurate. The experiments presented in this paper concentrate on designing transient detection algorithm for speech codec and assessing its performance.

2. Problems of transient detection in speech coding applications

Various algorithms for detection of transient states may be found in literature [4]. However, selection of algorithm suitable for speech codec used in VoIP systems is not straightforward. There are several issues related to this problem and they are discussed below.

1. The algorithm should have low complexity. The complete encoding scheme is very complex and transient detection is only a small part of the signal processing chain. Therefore, the detection algorithm should take as little system resources as possible and it should not introduce high delays into the processing scheme.

Manuscript received December 5, 2006.
Manuscript revised December 25, 2006.

2. High accuracy of transient detection in speech signal is required. Too many false positive decision will result in inappropriately high bit rate, as transient encoding algorithms usually produce high bit rate streams. Too many false negative decisions cause deterioration of signal quality, because transient are not encoded efficiently.
3. Speech codec receives and processes small packets (frames) of signal samples. Many transient detection algorithms operate in off-line mode and they require access to the complete signal [5], which is not possible in case of speech codec, so these methods cannot be used here. Therefore, the detection algorithm should be able to perform decisions using only information obtained from the current and previous signal frames.
4. Most of the transient detection methods found in literature are dedicated to musical signals [4], some of them are even named 'musical onset detection algorithms' [6]. These methods are not equally accurate for speech signals due to differences in nature of transients in musical and speech signals. In experiments performed by the authors, many algorithms that were accurate for music, failed to detect transients in speech. Therefore, the detection algorithm has to be adopted for speech signal characteristics.

To conclude, it is not recommended to simply take a detection algorithm that was proved to be accurate for musical signals and implement it in speech codec, because this will not yield the desired results.

The transient detection algorithm described in the following sections of the paper, was tested by the authors and selected for use in the proposed speech codec. The algorithm proved to be an effective and accurate method for transient detection in speech signals.

3. Speech transient detector

3.1 General design of the algorithm

Typical transient detection methods are based on evaluation of changes in signal energy and also on detection of changes in spectral distribution of speech signal. In speech, some transients are characterized only by relocation of the spectral energy into another frequency range, while the overall energy does not change significantly. This kind of transients will be referred to as spectral transients [7].

The general block diagram of the algorithm is presented in Fig 1. Signal samples are processed in frames. Packets of incoming samples are accumulated in the buffer and divided into frames 256 samples long, with 50% overlap. Each frame is multiplied by hanning window and FFT is computed.

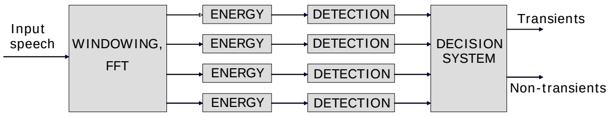

Fig. 1 Block diagram of the proposed transient detection algorithm, operating in four frequency bands.

In order to detect both normal and spectral transients, the detection should not be performed on the whole frequency range of the signal. Many detection algorithms divide the signal into several frequency bands, perform transient detection in each band separately, then combine detection values calculated for each band to make the final decision [5,6]. This approach is also used in the algorithm described here, but it is implemented in a different way. Many authors propose using filterbanks for successive division of the signal into half-bands. This approach has an advantage that analysis resolution is matched to signal characteristics in different frequency ranges [5]. However, these methods require access to the complete signal. Speech codec operates on small frames of input samples and does not know about future signal frames. Therefore, in the proposed algorithm, instead of filtering the signal using a filter bank, spectrum of the signal frame is computed using FFT and the resulting spectrum is divided into several bands.

In the early experiments, the signal with 44.1 kHz sampling rate was divided into 4 half-bands, simulating the filterbank structure used in some of the onset detection algorithms. However, it was found that information obtained for frequencies above 5.5 kHz is not useful for transient detection in speech. Therefore, detection was limited to range 0 – 5.5 kHz, divided into four frequency bands: 0 – 1.4 kHz, 1.4 – 2.8 kHz, 2.8 – 4.1 kHz and 4.1 – 5.5 kHz. This structure results in 8 FFT values for each band. The resolution of the analysis proved to be sufficient for transient detection.

For each frequency band, energy is calculated. Using the obtained energy values, detection functions are constructed and values of these functions are compared with local threshold values, for each band and for each frame. If value of detection function in a given band is below the threshold, this value is set to zero. Next, global detection value is computed as a weighted sum of local detection values, calculated for each band. Finally, this function is compared to the global threshold value and if this threshold is exceeded, the frame is selected as a transient.

