VOICE ACTIVITY DETECTION IN NOISY ENVIRONMENTS

J. Stadermann [1], *V. Stahl* [2], *G. Rose* [2]

[1]
University of Duisburg
Bismarckstr. 90, 47057 Duisburg
Phone:+49-203-379-4222,
stadermann@fb9-ti.uni-duisburg.de

[2]
Philips Research Laboratories Aachen
Weisshausstr. 2, 52066 Aachen
Phone: +49-241-6003-{563,557}
{Volker.Stahl,Georg.Rose}@philips.com

ABSTRACT

The subject of this paper is robust voice activity detection (VAD) in noisy environments, especially in car environments. We present a comparison between several frame based VAD feature extraction algorithms in combination with different classifiers. Experiments are carried out under equal test conditions using clean speech, clean speech with added car noise and speech recorded in car environments. The lowest error rate is achieved applying features based on a likelihood ratio test which assumes normal distribution of speech and noise and a perceptron classifier. We propose modifications of this algorithm which reduce the frame error rate by approximately 30% relative in our experiments compared to the original algorithm.

1. INTRODUCTION

A *voice activity detector* (VAD) is an algorithm which is able to distinguish between speech (usually distorted by noise) and noise only. The output from a VAD is a signal that possesses the information whether the input signal contains speech (e.g. output value 1) or noise only (e.g. output value 0). In order to make the problem more tractable we assume that the speech and the noise signal are stationary within a certain time interval. This assumption allows to apply conventional techniques of signal processing to this problem.

The most common features used by VAD algorithms are related to signal energy. Since the frame energy alone shows bad performance, a signal-to-noise ratio (SNR) is introduced that uses an estimation of the noise energy. Further improvement is achieved if the SNR computation is done separately for every spectral component and if probability densities are introduced for the spectral energy values. Other algorithms use features like the zero cross rate or the autocorrelation function to find discriminative features in the time domain. The classification is done by comparing results from the feature extraction with an adaptive threshold [1] or using classification algorithms from pattern recognition [2]. The algorithms in [3] and [4] combine several features to detect speech. [5] post-processes the VAD decision with so-called "hangover" methods. These methods use context (i.e. the classification result of previous frames) in order to achieve a more reliable results. This paper reviews some VAD feature extraction algorithms described in the literature and combines them with classification algorithms commonly used in pattern recognition. We propose a modification of one of the feature extraction algorithms, which improved the frame error rate by 30% relative. The paper is organized as follows: Section 2 describes our basic VAD setup and the most important VAD algorithms together with modifications, Section 3 summarizes the used classifiers, Section 4 presents results, and conclusions are given in Section 5.

2. VAD ALGORITHMS

All VAD algorithms considered in this paper are frame based. The incoming audio signal is sampled, quantized and divided into overlapping frames, then each frame is classified as either speech or non-speech. Throughout our experiments we used a frame length of 32 ms and a frame shift of 16 ms which results in an overlap of 50%.

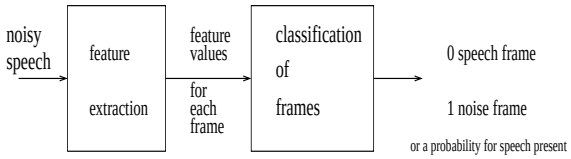

Fig. 1. Block diagram of a frame-based VAD

The basic structure of a VAD is depicted in Figure 1. The time samples of the observed signal are input to a feature extraction module, whose output are feature vectors which are classified as either speech or non-speech in the classification module. Context can be added by augmenting the feature vector by features of neighboring

frames. In the literature [1–6] many algorithms for VAD are presented. In the next section we summarize the ones which performed best in our experiments and propose some modifications and extensions. The basic idea of the presented algorithms is that the speech/non-speech decision is based on a likelihood ratio between the feature probability densities of speech and non-speech.

2.1. Likelihood ratio test based on Gaussian distribution of the input signal

In [5] and [6] it is assumed that the distribution of the time samples of speech and noise is *Gaussian* with zero mean. According to the central limit theorem (see [7]) this implies that the spectral coefficients are asymptotically independent Gaussian random variables. If we further assume that the noise is additive, we obtain $x(k) = s(k) + n(k) \circ\!\!-\!\!\bullet\; X(k) = S(k) + N(k)$. The block length of the Fourier transform is L time domain samples. Only half of the coefficients are needed in the frequency domain because we consider only real signals. Two hypotheses are set up:

Hypothesis H_0: Only noise is present, the observed time samples are $x(k) = n(k) \circ\!\!-\!\!\bullet\; X(k) = N(k)$. The probability density function (PDF) of the $\frac{L}{2}$-dimensional complex spectral vector $\vec{X} = (X(0), \ldots, X(\frac{L}{2}-1))^T$ is

$$p(\vec{X}|H_0) = \prod_{k=0}^{\frac{L}{2}-1} \frac{1}{\pi\sigma_N^2(k)} \mathrm{e}^{-\frac{|X(k)|^2}{\sigma_N^2(k)}} \qquad (1)$$

where $\sigma_N^2(k)$ is the variance (or power) of the k-th spectral coefficient of the noise signal.

