Low-Complexity and Efficient Classification of Voiced/Unvoiced/Silence for Noisy Environments

Tuan Van Pham and Gernot Kubin

Signal Processing and Speech Communication Laboratory
Graz University of Technology, Austria
v.t.pham@tugraz.at, g.kubin@ieee.org

Abstract

This paper describes a low-complexity and efficient speech classifier for noisy environments. The proposed algorithm utilizes the advantage of time-scale analysis of the Wavelet decomposition to classify speech frames into voiced, unvoiced and silence classes. The classifier uses only one single multidimensional feature which is extracted from the Teager energy operator of the wavelet coefficients. The feature is enhanced and compared with quantile-based adaptive thresholds to detect phonetical classes. Furthermore, to save memory, the adaptive thresholds are replaced by a slope tracking method on the filtered feature. These algorithms are tested with the TIMIT database and additive white, car, factory noise, and compared with other methods to demonstrate their superior performance and robustness.

Index Terms: robust phonetic/speech classification, quantile filtering, time-scale features, wavelet decomposition.

1. Introduction

The classification of speech signals into voiced/unvoiced/silence (V/U/S) classes is crucial in various types of speech applications, such as voice activity detection, noise suppression, automatic speech recognition and speech coding. In principle, V/U/S classification relies on different feature vectors which are extracted from the input speech frames. Some classification algorithms are based on single-dimensional features such as zero crossing rate, relative energy level, autocorrelation coefficients [1, 2], linear predictive coding (LPC) and glottal closure indices (GCI) [3], MEL frequency cepstral coefficients [4] and instantaneous frequency amplitude spectrum (IFAS) [5]. Another approach combines both time and frequency domains by using the short time Fourier transform (STFT) [6] or Wavelet transform [7]. These two-dimensional signal representations certainly improve the classification rate. In order to achieve high accuracy, most algorithms require many different input parameters.

In this paper, we propose an efficient speech classifier based on a single wavelet parameter and demonstrate the robustness of the proposed algorithm. First, every windowed overlapping speech frame, which has 32ms length and 8ms overlap, is decomposed by a Wavelet decomposition (WD) at the 3^{rd} scale. Then a multidimensional feature is calculated from the Teager energy operator (TEO) of the wavelet coefficients. The extracted feature is compressed by sigmoidal function and then filtered by median filtering to enhance its robustness against noise. Second, the enhanced feature is compared with a quantile-based adaptive threshold to classify each frame into V/U/S classes. Furthermore, by applying a slope tracking method on the processed feature instead of using the adaptive threshold, the V/U/S decisions are also obtained with lower delay and memory requirements. The proposed algorithms are tested with noisy speech database with additive white, car, and factory noise over a wide range of signal-to-noise ratios (SNRs). Separate evaluations are done to study the impact of gender dependence on the algorithm.

The paper is structured as follows: the next section describes the WD, TEO, and feature extraction. Section 3 presents the quantile-based adaptive threshold method. The slope tracking method is explained in section 4. The evaluations and discussion are shown after that. The final section presents a conclusion and future research.

2. Multidimensional feature extraction

2.1. Time-scale analysis

A discrete-time signal $x[k]$ can be represented as:

$$x[k] = \sum_m \sum_n \langle \psi_{m,n}, x \rangle \, \tilde{\psi}_{m,n}[k], \tag{1}$$

where $m, n, k \in \mathbb{Z}$. The discrete-time wavelet basis function $\psi_{m,n}[k]$ is constructed from iterated filters. Based on that method, the discrete-time signal $x[k]$ can be decomposed into the sum of an approximation plus L details at L resolution stages as:

$$\begin{aligned} x[k] = &\sum_{n=-\infty}^{\infty} X^{(L)}[2n] \cdot g_0^{(L)}[k - 2^L n] + \\ &\sum_{m=1}^{L} \sum_{n=-\infty}^{\infty} X^{(m)}[2n+1] \cdot g_1^{(m)}[k - 2^m n], \end{aligned} \tag{2}$$

where

$$\begin{aligned} X^{(L)}[2n] &= \left\langle h_0^{(L)}[2^L n - l], x[l] \right\rangle, \\ X^{(m)}[2n+1] &= \left\langle h_1^{(m)}[2^m n - l], x[l] \right\rangle, \end{aligned} \tag{3}$$

are the approximation coefficients (low-frequency part) and the detail coefficients (high-frequency part), respectively, at the output of the iterated filter bank with L stages. $g_0^{(m)}[k]$ is an equivalent filter obtained through m stages of lowpass synthesis filters $g_0[k]$, preceded by an upsampler by 2. We call $W_{m,i}(n)$ the sequence of all wavelet coefficients (i.e, the $X^{(L)}[2n]$ and $X^{(m)}[2n+1]$) which are derived by WD at the m^{th} scale of the i^{th} frame, n is the coefficient index, $i \in \mathbb{Z}$.

