NEW APPROACH TO VOICED ONSET DETECTION IN SPEECH SIGNAL AND ITS APPLICATION FOR FRAME ERROR CONCEALMENT

Catherine Lemyre, Milan Jelinek and Roch Lefebvre

Université de Sherbrooke, Sherbrooke (Québec) Canada
catherine.lemyre@usherbrooke.ca

ABSTRACT

This paper presents a new approach for accurate detection of voiced onsets in a speech signal. The knowledge of the precise location of the beginning of a voiced segment has several potential applications in speech compression (e.g. source-controlled variable bitrate coding, classification-based frame erasure concealment). The proposed onset detection algorithm is based on Teager's energy operator applied on subbands. The technique has been integrated in the signal classifier of the VMR-WB speech coding standard used for frame erasure (FE) concealment. A subjective listening test using MUSHRA methodology showed an improvement in frame error conditions with respect to the legacy VMR-WB classification.

Index Terms— Onset detection, frame erasure concealment, Teager's energy operator, signal classification, VMR-WB

1. INTRODUCTION

Signal classification, and onset detection in particular, is important in many speech processing applications. In speech coding, the source-controlled variable bitrate (VBR) concept has been used for many years and is a part of several 3GGP2 standards (SMV [1], VMR-WB [2]). The basic idea of VBR coding is that more bits are necessary to describe transient signals compared to stationary signals. As voiced onset frames are probably the most critical, they require the highest available bit rate, and accurate onset detection is thus important for high quality VBR coding.

Another application of accurate onset detection is in frame error (FE) concealment based on signal classification. Classification-based FE concealment has been successfully applied e.g. in the VMR-WB speech coding standard. The general rule of a classification-based FE concealment is that the signal energy and the spectral envelope should converge rapidly to background noise parameters if erasure happens in a non-stationary region, but should be maintained practically unchanged if erasure happens in stationary parts of the speech signal. It means that while classification-based FE concealment helps in general, a wrong classification might produce annoying artefacts. A wrong detection of an onset frame is probably the most critical case.

FER concealment techniques have advanced considerably in recent years and the synthesized speech quality is generally very good for erasures occurring in a quasi-stationary speech segment. This was reflected e.g. in the terms of reference in recent ITU standardizations where tested FER conditions were compared to clean-channel references. However, FE occurring during transition frames can still cause significant distortion. While voiced onsets erasures are not frequent (using our detector, approximately 6% of active speech frames are voiced onset frames), their impact is important and can affect the overall telecommunication quality perceived by users.

In this paper, we propose a new technique that allows for onset detection with a subframe resolution. The proposed detection uses Teager's energy operator [3] applied on a band-limited speech signal. The method has been implemented in VMR-WB and its performance in FE conditions is compared with the original VMR-WB classification.

The paper is organized as follows. Section 2 introduces Teager's energy operator. Section 3 presents the application of Teager's energy operator to onset detection. Section 4 demonstrates how the proposed onset detection can be applied to VMR-WB to improve classification for FE concealment. Finally, section 5 presents results and gives a comparison with the standard VMR-WB classification.

2. TEAGER'S ENERGY OPERATOR

In speech processing, signal energy is usually measured as the square of the amplitudes. Kaiser [3] proposes another way to quantify the signal energy. This definition is based on the energy needed to generate the signal. With the conventional energy definition, signals with different frequency but with same amplitude have equivalent energy. With Kaiser's approach, less energy is required to produce lower frequency signals than higher frequency signals if these two signals have the same amplitude. This definition is inspired by the motion of a mass suspended on a spring. Kaiser demonstrates that the energy of this simple oscillation system is proportionate

This project is financed by the Canadian NSERC and VoiceAge Corporation.

both to the square of the amplitude and to the square of the frequency of the oscillation:

$$E \propto A^2\omega^2 \quad (1)$$

In [3] Kaiser derives a discrete energy operator that approximately satisfies (1) for a discrete harmonic system, provided the system oscillation frequency is significantly lower than the sampling rate. This operator is called the discrete Teager energy operator $\Phi[x(n)]$ and can be easily calculated with equation (2).

$$\Phi[x(n)] \quad = \quad x^2(n) - x(n-1)x(n+1) \approx A^2\Omega^2 \quad (2)$$

From (2) it can be seen that Teager's energy operator gives a good approximation of the energy variation provided the oscillation frequency is constant or slowly varying. In case two harmonic components are present in the signal, Kaiser shows [3] that in addition to the sum of the energies of the two components, the result contains also a cross-term. For this reason, this operator is generally applied to a band-limited signal.