Hypothesis H_1: Speech and noise are present, the observed time samples are $x(k) = s(k)+n(k) \circ\!\!-\!\!\bullet\; X(k) = S(k) + N(k)$. The PDF can be written as

$$p(\vec{X}|H_1) = \prod_{k=0}^{\frac{L}{2}-1} \frac{1}{\pi(\sigma_N^2(k) + \sigma_S^2(k))} \mathrm{e}^{-\frac{|X(k)|^2}{\sigma_N^2(k)+\sigma_S^2(k)}} \qquad (2)$$

where $\sigma_S^2(k)$ is the variance (or power) of the k-th spectral coefficient of the speech signal. It remains to estimate the unknown parameters $\sigma_N^2(k)$ and $\sigma_S^2(k)$. $\sigma_S^2(k)$ is estimated from (2) using a maximum likelihood (ML) estimation. Supposing n observations of the random variable $X(k)$, the ML estimation is

$$\hat{\sigma}_S^{2\,ML}(k) = \mathrm{E}\left[|X(k)|^2\right] - \sigma_N^2(k) \qquad (3)$$

Since we only have one instantaneous observation $X(k)$ we approximate the expectation value by (see [5])

$$\hat{\sigma}_S^2(k) = |X(k)|^2 - \sigma_N^2(k). \qquad (4)$$

Substituting $\hat{\sigma}_S^2(k)$ in (2) we can introduce the generalized likelihood ratio Λ_g as

$$\begin{aligned} \ln \Lambda_g &= \ln \frac{p(\vec{X}|H_1)}{p(\vec{X}|H_0)}\Bigg|_{\sigma_S^2(k)=\hat{\sigma}_S^2(k)} \\ &= \sum_{k=0}^{\frac{L}{2}-1} \frac{|X(k)|^2}{\sigma_N^2(k)} - \ln\frac{|X(k)|^2}{\sigma_N^2(k)} - 1 \end{aligned} \qquad (5)$$

The noise parameter set $\sigma_N^2(k)$ is estimated by its conditional expectation value

$$\begin{aligned} \mathrm{E}\left[\sigma_N^2(k)|X(k)\right] &= \mathrm{E}\left[\sigma_N^2(k)|H_0\right] \cdot Pr(H_0|X(k)) \\ &\quad + \mathrm{E}\left[\sigma_N^2(k)|H_1\right] \cdot Pr(H_1|X(k)) \\ &= \frac{\mathrm{E}\left[\sigma_N^2(k)|H_0\right]}{1+\epsilon\Lambda(k)} + \frac{\epsilon\Lambda(k)}{1+\epsilon\Lambda(k)}\mathrm{E}\left[\sigma_N^2(k)|H_1\right] \end{aligned} \qquad (6)$$

where $\epsilon = \frac{Pr(H_1)}{Pr(H_0)}$ is the *a-priori* probability ratio of the two hypotheses and

$$\Lambda(k) = \frac{p_{X(k)|H_1}(X(k)|H_1)}{p_{X(k)|H_0}(X(k)|H_0)}.$$

It remains to estimate $\mathrm{E}\left[\sigma_N^2(k)|H_0\right]$ and $\mathrm{E}\left[\sigma_N^2(k)|H_1\right]$: A very simple estimation of $\mathrm{E}\left[\sigma_N^2(k)|H_0\right]$ is the instantaneous observation $|X(k)|^2$. (In the H_0 case the speech signal is absent, hence the spectral noise variance is equal to the spectral variance of the observed signal). The term $\mathrm{E}\left[\sigma_N^2(k)|H_1\right]$ can approximately be replaced by the noise variance estimation of the previous frame. [6] suggests to replace $\Lambda(k)$ by Λ_g from eq. (5) . So, the estimator for $\hat{\sigma}_N^{\,2}(k)$ is

$$\hat{\sigma}_N^{2\,(m)}(k) \approx \frac{|X(k)|^2}{1+\epsilon\Lambda_g^{(m)}} + \frac{\epsilon\Lambda_g^{(m)}}{1+\epsilon\Lambda_g^{(m)}}\hat{\sigma}_N^{2\,(m-1)}(k) \qquad (7)$$

