A Novel Voiced / Unvoiced / Silence Classification Scheme for Offline Speech Coding

C. Hoelper, A. Frankort, C. Erdmann, and P. Vary
Institute of Communication Systems and Data Processing (ind)
Aachen University of Technology
{hoelper,erdmann,vary}@ind.rwth-aachen.de

ABSTRACT

For use in offline speech processing systems a novel algorithm has been developed, that classifies clean speech segments robustly as voiced, unvoiced, or silence respectively. This decision is needed e.g. in source controlled speech coders which treat voiced, unvoiced, and silent segments differently, to increase the coding efficiency. The classifier is based on a combination of several features, extracted from the speech signal in the time domain. Besides the Energy, a novel measure, representing the unsteadiness of the speech signal, is proposed. Non-realtime coding allowes iterative refinement of the classification, taking the cepstral distance into account. The new classification scheme has been tested with the popular AMR codec.

1 Introduction: Automated Offline Speech Classification

Source controlled speech coders use different encoding schemes for voiced, unvoiced and silence, always yielding the best possible speech quality at minimum bit rate. For application in speech storage systems an experimental speech codec has been designed that aims at excellent speech quality with very low bit rates.

As the entire speech data is available for offline classification, the classification can be improved iteratively by repeated processing of the speech file. Another advantage is the adjustment of the decision thresholds to the properties of the individual speech file. In figure 1, the various steps of the classification process are illustrated. The different blocks will be explained in the remainder of this article.

2 Main Measures: Energy & Unsteadiness

Since one single parameter per frame does not reveal enough information about the speech as to be able to make a reliable classification, parameters have to be combined in the classification process. In this approach, two main parameters are made use of: first, the energy of the speech frame, and secondly, a measure of unsteadiness, “Ink”, which will be explained below. The applicability to the offline speech classifier of various other parameters, for example Campbell-Energy [3] and Zero Crossing Rate ZCR [8, 2], have been investigated, but the achieved results were not satisfactory. The parameter Energy is a measure which sizes the amplitude of the signal in each frame n:

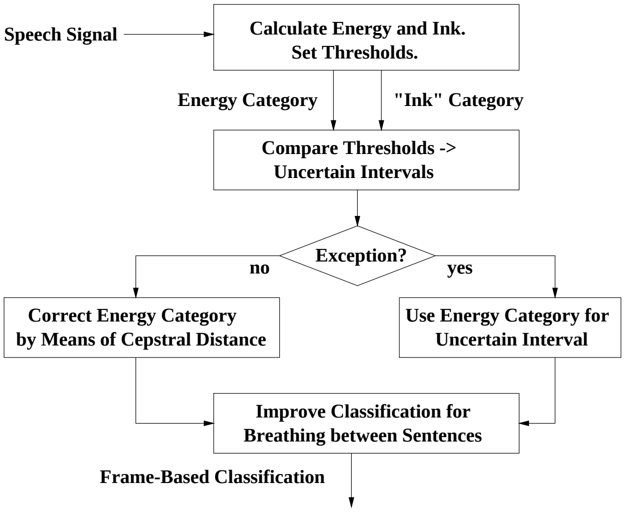
Speech Signal
Calculate Energy and Ink. Set Thresholds.
Energy Category
"Ink" Category
Compare Thresholds -> Uncertain Intervals
Exception?
no
yes
Correct Energy Category by Means of Cepstral Distance
Use Energy Category for Uncertain Interval
Improve Classification for Breathing between Sentences
Frame-Based Classification

Figure 1: *Block Diagram of the Classification Process*

$$E_n = \sum_{k=1}^{160} s^2(k,n)$$

As, in the majority of cases, voiced speech has a higher amplitude than unvoiced speech, and silent regions are distinguished by almost no amplitude, the Energy is very suitable for a first classification of speech. The parameter “Ink” is used to iteratively improve this first classification. It measures the optical density of the plotted speech signal (e.g. in the lower part of figure 3 the optical density is much higher in the right part than in the left part) and mirrors the unsteadyness, which results from number and frequency of outliers and number of zero-crossings. The parameter “Ink” can be interpreted as follows: imagining the speech signal with normalized envelope being drawn with a fountain pen, the parameter “Ink” would represent the amount of ink required.