Compared to e.g. an energy detection using a sliding window, the Teager's energy operator has two interesting properties: it is independent of the initial phase of an energy event, and is capable of responding very rapidly to changes in the amplitude and the frequency of a signal. If we consider a beginning of a harmonic signal as theoretical example, the Teager's energy operator will reach in two sampling instants and remain set during its duration. Since a voiced speech onset can be considered to some extent as a beginning of a harmonic oscillation, these properties make Teager's operator very interesting for onset detection.

3. VOICED ONSET DETECTION USING TEAGER'S ENERGY OPERATOR

Based on the characteristics described in section 2, our goal is to use Teager's energy operator to locate the beginning of a voiced onset. In our approach the speech signal is separated in subbands covering only the typical range of the fundamental frequency of speech and Teager's energy operator is applied to each band. The bandwidth of the frequency bands is chosen sufficiently narrow (typically 50 Hz) to avoid multiple pitch harmonics in the same band. This reduces false alerts due to energy events other than beginnings of voiced onsets, and reduces Teager's operator cross-term effect due to the superposition of several signal components.

To determine the beginning of an onset, the output of Teager's energy operator is compared to a threshold T in each subband. As Teager's operator is based on absolute energy, it is however sensitive to the input signal level. To overcome this sensitivity, the threshold T is made adaptive with the long term active speech energy average E_{LT}:

$$T = 0.95 * E_{LT} \quad (3)$$

The value 0.95 has been found experimentally.

To be able to use the same threshold for each subband, a normalization is further applied to the result of Teager's operator. This normalization factor is equal to the central frequency of each subband (ω_c^2).

An example of Teager's energy operator is shown in Fig. 1 for few subbands with the corresponding speech signal at the top. Sixteen frequency bands cover the pitch range of the speech signal (0 to 800 Hz).

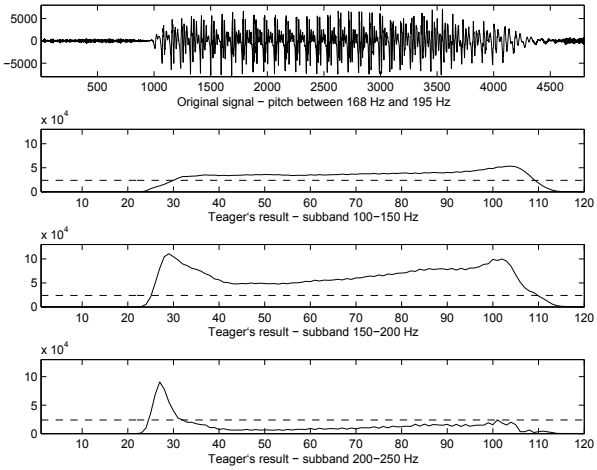

Fig. 1. Teager's energy holds over threshold during more than 1 frame

As mentioned previously, one of the interesting properties of Teager's energy operator is its rapidity in reacting to the changes in the amplitude and the frequency of the signal. This gives it an advantage compared to classical signal classification techniques usually based on a correlation analysis and an estimation of the ratio between high and low frequency components. Both these approaches need a significant time interval to provide a robust classification result.

4. APPLICATION TO VMR-WB FE CONCEALMENT

We have implemented the proposed onset detection within the signal classification used for FE concealment in the VMR-WB speech coding standard. That classification uses a merit function given by a weighted average of several normalized parameters (normalized correlation, spectral tilt, SNR in the weighted speech domain, pitch stability counter, relative frame energy and zero-crossing counter)[2].The VMR-WB concealment depends on the signal classification and distinguishes the following frame classes: Voiced, Unvoiced, Voiced Transition, Unvoiced Transition and Onset frames. The concealment following transitions can be summarized as a rapid attenuation towards the characteristics of the background noise. Erasures following Unvoiced frames are concealed by filtering a random excitation signal, with energy similar to the en-

ergy of the past frame excitation, through a linear prediction (LP) synthesis filter of the previous frame. Onset class comprises voiced frames with stable characteristics at the end of the frame (i.e. at least one well-built complete pitch period can be found at the frame end) following Unvoiced or Unvoiced transition frame. From the concealment perspective, erasures following Voiced or Onset frames are processed in the same manner as stationary voiced sounds. Basically, the last pitch period of the previous frame excitation is repeated with some attenuation, and the spectral envelope is practically kept unchanged.