Based on equations (5) and (7) the following algorithm can be devised:

1. Set frame number $m := 0$ and initialize $\hat{\sigma}_N^{\,2\,(m)}(k) := |X^{(m)}(k)|^2$ for all k.
2. Evaluate equation (5), frame index $m := m + 1$
3. Evaluate eq. (7) for $k = 0, \ldots, \frac{L}{2} - 1$ and store the obtained value of $\ln(\Lambda_g)$ in the feature vector
4. go back to 2

We propose the following modification:
Eq. (7) approximates (6) by replacing the frequency-dependent likelihood ratio $\Lambda(k)$ with the generalized likelihood ratio Λ_g that is common for all frequency indices. We now stick to eq. (6) and compute a generalized log-likelihood ratio for each frequency index k:

$$\ln \Lambda_g(k) = \frac{|X(k)|^2}{\sigma_N^2(k)} - \ln\frac{|X(k)|^2}{\sigma_N^2(k)} - 1$$

then we obtain a more accurate estimation of $\hat{\sigma}^{\,2}_{N}{}^{(m)}(k)$. The generalized log-likelihood ratio needed for the speech/non-speech decision can again be computed by summing up over all frequencies:

$$\ln \Lambda_g = \sum_{k=0}^{\frac{L}{2}-1} \ln \Lambda_g(k) \tag{8}$$

2.2. The use of a hidden Markov model (HMM)

The speech/silence segmentation of a spoken utterance is usually characterized by a burst of speech frames followed by a burst of silence frames. This behavior can be taken into account if the feature value developed in the last section is modified by a first order *hidden Markov model* [5] which interprets the densities $p(\vec{X}|H_0)$ and $p(\vec{X}|H_1)$ as emission distributions of the state variables.

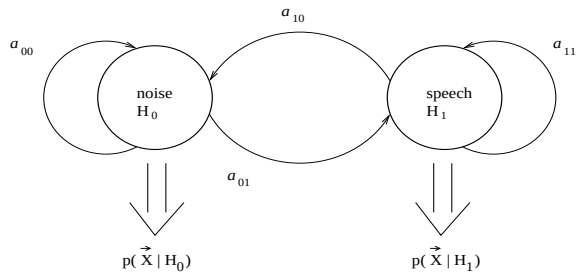

Fig. 2. Two state (speech/noise) hidden Markov model

The resulting likelihood ratio for frame m is according to [5]:

$$\Lambda_{\mathrm{HMM}} = \frac{1}{\epsilon}\,\Gamma(m) \tag{9}$$

with $\epsilon = \frac{Pr(H_1)}{Pr(H_0)}$ and $\Gamma(m) = \frac{a_{01}+a_{11}\Gamma(m-1)}{a_{00}+a_{10}\Gamma(m-1)} \cdot \Lambda_g(m)$.

3. CLASSIFIER

The classification of the feature values that are computed by the algorithms discussed in the previous section is done by two well known pattern classification algorithms:

- The least-squares classifier (LSC). The disadvantage of the LSC is that the decision boundary is determined mostly by data points which are far away from the boundary.
- The perceptron algorithm [8]. Like the simplex algorithm and support vector machines, this algorithm has the property that the decision boundary is determined mostly by data points which are near the boundary.

4. RESULTS

The evaluation of the VAD algorithms consists of two steps: The first step is to estimate the classifier coefficients based on a segmented training set. In the second step the frame error rate of the VAD is evaluated on a disjoint test set with a given segmentation. The databases used in the evaluation are the TIMIT corpus (speaker-independent, hand-segmented speech recorded in offices (high SNR)) and the CSDC-database (speaker-independent, segmented by a speech recognizer, spoken single digits or digit chains recorded in cars (low SNR), see [9]). A third database called cTIMIT was created by adding car noise at 6 dB SNR to the TIMIT database.

Table 1 presents frame error rates obtained without context information. The investigated algorithms are

- (A) simple frame energy - the frame energy is computed according to $E_{frame} = \sum_{k=0}^{L-1} x^2(k)$, where x is the observed time signal
- (B) spectral entropy (fullband, upper and lower halfband) - this algorithm is described in [3]
- (C) zero-crossing rate $\mathrm{zcr} = \frac{1}{L}\sum_{k=1}^{L-1} \left|\mathrm{sign}\left(x(k)\right) - \mathrm{sign}\left(x(k-1)\right)\right|$
- (D) log likelihood - see eq. (5)
- (E) modified log likelihood - see eq. (8)
- (F) log likelihood with HMM - see eq. (9)