3.2 Algorithm based on energy comparison

The algorithm presented here aims to detect rapid changes in signal energy, typically associated with transient states. As described in the previous section of the paper, detection is performed in frequency bands, so it is possible to detect increase in total signal energy, as well as changes in spectral distribution of energy (if energy raises in a frequency band).

In this algorithm, very simple detection function was used. This function is calculated as a difference of energy of the current and previous frame. For a given frame *n*, in each frequency band *m*, energy of the signal frame $E_{n,m}$ is calculated as:

$$E_{n,m} = \sum_{k=kfirst}^{klast} \left| X_k \right|^2 , \qquad (1)$$

where X_k is the *k*-th component of FFT of the windowed frame containing signal samples, *kfirst* and *klast* are the first and the last FFT component belonging to the frequency band *m*, respectively.
The detection function for each band is computed as:

$$F_{n,m} = E_{n,m} - E_{n-1,m} \ . \qquad (2)$$

Values of detection function in each band are compared to local thresholds T_m and the global detection function GF_n is calculated:

$$GF_n = \sum_{m=1}^{M} w_m F'_{n,m} \ , \qquad (3)$$

where w_m is a weighting factor, *M* is the number of frequency bands and $F'_{n,m}$ is defined as:

$$F'_{n,m} = \begin{cases} F_{n,m} & \text{if } F_{n,m} \geq T_m \\ 0 & \text{if } F_{n,m} < T_m \end{cases} , \qquad (4)$$

assuming that T_m is always a positive value.
Finally, if value of global detection function is greater than or equal to global threshold value, the whole signal frame is marked as a transient.

This algorithm is very simple and one cannot expect that it will provide accurate detection of speech transient. However, it was implemented and examined in order to find out what modifications are needed for improvement of performance of this algorithm. The example of detection results obtained using this method is shown in Fig. 2. It is evident that this simple detection function makes it very hard to set threshold values that would result in accurate transient detection. Due to high frequency of changes of signal energy, the detection function also changes very rapidly. Therefore, the main conclusion of this experiment is that the detection function needs further processing.

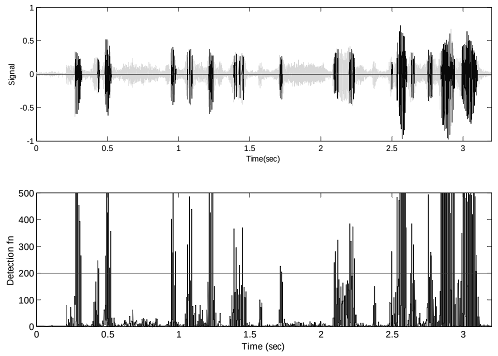

Fig. 2 Results of transient detection in speech signal using basic algorithm, calculating difference of energy between two adjacent frames. Upper plot: speech signal separated into transient (black) and non-transient (gray) frames. Lower plot: transient detection function and threshold (flat line)

3.3 Smoothing of detection function

The plot of detection function obtained using the basic form of the detection algorithm, shows that this function is very noisy and changes rapidly. Therefore, the first modification introduced to the algorithm is smoothing of the function. This may be achieved simply by applying a low order FIR filter. It is possible to filter the detection function calculated using Eq. 2. However, it is more computationally efficient to use filtering operation during construction of the detection function, so that Eq. 2 is changed to:

$$F_{n,m} = E_{n,m} - \sum_{p=1}^{P} \left(E_{(n-p),m} / p \right) . \qquad (5)$$

The detection function now depends on energy of the current frame and *P* previous frames. Due to introduction of weighting factor (1/*p*), more recent frames have greater influence on the value of detection function, but older frames also have some effect on the detection. This operation averages the measured energy in the frames, but it may also be treated as low pass filtering of the detection function, resulting in much more smooth shape of this function.

The plot of the final detection function averaged as in Eq. 5, with *P* = 10, is shown in Fig. 3. Tenth order FIR filter is sufficient to obtain smoothed detection function. It can be seen that this simple modification of the original algorithm significantly increased efficiency and accuracy of transient detection. It is now much easier to set threshold values that result in proper transient detection.

3.4 Selection of transient spectral components

The algorithm presented in previous sections of the paper is able to detect changes in total energy and in spectral distribution of signal energy. Some algorithms used for musical signals use also phase dependencies in order to divide the signal into steady state and transient parts. The idea is based on computation of instantaneous frequency. For each FFT component, phase difference between the current and previous frame is calculated. It is assumed that changes of instantaneous frequency indicate the transient state [6].