2.2. Teager Energy Operator

As shown in [9], the TEO is an efficient nonlinear operator for many speech processing algorithms and speech applications. It can enhance the discriminability of speech components from noise [8]. In our research, the TEO expands the difference between the approximation subband and detail subbands. This improvements is very useful in case of unvoiced frames dominated by strong noise. The TEO coefficients $T_{m,i}$ are calculated by the discrete form of the TEO introduced in [9] as follows:

$$T_{m,i}(n) = W_{m,i}^2(n) - W_{m,i}(n+1)W_{m,i}(n-1). \quad (4)$$

2.3. Sigmoidal delta feature

As observed in Fig. 1, the power of the voiced frames is mostly contained in the approximation subband and much less in the detail subbands, and vice versa for the unvoiced frames. A relatively equal power distribution occurs for the silence frames. In detail, from the statistical properties of speech sounds, we observe that the spectrogram power in the range 0-1 kHz is very low for unvoiced fricative frames in comparison with voiced frames consisting of vowels and voiced fricatives. In this research, decomposition scale is chosen as $m = 3$ to consider the relation between frequency band $0 - 1kHz$ and other higher frequency bands. A delta parameter which is the power difference between approximation subband and detail subbands is extracted as:

$$D(i) = \frac{1}{N_a}\sum_{n=1}^{N_a} T_{m,i}^2(n) - \frac{1}{N_d}\sum_{n=1}^{N_d} T_{m,i}^2(n). \quad (5)$$

where $N_a = \frac{N}{2^m}$ and $N_d = N - N_a$ are the length of the approximation and detail parts, respectively, and N is number of samples in one speech frame.

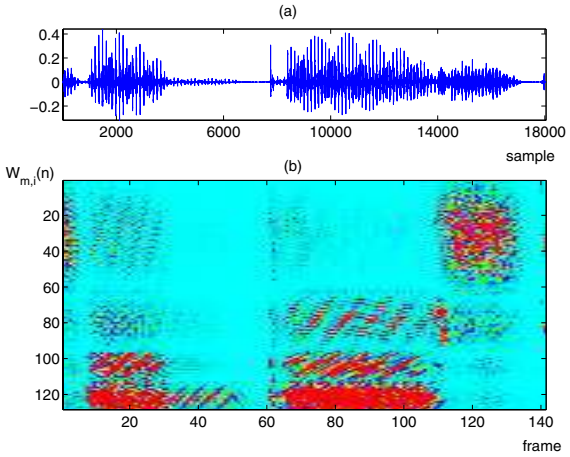

Figure 1: Waveform of speech segment (a), approximation and details at 3^{rd} scale decomposition (b).

The weak voiced or unvoiced frames result in small values of the delta D while, in general, the voiced and unvoiced frames give very high values of D with positive and negative sign, respectively. In order to balance the impact of the large range of values of D during processing, the sigmoidal function is applied on $D(i)$ as:

$$D_s(i) = \frac{2}{1 + e^{-2D(i)}} - 1. \quad (6)$$

Because of the strong noise at low SNR, the parameter D_s fluctuates with high variance even during the flat segments of silence frames. To make the classifier robust against noise, the parameter D_s is further smoothed by median filtering with window length of 4 frames still keep a low delay of the output.

3. Quantile-based adaptive threshold

We observed that the energy in each wavelet subband, and therefore the delta parameter D_s, is at the noise level over a significant part of the time. Thus, we develop a quantile-based method to estimate the adaptive thresholds related to the noise level. First, the delta values $D_s(i)$ are sorted in ascending order over a buffer of one second length with one frame shiting, then the threshold T_q is determined by taking the q^{th} quantile as show in Fig. 2. The quantile $q = 0.3$ has been selected experimentally over the range of possible values $q = 0.0, 0.1, \ldots, 1.0$.

