Proceedings of the 6th Nordic Signal Processing Symposium - NORSIG 2004
June 9 - 11, 2004
Espoo, Finland

SPEECH PRESENCE DETECTION IN THE TIME-FREQUENCY DOMAIN USING MINIMUM STATISTICS

Karsten Vandborg Sørensen and Søren Vang Andersen

Department of Communication Technology
Aalborg University, DK-9220 Aalborg Ø, Denmark
{kvs,sva}@kom.aau.dk

ABSTRACT

The contribution of this paper is a time-frequency domain speech presence detection method that classifies power bins in the time-frequency domain as containing speech or not. An initial decision rule is based on ratios between optimally time-smoothed signal-plus-noise periodograms and weighted noise periodogram estimates, obtained from minimum statistics as proposed by Martin [1]. The initial decision rule is generalized into a weighted decomposition where the weights are obtained from off-line training by means of an artificial neural network. Experiments show that the method can be configured to be very sensitive to speech presence even in very high levels of noise and without classifying much of the noise as speech. It is shown that a fixed set of weights gives good performance at different signal-to-noise ratios indicating that the terms in the decision rule have been adequately chosen.

1. INTRODUCTION

Methods for speech enhancement are often developed explicitly without the use of speech presence detection methods. A consequence hereof is that most of these methods suffer from musical noise in the speech estimate. MMSE-LSA by Ephraim and Malah [2] handles this problem very well by both attenuating the power and lowering the dynamics of rapid changing power at bins with low signal-to-noise ratios, but some noise still remains in the estimated speech. Cohen [3] has shown that by including signal presence uncertainty in the estimation the speech quality can be improved. In this paper we propose a speech presence detection method aimed for use in any time-frequency domain speech enhancement method such that different attenuation rules can be applied for different speech presence decisions. The approach taken utilizes a trained linear combination of a few terms in a logarithmic domain. These terms are available from standard minimum statistics calculations [1]. Therefore the resulting decision rule can be implemented with very little overhead as an addition to a minimum statistics based method. We approach the problem from a simple threshold on the ratio between estimated signal-plus-noise power spectral densities and estimated noise power spectral densities and decompose this expression as terms in a logarithmic domain. This expression is easily generalized and optimized using neural network methods. This generalized framework provides a method to evaluate the importance of the different factors of minimum statistics for speech presence detection. The resulting binary decision rule can be incorporated in speech enhancement methods such that two different attenuation rules can be applied.

The remainder of this paper is organized as follows. In Section 2 we introduce the signal model and describe the terms that are used in the decision rule. The decision rule and how it is generalized and optimized using neural network methods is described in Section 3. Section 4 contains a brief description and results of the experiments and in Section 5 we give a discussion of the proposed method.

This work was supported by The Danish National Centre for IT Research, Grant No. 329.

2. SIGNAL MODEL

We assume that an observation $y(i)$ at sampling time index i consists of speech $s(i)$ and additive noise $n(i)$. We further assume that the signals are zero mean and statistically independent. For time-frequency analysis of $y(i)$ the N-point Short-Time Fourier Transform (STFT) is applied, i.e.

$$Y(\lambda,k)=\sum_{\mu=0}^{L-1} y(\lambda R+\mu)h(\mu)\exp\left(-j2\pi k\mu/N\right), \quad (1)$$

where $\lambda \in [-\infty;\infty]$ is the (sub-sampled) time index, $k \in [0;N-1]$ is the frequency index, L is the window length (in this paper we have that $L=N$), R is the number of samples that successive frames are shifted, and $h(\mu)$ is a unit power window function, i.e. $\sum_{\mu=0}^{L-1} h^2(\mu)=1$. In the time-frequency domain we have that

$$Y(\lambda,k)=S(\lambda,k)+N(\lambda,k). \quad (2)$$

To lower variance of the periodogram power spectral density estimates the observation periodograms $P_Y(\lambda,k) \triangleq |Y(\lambda,k)|^2$ are recursively smoothed over time with time varying smoothing parameter $\alpha(\lambda,k)$ at each frequency, i.e.

$$P(\lambda,k) = \alpha(\lambda,k)P(\lambda-1,k) + (1-\alpha(\lambda,k))P_Y(\lambda,k). \quad (3)$$

The optimum smoothing parameter is both derived and modified to suit a practical implementation by Martin [1]. This particular smoothing method has the ability to both lower the smoothing parameter and follow power changes when the power change is large and in particular when the total power between frames changes and at the same time provide a high degree of smoothing elsewhere for each frequency. This motivates its use in speech presence detection where the distinct edges provides easy distinction between periods with and without speech presence.