Related to the Onset frame detection, there are the following three critical concealment situations. First situation arises when the Onset frame is detected too soon and Unvoiced frame is classified as Onset. If the following frame is lost, the excitation is repeated with a meaningless pitch period. Second situation arises when an onset is indeed present in a frame classified as Onset, but is not yet built well enough at the end of the frame to be periodically repeated in case the following frame is lost. An incomplete pitch waveform is then repeated with often a wrong pitch period. Third, if the Onset frame is detected too late, the true Onset frame is classified as Unvoiced or as Unvoiced Transition. The concealment will incorrectly move the signal parameters rapidly to the parameters of the background noise causing a perceptually annoying energy event.

In our implementation, the proposed Teager onset detection has priority over the original VMR-WB onset detection. It means that if the new Teager detection finds an Onset frame, the classification is forced to Onset. If it decides to delay the Onset decision to the next frame, Unvoiced or Unvoiced Transition frame is forced in the current frame. The rest of the VMR-WB FE classification is kept intact. The decision between Unvoiced and Unvoiced Transition is important as the VMR-WB concealment will generate a random excitation practically without attenuation in the first case, and will rapidly attenuate the signal in the second case. The decision depends on the position of the onset detected by the Teager's operator and the estimated pitch period given by the VMR-WB open-loop pitch estimation [2].

VMR-WB uses a 12.8 kHz internal sampling rate and 20 ms frames (256 samples). For the purpose of the proposed onset detection, the input signal is decomposed in bands of 50 Hz each within the range of the speech fundamental frequency. In our implementation the frequencies covered vary from 50 Hz to 800 Hz. Teager's energy operator is then applied to each of these subbands using equation (2), normalized with the square of the subband central frequency, and averaged over blocks of 32 samples. It results in 8 values of Teager's energy operator per band per frame. To detect an onset, the output of Teager's operator must be over the threshold of Eq. 3 in at least one subband and the previous frame needs to be classified as Unvoiced or Unvoiced-Transition. This is because Teager's energy operator will stay over the threshold as long as a frequency component is present in the band.

In the VMR-WB FE concealment framework, the proposed detection does not guarantee that at least one pitch period is well-built at the end of the frame, and the concealment is based on the assumption that this is needed for the voiced concealment to work properly in case the following frame is lost. If less than one pitch period is present, it is better to attenuate the signal since a periodic extension of the pitch period fraction would likely cause a more significant artifact. In particular, the pitch length must be lower than the number of samples corresponding to the difference between the end of the subframe where Teager's energy operator detected an onset and the end of the frame to declare an Onset frame. If pitch length is greater than Teager's energy operator position, the frame is declared Unvoiced (if Teager's position is near the end of the frame) or Unvoiced Transition (if Teager's position is at the beginning of the frame).

5. RESULTS

The goal of implementing our method in the VMR-WB framework was to see if we can correct situations where the VMR-WB FE classification fails to correctly detect Onset frames and whether correcting those cases has a perceptual impact in FE error conditions. An example of frame erasure following the Onset frame detected too early is illustrated in the second plot of Fig. 2. The third plot shows a result when the proposed onset detection is used instead. In this case, the third frame is correctly classified as Onset frame.

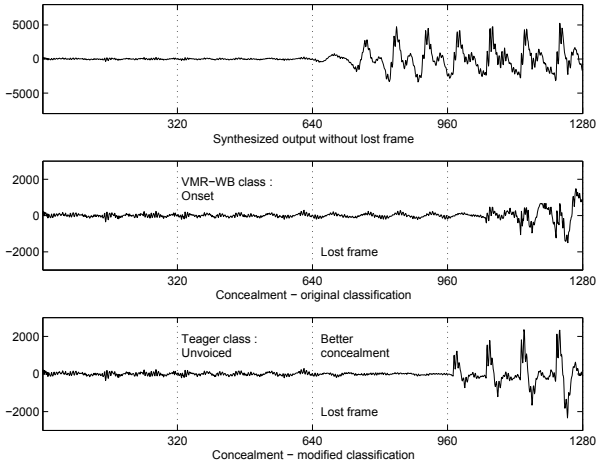

Fig. 2. Example of the Onset frame classified too early, and its correction with the presented method

The second example shown in the second plot of Fig. 3 illustrates the case when VMR-WB FE classification detects the Onset frame too late. The third plot shows again the concealment in case the proposed classification is used instead.