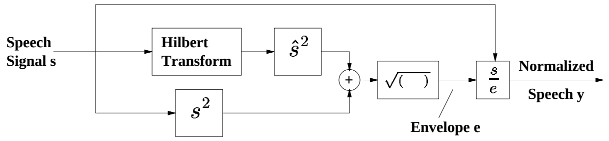

Figure 2: *Principle of Envelope Normalization with Hilbert Transform*

The envelope normalization to signal amplitudes of ±1 is necessary to prevent the amplitude from affecting the "Ink" measure, so that only the inquietude of the speech signal is recorded. This way, the energy of the signal has no influence on the parameter"Ink".

For envelope normalization of the speech, a Hilbert transformation [6] is performed, which generates the envelope of the signal, i.e. the magnitude of the analytical signal $s(k) + j\hat{s}(k)$ (see figure 2). After division of the speech signal by this envelope the normalized speech results, which fluctuates between +1 and -1 (figure 3).

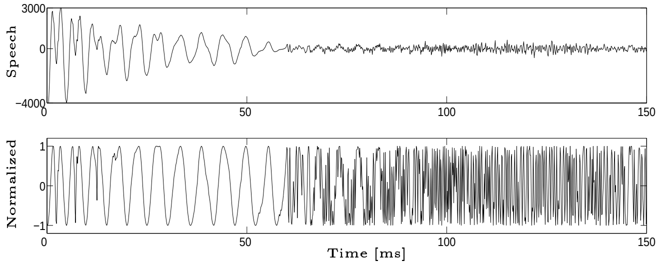

Figure 3: *a) Speech Signal s and b) Speech Signal y after Envelope Normalization*

Now the (geometric) distances d between subsequent normalized samples $y(k)$ are calculated (compare figure 4) and summed up for each frame n of 160 samples. The result is the parameter "Ink":

$$Ink_n = \sum_{k=1}^{160} \sqrt{[y(k) - y(k-1)]^{\ 2} + 1}$$

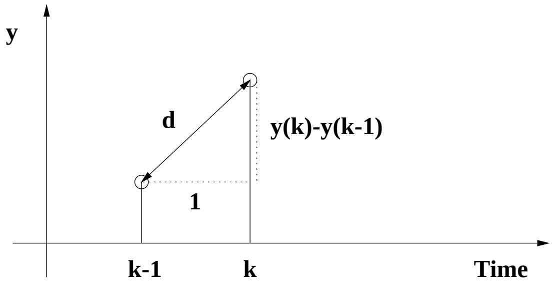

Figure 4: *Calculation of "Ink"*

As unvoiced regions are far more unsteady than voiced regions, the required amount of imaginary ink for drawing the normalized speech is much higher here. This results in the parameter "Ink" being larger in unvoiced regions than in voiced regions, whereas the parameter Energy is smaller in unvoiced regions. Thus, "Ink" and Energy behave complementary in voiced and unvoiced regions.

3 Thresholding the Parameters

After calculation of Energy and "Ink", the next step in the classification process is the setting of thresholds over the parameters (see figure 1). For the Energy, two thresholds are set, whereas the parameter "Ink" is compared with one threshold. As the range of the amplitude especially of the parameter Energy is very large, the calculations are performed on a logarithmic scale. Figure 5 qualitatively shows the behavior of the two main parameters.