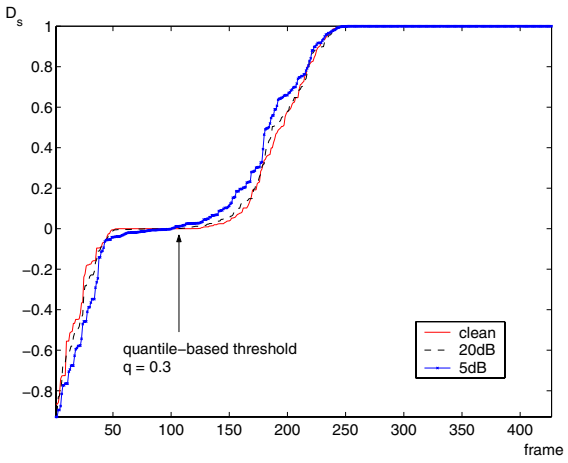

Figure 2: Quantile-based adaptive threshold for different SNRs.

To make the V/U/S decisions, the compressed delta parameter $D_s(i)$ of each input speech frame is calculated and compared with the determined threshold by the following rule:

$$D_s(i) = \begin{cases} V, & \text{if } D_s(i) > T_q \\ U, & \text{if } D_s(i) < -T_q \end{cases}. \quad (7)$$

4. Slope tracking method

The disadvantages of the quantile method are the need of memory for storing delta values in the buffer and one frame delay. To lower the memory and delay requirements, a slope tracking method on the processed parameter is proposed.

4.1. Slope generation

The frame-based values $D_s(i)$ of a speech signal are filtered by a one-pole IIR filter as:

$$D_f(i) = D_s(i) + D_f(i-1). \quad (8)$$

Although this filter is unstable, it is useful to distinguish between silence and unvoiced sounds which are visible as flat or downward slopes, respectively, at the output of filter. By using finite-length buffers for processing of input segments, the steady increase towards infinity, which results from the unstable filter,

can be contained. In general, the parameter D_s has positive values for voiced frames, negative values for unvoiced frames and approximates zero for silence frames. Because the filter operates as the cumulative sum of the elements of D_s, it results in the output parameter D_f as upward slope, downward slope, and almost flat regions for voiced, unvoiced, and silence classes, respectively (depicted in Fig. 3). In noisy environments, the filtered parameter D_f still shows the same slope characteristics as clearly even at very low SNR of Fig. 4. From that, the phonetic segments can be classified by a slope detection method which is described later.

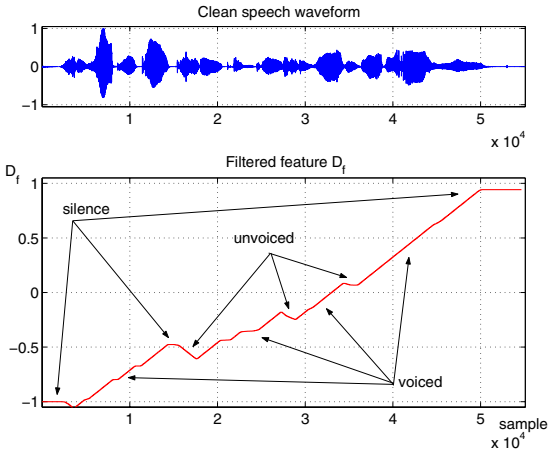

Figure 3: The filtered parameter D_f with phonetic segments for clean speech.

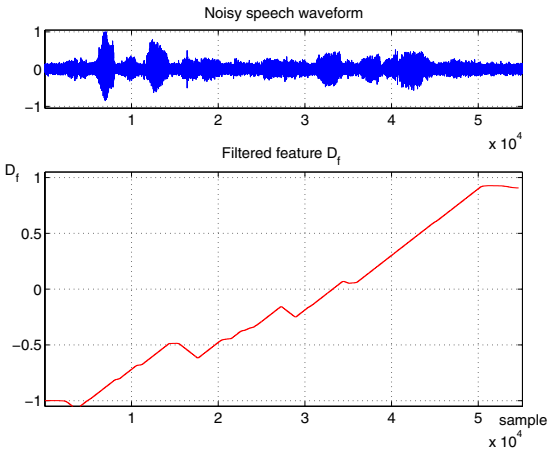

Figure 4: The filtered parameter D_f with phonetic segments for noisy speech at SNR = 5dB.