In order to assess whether inclusion of phase dependencies in the detection algorithm improves accuracy of the detection, the algorithm was modified as follows. After FFT of the windowed signal frame is computed, phase of each FFT component is predicted using phase information obtained for the previous two frames. In other words, for a given spectral component, phase increment between each two adjacent frames should be identical. If this is not the case, a high value of prediction error will be observed. Prediction error is calculated as

$$\varepsilon = \phi_{n,k} - 2\phi_{(n-1),k} + \phi_{(n-2),k} \ , \qquad (6)$$

where k is index of FFT component calculated for n-th frame, $\phi_{n,k}$ is phase of this component. Therefore, FFT components with value of prediction error ε lower than a set threshold, are considered to be non-transient and they are discarded from further analysis. Energy of the signal frame in each frequency band is now computed using Eq. 1, taking into account only transient FFT components that remain after phase analysis. Further operations are identical to the original algorithm.

The results of transient detection obtained using the modified algorithm are shown in Fig. 4. Comparing Figs. 3 and 4 it can be seen that the shapes of detection functions are different. Values of detection function for the modified algorithm are higher than in the original one for some of the detected transients, while for other transient frames the opposite effect is observed. However, maxima of both detection functions generally occur in the same frames and by appropriate setting of the threshold values, similar accuracy is achieved using both algorithms.

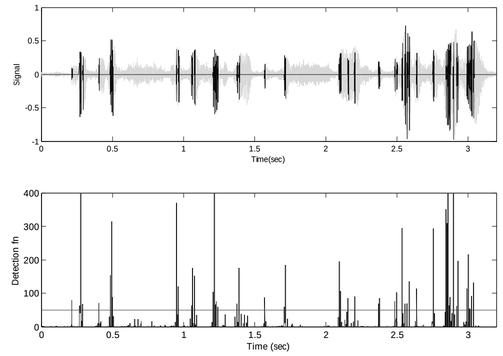

Fig. 3 Results of transient detection in speech signal using modified algorithm, smoothing detection function with 10th order FIR filter. Upper plot: speech signal separated into transient (black) and non-transient (gray) frames. Lower plot: transient detection function and threshold (flat line)

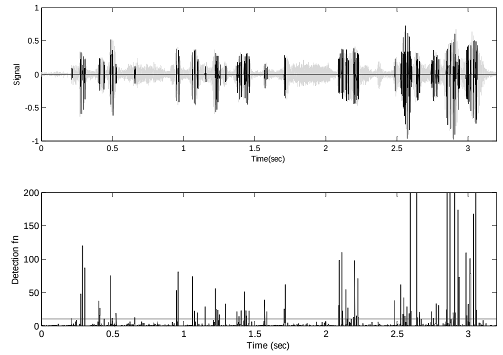

Fig. 4 Results of transient detection in speech signal using algorithm with additional selection of transient spectral components. Upper plot: speech signal separated into transient (black) and non-transient (gray) frames. Lower plot: transient detection function and threshold (flat line)

4. Discussion

The results of the experiments proved that transient detection algorithm based on simple comparison of energy calculated for signal frames, with detection performed in several frequency bands and with simple smoothing of the obtained detection functions, is able to fulfil the requirements stated in Section 2. The algorithm performs transient detection in speech signal, operating on incoming packets of signal samples, with satisfactory accuracy.

The complexity of the proposed algorithm (without selecting transient components using phase information) is not high. Windowing of the signal and FFT computation are typical signal processing operations and they are performed in speech codec even if transient detection is not performed. Calculation of frame energy, detection functions computation (with smoothing using FIR filter of low order) and thresholding also do not require excessive processing power. Therefore, implementation of this transient detection algorithm in the developed speech codec does not increase complexity of this codec in a significant way.

The accuracy of this detection method is satisfactory. It is important to stress that the developed speech codec may impose some limit on number of frames that are encoded using the algorithm dedicated for transients. If this is the case, it is important that the algorithm detects all the most evident transients. If some weaker transients are missed, it should not deteriorate the final signal quality. The algorithm presented here performs the task of selecting the strongest transients with accuracy that is sufficient for speech codec application.

Assessment of the modified transient detection algorithm that selects spectral components basing on phase dependencies, is problematic. This algorithm did not improve the overall accuracy of transient detection as the authors expected. At the same time, computational complexity of the algorithm is increased. More experiments need to be carried out in order to assess whether the expanded model provides sufficient increase of detection accuracy that justifies greater complexity of the algorithm. However, basing on the experiments performed so far, it seems that the algorithm without additional selection of transient components is an optimal choice for speech codec, as detection accuracy and computational complexity are properly balanced.