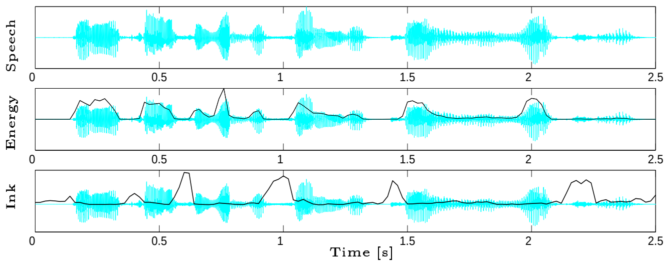

Figure 5: *Speech Signal, Parameters Energy and "Ink"*

As an upper and a lower threshold are set over the parameter Energy, the obtained Classification can adopt the values 0 (silence), 1 (unvoiced), or 2 (voiced) (see figure 6).

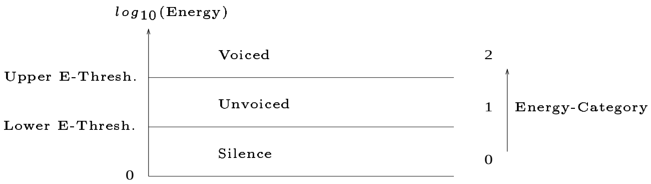

Figure 6: *Formation of the Energy-Category*

To determine suitable thresholds, tests with an experimental speech codec using different encoding schemes for voiced, unvoiced and silence have been carried out. Best results were achieved with the mean of the logarithm of Energy of the whole speech signal, $\overline{log_{10}E}$, as upper threshold for continuously spoken text. However, if much silence occurs, the upper threshold should be notably lower than $\overline{log_{10}E}$. The lower threshold, defining the boundary between unvoiced and silence, should be set at circa 2 $\sqrt{\overline{log_{10}E}}$. The plot of the Energy classification with the three classes voiced, unvoiced, and silence is illustrated in figure 7. The distinction between the classes is fairly good, though not yet error free, and can be improved in the next steps of the classification process.

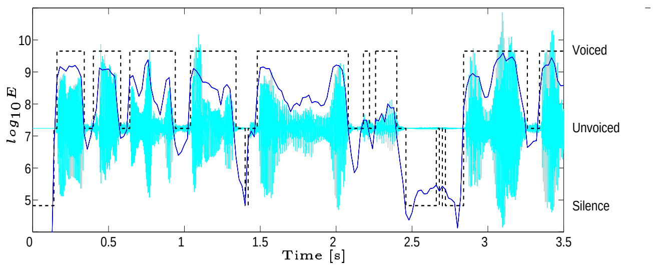

Figure 7: *$log_{10}E$, the Underlying Speech Signal and the Energy-Category with its Three Classification Classes (Dashed Line)*

There is only one threshold for the parameter "Ink". The reason for setting only one threshold is the normalization of silent regions, which causes the remaining background noise to gain the same amplitude as voiced and unvoiced regions. This normalized background noise is misinterpreted as unvoiced because of its inquietude. Thus, the parameter "Ink" only distinguishes between voiced and unvoiced segments. For the classification of silence, only the parameter Energy is taken into account. A value of $\overline{log_{10}(Ink)}$ has proven to be a useful threshold for the parameter "Ink".

4 Uncertain Intervals

As the classifications based on Energy and based on "Ink" contradict in some frames, the true classification of such frames is not known. Contradiction particularly exists at transitions from one class to the other. Intervals, in which the classifications contradict, are called uncertain intervals (see figure 8). Within these uncertain intervals, a measure of spectral distance corrects the existing Energy classification. As suggested by Hagen [4] not the common spectral distance, SD,