Table 2 includes context: 4 frames in the past, 4 frames in the future and the current frame. Future frames are introduced by a time delay of the computation. Since the features of all frames are combined in one large feature vector the dimension of this feature vector is increased by a factor of 9 compared to the vector in Table 1. The results based on the TIMIT and cTIMIT corpus are not comparable with results based on the CSDC corpus because of a different ratio of speech and noise frames (CSDC: approx. 43% speech frames, TIMIT: approx. 77% speech frames). The reduction of the frame error rate due to the proposed modification of the VAD algorithm is more than 30% relative (standard log likelihood ratio compared to modified log likelihood ratio, both with perceptron classifier) on tests with the CSDC and with the TIMIT corpus. The introduction of context to the VAD again reduces the frame error rate by about 30% relative (regarding the best results on the CSDC corpus, this reduction is independent of the algorithm). The perceptron classifier clearly outperforms the LSC in all tests.

algorithms	dim	err	corpus	class.
A	1	41.7	CSDC	LSC
D	1	41.3	CSDC	LSC
B	3	24.5	CSDC	LSC
E and F	2	19.9	CSDC	LSC
C	1	32.3	CSDC	Perc
D	1	22.2	CSDC	Perc
E	1	15.0	CSDC	Perc
E and F	2	13.7	CSDC	Perc
E and F	2	13.0	cTIMIT	Perc
D	2	13.0	TIMIT	Perc
E	2	8.5	TIMIT	Perc

Table 1. VAD frame error rates without context (abbreviations: dim - number of components in the feature vector; err - frame error rate in %; LSC - least squares classifier; Perc - perceptron classifier)

algorithms	dim	err	corpus	class.
B	27	20.2	CSDC	LSC
E and F	18	15.2	CSDC	LSC
C	1	29.7	CSDC	Perc
D	9	14.5	CSDC	Perc
E	9	9.8	CSDC	Perc
E and F	18	9.6	CSDC	Perc
E and F	18	9.9	cTIMIT	Perc
D	9	12.6	TIMIT	Perc
E	9	5.8	TIMIT	Perc

Table 2. VAD frame error rates with context (abbreviations: dim - number of components in the feature vector; err - frame error rate in %; LSC - least squares classifier; Perc - perceptron classifier)

5. CONCLUSION

In this paper we compared several algorithms for *voice activity detection* (VAD). The performance of the algorithms was evaluated on the CSDC corpus (spoken digits with car noise), the TIMIT corpus (phonetically rich sentences in the office environment) and the cTIMIT corpus (TIMIT data with added car noise). Each algorithm was tested under the same conditions using frame based signal processing.

We investigated several feature extraction algorithms for VAD and combined them to generate larger feature vectors which were subjected to a perceptron and a least squares classifier. Context modeling was accomplished by a two state HMM.

We proposed a modification of a likelihood ratio based feature extraction algorithm, which reduced the frame error rate by over 30% relative compared to the original algorithm in our experiments.

6. ACKNOWLEDGEMENTS

We would like to thank Prof. H. Luck and Dr. T. Kaiser from the University of Duisburg and the members of the speech recognition group at the Philips Research Laboratories Aachen for scientific advice and technical support during this project.

7. REFERENCES

[1] D. K. Freenab, G. Cosier, C. B. Southcott, and I. Boyd, "The voice activity detector for the pan-european digital cellular mobile telephone service," in *ICASSP*, 1989, pp. 369–372.

[2] A. Benyassine, E. Shlomot, Y. S. Huan, and E. Yuen, "A robust low complexity voice activity detection algorithm for speech communication systems," in *IEEE Workshop on Speech Coding for Telecommunications Proceedings*, 1997, pp. 97–98.

[3] S. McClellan and J. D. Gibson, "Variable-rate celp based on subband flatness," in *IEEE Transactions on Speech and Audio-Processing*, 1997, vol. 5, pp. 120–130.

[4] J. D. Hoyt and H. Wechsler, "Detection of human speech in structured noise," in *ICASSP*, 1994, pp. 237–240.

[5] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," in *IEEE Signal Processing Letters*, 1999, vol. 6, pp. 1–3.

[6] J. Sohn and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," in *ICASSP*, 1998, pp. 356–368.

[7] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," in *ICASSP*, 1980, vol. 28, pp. 137–144.

[8] R. O. Duda and P. E. Hart, *Pattern classification and scene analysis*, pp. 138–143, Wiley, 1973.

[9] D. Langmann, T. Schneider, R. Grudszus, A. Fischer, T. Crull, H. Pfitzinger, M. Westphal, and U. Jekosch, "CSDC - The MoTiV Car-Speech Data Collection," in *First International Conference on Language Resources and Evaluation*, 1998.