VOICED/UNVOICED/SILENCE CLASSIFICATION FOR OFFLINE SPEECH CODING

C. Hoelper, A. Frankort, C.Erdmann

Institute of Communication Systems and Data Processing (ind)
Prof. Dr.-Ing. Peter Vary
Aachen University, D-52056 Aachen, Germany
{hoelper,erdmann}@ind.rwth-aachen.de

INTRODUCTION: AUTOMATED OFFLINE SPEECH CLASSIFICATION

For use in offline speech processing systems a novel algorithm has been developed, that classifies clean speech segments 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, always yielding best speech quality at minimum bit rate. The classifier is based on a combination of features, extracted from the signal in the time domain. Besides the energy, a novel measure, representing the unsteadiness of the speech, is proposed. Non-realtime coding allows iterative refinement of the classification using the cepstral distance. Another advantage is the adjustment of the decision thresholds to the properties of the individual speech file. The new classification scheme was tested in speech storage systems with the AMR codec. In Fig. 1, left, the various steps of the classification process are illustrated. The different blocks will be explained in the remainder of this article.

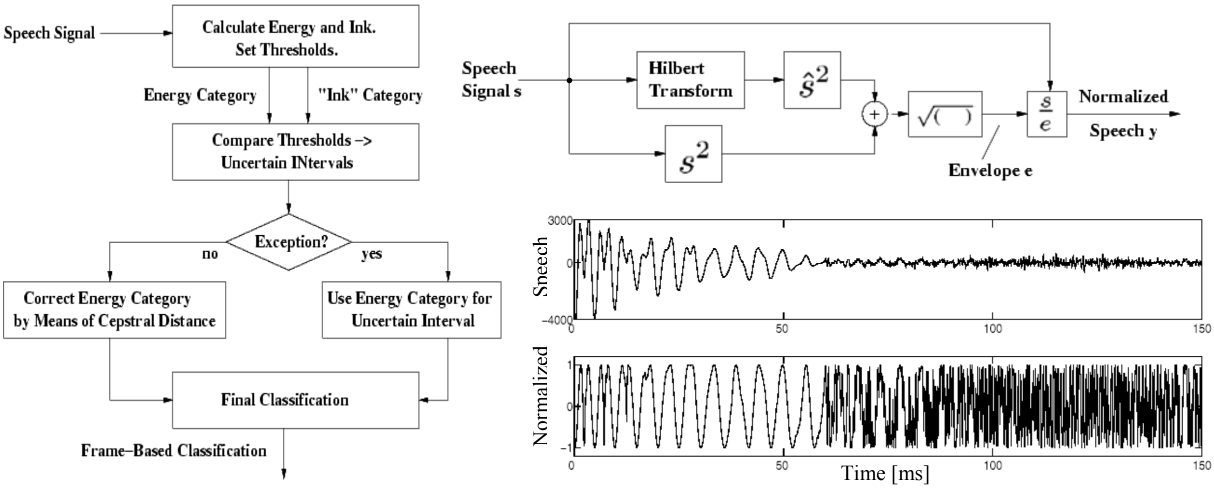

Figure 1: Block Diagram (left), Envelope Normalization (right)

CLASSIFICATION MEASURES AND REFINEMENT

One single parameter per frame does not reveal enough information about speech as to be able to acquire a reliable classification. In this approach, two main parameters are made use of: first, the energy of the speech frame, and secondly, a measure of unsteadiness, “Ink”. 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. The applicability of various other parameters was not satisfactory. Voiced speech usually has a higher amplitude than unvoiced speech, and silent regions are distinguished by almost no amplitude, which makes the energy very suitable for a first classification. 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 Fig. 1, lower right, the optical density is much higher in the right part than in the left part) and mirrors the unsteadiness, which results from number and frequency of outliers and number of zero-crossings. 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. For envelope normalization of the speech, a Hilbert transformation [4] is performed, which generates the envelope, i.e., the magnitude of the analytical signal $s(k) + j\,\hat{s}(k)$ (see Fig. 1, upper right). After division of the signal by this envelope the normalized speech results. The (geometric) distances d between subsequent normalized samples $y(k)$ are calculated and summed up for each frame to result in the parameter "Ink".

After thresholding both parameters, the classification obtained from Energy can adopt the values 0 (silence), 1 (unvoiced), or 2 (voiced). The parameter "Ink" only distinguishes between voiced and non-voiced segments, as unvoiced and silence are indistinguishable after normalization. To classify silence only the energy is taken into account. Classifications based on Energy and based on "Ink" may contradict. Contradiction particularly exists at transitions from one class to the other. Intervals, in which the classifications contradict, are called uncertain intervals. Within these, a measure of spectral distance corrects the existing classification. As suggested by Hagen [2] not the common spectral distance but the squared error in the cepstral domain d_{cep} is used, which is equivalent. The cepstrum describes the envelope of the logarithmized spectrum. Cepstral coefficients $c(i)$ of a speech frame are calculated as given in [2]:

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

with $a(i)$ as LPC-coefficients and N_p as filter order of the linear prediction. The cepstral distance d_{cep} is calculated by comparing the cepstrum of one frame with that of the other:

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

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. 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}. The classification 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. Not every uncertain interval can be corrected by means of preceding-d_{cep} and succeeding-d_{cep}. Exceptions are classifiying as "silence", while the Energy classification definitely indicates non-silence; long uncertain intervals, as the correlation between most of the frames to be compared is lost; preceding and succeeding frame having the same Energy classification and Energy classification changing its value twice or more within the uncertain interval.

CONCLUSIONS

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 7.95kbit/s [1] for voiced mode and noise excitation for unvoiced and silence [3], excellent speech quality was gained at a bit rate of less than 4kbit/s. The algorithm improves the rate of correct classifications by about 10% in comparison to a classification algorithm based only on the energy measure [5].

REFERENCES

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

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

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

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

[5] Hoelper, C.; Frankort, A.; Erdmann,C.; Vary,P., *A Novel Voiced / Unvoiced / Silence Classification Scheme for Offline Speech Processing*, European Signal Processing Conference, Toulouse, 2002