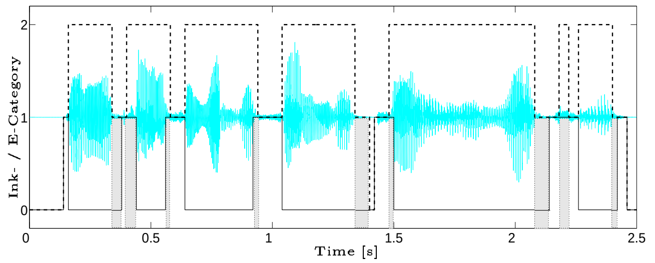

Figure 8: *Uncertain Intervals (Grey Fields) Exist at Contradiction of Energy (Dashed Line) and Ink Classification (Solid Line). 0: Silence, 1: Unvoiced, 2: Voiced.*

$$SD^2 = \frac{20^2}{2\pi} \int_{-\pi}^{\pi} \left[\log \left| H(e^{j\omega}) \right| - \log \left| \hat{H}(e^{j\omega}) \right| \right]^2 d\,\omega$$

with the production filter of the linear prediction model $H(z) = A^{-1}(z)$, is used, but the squared error in the cepstral domain, d_{cep}, which is equivalent. The cepstrum describes the envelope of the logarithmized spectrum. Cepstral coefficients of a speech frame are calculated as given e.g. in [4, 7, 9]:

$$c(i) = a(i) + \frac{1}{i} \sum_{k=1}^{i-1} k\, c(k)\, a(i-k), \quad 0 < i \leq N_p$$

with $a(i)$ as LPC-coefficients and N_p as filter order of the linear prediction, which is set to 20 in the classification process in order to obtain a high accuracy. The cepstral distance d_{cep} is calculated by comparing the cepstrum of one frame with that of the other:

$$d_{cep} = 10\, log_{10}(e) \sqrt{2 \sum_{i=1}^{N_p} (c(i) - c'(i))^{\,2}} \quad [dB]$$
$$0 < i \leq N_p$$

In this way, the cepstral distance measure d_{cep} is a measure for the spectral similarity of two frames. If the cepstrum of one frame resembles the other, the distance measure d_{cep} is small and one can expect the classification of these frames to be the same.

4.1 Correction of Uncertain Intervals by means of Cepstral Distances

Cepstra are calculated for every frame of 160 samples in every uncertain interval. By means of d_{cep}, two comparisons are performed: first to the cepstrum of the frame, that precedes the uncertain interval, which results in the measure preceding-d_{cep}, and second, to the cepstrum of the frame that succeeds the uncertain interval, which results in the measure succeeding-d_{cep}. For every frame in the uncertain interval, it must be decided, whom it resembles more: either the frame that precedes the uncertain interval, or the one that succeeds it. The solution is given by the smallest of the distance measures preceding-d_{cep} and succeeding-d_{cep}, which indicates the most similar frame. The frame in the uncertain interval then adopts the Energy classification of the most similar of the preceding and succeeding frame. It should be noted that for our experimental speech codec a misclassification as voiced only affects the gross bit rate. The voiced coding scheme does not degrade unvoiced speech.

4.2 Exceptions for the Correction of Uncertain Intervals

Not every uncertain interval can be corrected by means of preceding-d_{cep} and succeeding-d_{cep}. It would for instance be a mistake to adopt the classification "silence" for a frame in the uncertain interval, while the Energy classification definitely indicates non-silence. This case is intercepted by taking over the Energy classification without any change for such a frame, as the Energy classification already is a fairly good categorization. Another case of not modifying the uncertain interval occurs, if it contains more than 10 frames, as the correlation between most of the frames to be compared is lost.

Further, the classification of the preceding or succeeding frame cannot be taken over, if the preceding and succeeding frame have the same Energy classification, e.g. both have been marked as "unvoiced". The classification of the uncertain interval, which is enclosed by frames of the same category, would simply be overwritten, which might lead to errors. The fourth case, in which the classification of uncertain intervals cannot be corrected by means of preceding-d_{cep} and succeeding-d_{cep}, occurs, when the Energy classification changes its value twice or more within the uncertain interval. As the classification of these frames would simply be overwritten, the Energy classification is adopted without a change.