4.2. Slope detection and phonetic-smoothing

To detect the rising and falling slopes as well as the flat regions of the parameter D_f, we propose a three-step method to detect the beginning and end points of the phonetic segments as follows:

* First, if the magnitude difference of the parameter D_f between the current frame and the previous frame is bigger than a positive threshold $T_p = 0.5$ or smaller than a negative threshold $T_n = -0.1$, then the index of the previous frame is marked as the beginning point of a non-silence phonetic segment and saved in memory.
* Second, the described procedure is repeated for every couple of the neighboring frames untill it fails. That means the difference is smaller or larger than the selected thresholds. Then the total magnitude difference D_T of the smoothed parameter D_f between the beginning frame and the current frame is calculated and compared with another threshold $T_c = 10/1.5$ for voiced/unvoiced classses, respectively. If D_T is higher than T_c, the position of the current frame is marked as the end point and the segment is labeled with the corresponding class. Otherwise, the beginning point is replaced by the current point and the process is continued till the end of the buffer.
* Third, after all voiced and unvoiced segments are found, the remaining segments will be marked as silence automatically.

Finally, occasionally occurring incorrect decisions, which are due to non-stationary noise and transient sounds, are repaired by a frame-based smoothing method which enforces sequential consistency of speech sounds such as: VVSVV → VVVVV, etc. The final V/U/S labeling of a noisy recording is illustrated in Fig. 5

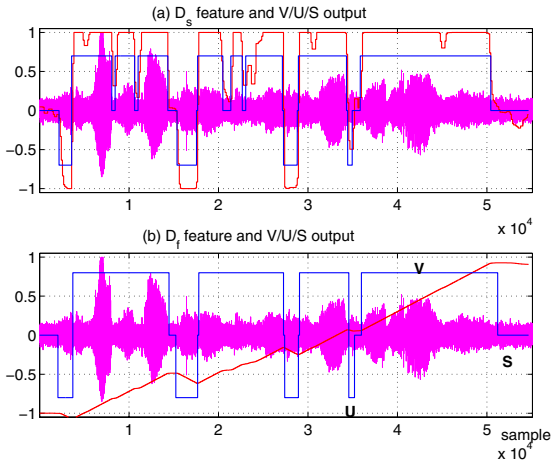

Figure 5: Sigmoidal delta D_s and V/U/S labeling (a), filtered delta D_f and V/U/S labeling (b) for white noise at 5dB SNR.

The advantage of processing with the running integrator is the saving of memory and lower delay. Only the beginning and end points need to be stored in memory. In addition, it exploits the continuity of speech sounds to classify long segments into phonetic classes without separate frame by frame detection as in most other current algorithms. This provides highly reliable performance for continuous sounds.

5. Evaluations and discussions

The speech data used to build the experiment dataset is extracted from the TIMIT database. Two gender-dependent sets of female (F) and male (M) are selected with 50 continuous utterances for each set (ca. 2000 frames). The speech signals (F_s=16kHz) are artificially corrupted with additive white, car, and factory noise over the SNR(dB) range of [30 20 10 5]. In this research, the closure and release frames of plosives are not counted because they cannot be determined as voiced or unvoiced sounds clearly. The reference labels (V/U/S) of the input speech frames which are derived from the TIMIT transcriptions are compared with the phonetic labels at the output of three classifiers which are based on the adaptive threshold (ADP), slope tracking (SLO), and GCI in [3].

Tabs. 1-3 show the average classification error rates calculated from confusion matrices over the total of frames of three classes V/U/S for the female and male datasets with clean and noisy recordings. For all three different types of noise, the ADP method always provides up to 2.5% lower error rate than SLO method for factory noise case. It is expected because the ADP method is based on a globally adaptive threshold while the SLO method is not. However, the latter method with the running integrator saves required memory during the tracking of phonetic classes. We observe that these wavelet-based methods have lower average error rates than the GCI-based method with clean speech, and noisy speech with SNRs down to 5dB. Due to the highest complexity of factory noise which includes transient, colored, and non-stationary noise, the outputs in this case drop down about 6% and 3% in comparison with the outputs of the white and car noise cases, respectively.

		Clean	30	20	10	5
ADP	F	6.23	6.62	7.63	8.47	12.49
	M	6.78	7.11	8.33	9.98	13.47
SLO	F	6.47	7.08	7.88	9.23	13.64
	M	7.63	8.98	8.92	11.17	14.23
GCI	F	7.12	7.39	8.63	10.39	15.81
	M	8.65	9.12	9.82	12.24	16.91

Table 1: Avarage error rates (%) for white noise.

		Clean	30	20	10	5
ADP	F	6.23	7.00	8.16	9.93	14.56
	M	6.78	8.26	10.02	11.25	15.98
SLO	F	6.47	7.46	8.89	11.21	16.34
	M	7.63	8.93	10.51	12.47	17.89
GCI	F	7.12	8.09	9.53	13.85	19.06
	M	8.65	9.78	11.37	11.26	21.23

Table 2: Avarage error rates (%) for car noise.