The main problem that remains to be solved in the next experiments is development of a method for optimal setting of the threshold values for detection. Constant threshold values do not provide equal detection accuracy for speech signals having different character (e.g. differing in pitch or speech tempo). A better solution is an algorithm that adapts the thresholds to the signal. This modification may increase the complexity of the algorithm, but it is expected that adaptive thresholding will yield more accurate results of transient detection.

5. Conclusions

The algorithm for transient detection in speech signals, intended for use in speech coding applications, was proposed and examined. The algorithm fulfils the main requirements: (a) sufficient detection accuracy, (b) low complexity, (c) processing of the signal incoming in packets. The accuracy of transient detection in speech signal is achieved by: (1) limiting the frequency range for detection to 0 – 5.5 kHz range, where most of the speech signal energy is cumulated; (2) detection based on comparison of signal energy in successive frames, (3) detection performed in several frequency bands, allowing detection of both normal and spectral transients; (4) computing the detection function as a weighted average of energy of several recent frames, which also results in smoothed detection function; (5) making the final 'transient/non-transient' decision by combining the decisions made in each frequency band. Further improvement of accuracy of the algorithm may be obtained by adapting the threshold values to character of the processed signal. Using the phase information may be beneficial for transient detection, although this requires further studies. The algorithm presented here is currently being implemented in the speech codec and the results of transient detection are used by other processing blocks of the codec. An optimal encoding algorithm will be used for parts of the signal selected as transients, resulting in enhanced subjective quality of the speech encoded using the proposed codec.

Acknowledgments

Research funded by the Polish Ministry of Education and Science within the Grant No. 3 T11D 004 28.

References

[1] P. Ojala, A. Lakaniemi, H. Lephanaho, M. Joakimies, "The Adaptive Multirate Wideband Speech Codec: System Characteristics, Quality Advances, and Deployment Strategies", IEEE Communication Magazine, vol. 44, no. 5, pp. 59-65, May 2006.

[2] M. Kulesza, G. Szwoch, A. Czyżewski, "A Hybrid Speech Codec Employing Parametric and Perceptual Coding Techniques", 121st AES Convention, preprint 6956, San Francisco 2006.

[3] S.N. Levine, J.O. Smith, "A Sines+Transients+Noise Audio Representation for Data Compression and Time/Pitch Scale Modifications", 105th AES Convention, preprint 4781, San Francisco 1998.

[4] L. Daudet, "A Review on Techniques for the Extraction of Transients in Musical Signals", Proc. CMMR'05, Pisa 2005.

[5] C. Duxbury, M. Davies, M. Sandler, "Separation of Transient Information in Musical Audio Using Multiresolution Analysis Techniques", Proc. COST G-6 Conference on Digital Audio Effects (DAFX-01), Limerick 2001.

[6] C. Duxbury, M. Sandler, M. Davies, "A Hybrid Approach to Musical Note Onset Detection", Proc. 5th Int. Conference on Digital Audio Effects (DAFX-02), Hamburg 2002.

[7] V.S. Babu, A.K. Malot, V.M. Vijayachandran, M.K. Vinay, "Transient Detection for Transform Domain Coders", 116th AES Conv., preprint 6175, Berlin 2004.

Grzegorz Szwoch received his M.Sc. degree in Sound Engineering from Gdansk University of Technology in 1996. Since then, he has been working in Multimedia Systems Department in the same university. Initially, his main topic of research was waveguide modeling of hearing aid and he received his Ph.D. degree in 2004. Currently, he is involved in several research projects, including development of new speech codec architecture and other sound processing applications.

Maciej Kulesza received his M.Sc. degree in Sound Engineering from Technical University of Gdansk in 2003. He is interested in digital audio processing techniques and DSP based electronics systems engineering. Now he is a Ph.D student at the Multimedia Systems Department. His thesis is devoted to developing of effective combined parametric-perceptual method of speech coding for VoIP applications.

Andrzej Czyzewski received his M.Sc. degree in Sound Engineering from the Gdansk University of Technology in 1982, his Ph.D. degree in 1987 and his D.Sc. degree in 1992 from the Cracov Academy of Mining and Metallurgy. Prof. Czyzewski joined the staff of the Sound Engineering Department of the Gdansk University of Technology in 1984. In December 1999 Mr. President of Poland granted him the title of Professor. In 2002 the Senate of his University approved him to the position of Full Professor.