To asses the perceptual impact of the modification, we

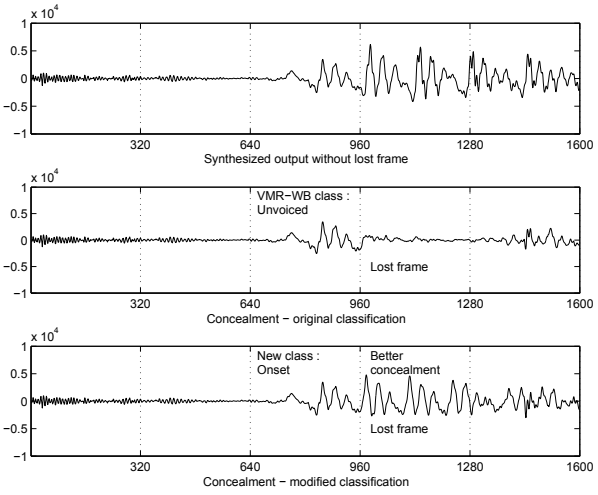

Fig. 3. Example of Onset frame classified too late, and its correction with the presented method

have designed two Mushra tests [4] evaluating scenarios corresponding to Fig. 2 and 3. First we have manually labelled Onset frames, then we have selected for erasures only frames where the new onset detection differs from the VMR-WB classification. In the first test, we have erased the manually labelled Onset frames. In the second test, we have erased the frames following the manually labelled Onset frames. Our goal was to show that while our modification affects only a small fraction of the frames, a frame error hitting one of those frames has a significant impact on the perceived quality (only one or two frames were generally erased per listening sentence in this test setting.)

Fig. 4 gives the results when the Onset frame is selected too soon by at least one of the approaches, and the true Onset frame is missing. Fig. 5 shows the results when the Onset frame has been detected too late and the frame following the true Onset frame is missing. It can be seen that the second case has generally greater impact on the perceived quality and consequently the improvement brought by more accurate onset detection is also higher. (Note these two frame erasure scenarios are the only scenarios where the concealment is significantly different for both approaches).

The test set-up consisted of 45 sentences grouped into 3 blocks and distributed among nine listeners. In total, 180 votes were collected per experiment.

6. CONCLUSION

We have introduced a new method for voiced onset detection. By using Teager's energy operator applied to the speech signal split in frequency bands, it is possible to achieve a more accurate detection of onsets. The method based on Teager's operator has been implemented in the VMR-WB speech coding standard to enhance the classification used for frame error concealment. Cases were illustrated where the new method corrects some of the VM-WB onset detection errors. Subjective tests showed a perceptual improvement for the new technology in case of frame erasures affecting frames following voiced onsets.

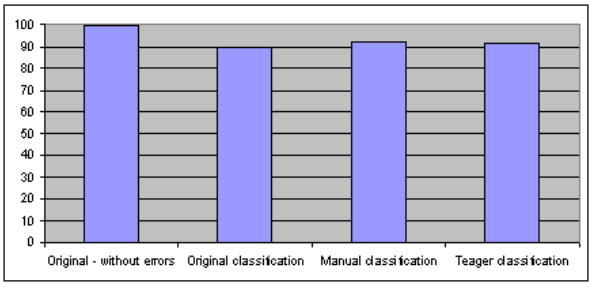

Fig. 4. Mushra results if Onset frame is selected too soon

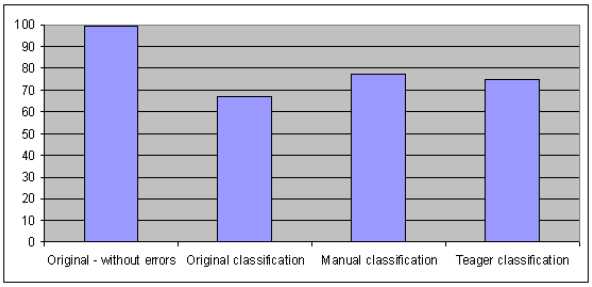

Fig. 5. Mushra results if Onset frame is selected too late

7. REFERENCES

[1] Yang Gao, Eyal Shlomot, Adil Benyassine, Jes Thyssen, Huan yu Su, and Carlo Murgia, "The SMV algorithm selected by TIA and 3GGP2 for CDMA applications," in *ICASSP'2000*. IEEE, 2000, vol. 2, pp. 709–712.

[2] M. Jelinek and R. Salami, "Wideband speech coding advances in VMR-WB standard," *IEEE Transactions on Audio, Speech and Language Processing*, pp. 1167–1179, May 2007.

[3] James F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," in *ICASSP'90*. IEEE, 1990, vol. I, pp. 381–384.

[4] Recommendation ITU-R BS.1534, "Method for the subjective assessment of intermediate quality level of coding systems," .