5 Improvement of the Classification for Breathing Noises between Sentences

Breathing between sentences is a noise, which does not contribute to the understandability of the speech, and therefore it shall be classified as silence only. Breaks, in which the speaker breathes loudly, are usually partly classified as unvoiced and partly as silence, notwithstanding the correction for uncertain intervals. The cause of this is a certain amount of energy coming along with the breathing sometimes, which is misinterpreted as unvoiced. In other frames though, breathing has no energy at all, which is interpreted as silence. As both classifications occur within the same breathing break, the classification fluctuates, which could be disturbing after synthesis of the speech. This problem is solved by iterating the algorithm using a lower threshold, which sets the border between "voiced" and "silence", for the classification of breathing breaks. $\sqrt{4.5 log_{10} E}$ has proven to be a good choice for the adapted threshold.

After the completion of this calculation for all breathing intervals in the speech, the thus obtained classification of breathing breaks is processed once more in order to set isolated single silence frames to "unvoiced" again. This prevents frequently audible toggeling of the classification.

6 Results

The new classification algorithm proved to correctly classify about 98% compared to hand labeled speech data taken from a German audio book. If only measures of energy are applied the correct classification rate reaches only between 85% and 90%. Working with an experimental speech codec that uses the original AMR codec mode 7.95 kbit/s [1] for voiced mode and noise excitation as suggested by Kubin, Atal and Kleijn [5] for unvoiced and silence, excellent speech quality was gained at about 4 kbit/s. Table 1 shows the distributions of the classification classes for a speech file with a length of 120 seconds. From these, it is apparent that slightly more than half of the speech frames are voiced, while about one third is unvoiced and one seventh is silent.

Voiced	54.42%
Unvoiced	31.11%
Silence	14.47 %

Table 1: Distribution of Classification for a Long Speech Signal Taken from an Audio Book

7 Conclusions

A new classification algorithm based on a combination of energy and a measure of the unsteadiness of the speech signal that is iteratively refined by means of cepstral distance has been presented. This algorithm improves the rate of correct classifications by about 10% in comparison to a classification algorithm based only on the energy measure.

References

[1] European Telecommunication Standard Institute *Digital Cellular Telecommunications System; Adaptive Multi Rate (AMR) Speech Transcoding*, GSM 06.60, Sep. 1998.

[2] Atal, B.S.; Rabiner, L.R.: *A Pattern Recognition Approach to Voiced-Unvoiced-Silence Analysis for Telephone Quality Speech"*, IEEE Trans. Acoust., Speech, Signal Processing, vol. 24, pp201-212, June 1976.

[3] Campbell Jr., J.P.; Tremain, T.E.: *Voiced/Unvoiced Classification of Speech with Application to the US Government LPC-10e Algorithm*, Proc. Int. Conf. Acoust., Speech, Signal Processing, ICASSP, Tokyo, Japan, pp473-476, 1986.

[4] Hagen, R.: *Spectral Quantization of Cepstral Coefficients*, ICASSP 94, Adelaide, Australia, April 1994.

[5] Kubin, G.; Atal, B.; Kleijn,W.B.: *Performance of noise excitation for unvoiced speech*, Proc. IEEE Workshop on Speech Coding for Telecommunications, 1993

[6] Oppenheim, Alan V.; Schafer, Ronald W.: *Discrete-Time Signal Processing*, 2nd ed., Prentice-Hall, Upper Saddle River, New Jersey, 1998.

[7] Paliwal, K.; Atal, B.: *Efficient Vector Quantization of LPC Parameters at 24 Bits/Frame* IEEE Trans. Speech and Signal Processing, vol1, pp3-13, January 1993

[8] Paulus, J.: *Codierung breitbandiger Sprachsignale bei niedriger Datenrate*, Dissertation, Institut für Nachrichtengeräte und Datenverarbeitung, RWTH Aachen, 1997.

[9] Saito,S.; Nakata, K.: *Fundamentals of Speech Signal Processing* Academic Press, 1985.