		Clean	30	20	10	5
ADP	F	6.23	7.89	9.87	13.03	18.38
	M	6.78	8.78	11.53	15.09	17.17
SLO	F	6.47	8.82	11.93	14.67	20.09
	M	7.63	8.34	13.81	16.21	23.27
GCI	F	7.12	9.78	13.05	17.96	23.81
	M	8.65	10.69	15.84	20.52	25.01

Table 3: Avarage error rates (%) for factory noise.

By analyzing the confusion matrix of the output of the wavelet-based method, we recognize that the error rate of the silence class increases by the effect of strong noise while the classification error rates of voiced and unvoiced classes are somewhat robust in noisy conditions. The performance of the proposed algorithms for clean speech are lower than the one obtained by neural network based method in [7]. This is expected because the NN-based method is more complex with 6 input features and trained only for classifying the clean speech signal.

As observed from Tabs. 1-3, it seems that the female dataset provides lower error rate than the male dataset for all three classifiers. The average difference of the average error rates over the determined range of SNRs obtained by the wavelet-based methods is lower than the result obtained from the GCI-based method, (1.33% compared with 1.77%), but higher than the one in [7] (1.14%).

6. Conclusions and outlook

High-performance phonetic classification algorithms are developed based on discrete-time Wavelet decomposition and the Teager energy operator. They exhibit a very low-complexity as the classifiers use only a single parameter. The results presented in the paper illustrate the effectiveness of the time-scale feature extracted from the wavelet coefficients. The quantile-based adaptive threshold method provides quite robust performance but requires a buffer of frames to determine the quantile threshold, and thus results in delay. The proposed three-step method for slope detection overcomes this shortcoming by using a running integrator to save memory and delay, and produces better performance than the compared methods. The testing on a larger database is necessary to validate the robustness of the proposed algorithm. Furthermore, the application of the proposed algorithms in designing VAD for robust speech recognition systems in [11] should be investigated to evaluate their effectiveness with respect to the recognition rate.

7. Acknowledgements

We kindly acknowledge the support of the European Union through the FP6 IST STREP SNOW (Services for NOmadic Workers, FP6-511587), http://www.snow-project.org/.

8. References

[1] L. Liao, M. A. Gregory, "Algorithms for speech classification", *5th ISSPA*, pp. 623-627, Australia, 1999.

[2] S. G. Tanyer, "Voice activity detection in nonstationary noise", *IEEE Trans. on Speech and Audio Process.*, Vol. 8, pp. 478-482, 2000.

[3] D. G. Childers, "Speech processing and synthesis toolboxes", *John Wiley*, USA, 2000.

[4] Z. Xiong, T. Huang, "Boosting speech/non-speech classification using averaged Mel-frequency cepstrum", *Proc. IEEE Pacific-Rim Conference on Multimedia*, Taiwan, 2002.

[5] D. Arifianto, T. Kobayashi, "Voiced/unvoiced determination of speech signal in noisy environment using harmonicity measure based on instantaneous frequency", *Proc. ICASSP*, Vol. 1, pp. 877-880, Philadelphia, USA, 2005.

[6] H. Yang, S. V. Vuuren, H. Hermansky, "Relevancy of time-frequency features for phonetic classification measured by mutual information", *Proc. ICASSP*, Vol. 1, pp. 225-228, Arizona, 1999.

[7] T. V. Pham, G. Kubin, "DWT-based phonetic groups classification using neural networks", *Proc. ICASSP*, Vol. 1, pp. 401-404, Philadelphia, USA, 2005.

[8] M. Bahoura, J. Rouat, "Wavelet speech enhancement based on the Teager energy operator", *Signal Proc. Letters*, Vol. 8, Iss. 1, pp. 10-12, 2001.

[9] J. F. Kaiser, "On a simple algorithm to calculate the energy of a signal", *Proc. ICASSP*, Vol.1, pp. 381-384, 1990.

[10] V. Stahl, A. Fischer, "Quantile based noise estimation for spectral subtraction and Wiener filtering", *Proc. ICASSP*, Vol. 3, pp. 1875-1878, Istanbul, Turkey, 2000.

[11] E. Rank, T. V. Pham, G. Kubin, "Noise Suppression Based On Wavelet Packet Decomposition and Quantile Noise Estimation For Robust Automatic Speech Recognition", *Proc. ICASSP*, pp. 477-480, Toulouse, France, 2006.