As the first step towards finding an estimate of the noise periodogram a minimum search window is applied to the smoothed observation periodograms to find biased, but representative values $P_{min}(\lambda,k)$ for the noise. The length D of the minimum search window should be chosen such that successive frequency bins with speech presence are 'bridged' by the minimum search window. This way the minimum tracked values will be unaffected by presence of speech. In order to compensate for the bias towards lower values that results from a minimum tracking of the smoothed periodograms we use the method proposed by Martin [1] where $P_{min}(\lambda,k)$ is multiplied with a bias compensation factor $B_{min}(\lambda,k)$ obtained from experiments and parametrized in the length of the minimum search window D and the estimated variances of the noise power spectral density and signal-plus-noise power spectral density. A factor $B_c(\lambda,k)$ improves the noise periodogram estimate when these estimated variances themselves have large variances. Multiplying with $B_c(\lambda,k)$ corresponds to an addition proportional to the normalized standard deviation of $P(\lambda,k)$. This leads to an estimation $\widehat{P}_N(\lambda,k)$ of the noise periodogram $P_N(\lambda,k) \triangleq |N(\lambda,k)|^2$ given by

$$\widehat{P}_N(\lambda,k) = B_c(\lambda,k)B_{min}(\lambda,k)P_{min}(\lambda,k). \quad (4)$$

3. SPEECH PRESENCE DETECTION

The main contribution of this paper is how the properties of the smoothed periodograms $P(\lambda,k)$ are exploited in a binary decision rule for the detection of speech presence. Because the presence of speech will cause a power increase in $P(\lambda,k)$ at a particular frequency it can be assumed to be higher than an estimated noise periodogram $\widehat{P}_N(\lambda,k)$ for the same time-frequency location, thus this ratio yields a robust measure (due to the smoothing) of the signal-plus-noise to noise ratio in the time-frequency bins. The smoothing will ensure that fluctuations in the speech power does not affect the speech presence detection.

3.1. Initial Binary Decision Rule

As a rule to decide between the two hypotheses $H_0(\lambda,k)$: 'No Speech Present' and $H_1(\lambda,k)$: 'Speech Present', i.e.

$$H_0(\lambda,k):\ Y(\lambda,k) = N(\lambda,k) \quad (5)$$
$$H_1(\lambda,k):\ Y(\lambda,k) = N(\lambda,k) + S(\lambda,k), \quad (6)$$

we use a binary decision rule where the smoothed observation periodograms $P(\lambda,k)$ are compared with the estimated noise periodograms $\widehat{P}_N(\lambda,k)$ weighted with γ, i.e.

$$P(\lambda,k) \underset{H_0(\lambda,k)}{\overset{H_1(\lambda,k)}{\gtrless}} \gamma\widehat{P}_N(\lambda,k). \quad (7)$$

3.2. Generalized Binary Decision Rule

The initial decision rule, with γ replaced by 10^a for later convenience, is generalized using the decomposition (4) and weights are applied as follows,

$$P(\lambda,k) \underset{H_0(\lambda,k)}{\overset{H_1(\lambda,k)}{\gtrless}} 10^a B_c^b(\lambda,k)B_{min}^c(\lambda,k)P_{min}^d(\lambda,k). \quad (8)$$

We apply the logarithm (base 10) on both sides and end up with a sum of weighted terms, i.e.

$$\log(P(\lambda,k)) \underset{H_0(\lambda,k)}{\overset{H_1(\lambda,k)}{\gtrless}} a + b\cdot\log(B_c(\lambda,k)) + c\cdot\log(B_{min}(\lambda,k)) + d\cdot\log(P_{min}(\lambda,k)), \quad (9)$$

thus well suited for training of the weights by means of an artificial neural network.

3.3. Artificial Neural Network Training

Fig. 1 illustrates the artificial neural network that is used to train the weights of the generalized binary decision rule. We have used a substitution for generality, i.e.

$$\begin{aligned} x_1 &= \log(P(\lambda,k)), & w_1 &= 1, & bias &= -a, \\ x_2 &= \log(B_c(\lambda,k)), & w_2 &= -b, & & \\ x_3 &= \log(B_{min}(\lambda,k)), & w_3 &= -c, & & \\ x_4 &= \log(P_{min}(\lambda,k)), & w_4 &= -d. & & \end{aligned} \quad (10)$$

3.3.1. Search for Initial Weights

Different costs are assigned to the two different types of errors, i.e. the cost of taking decision $D_0(\lambda,k)$ when $H_1(\lambda,k)$ is true is C_{01} and the cost is C_{10} for the opposite case. No cost is assigned for correct decisions. The neuron function $\varphi(\cdot)$ and the cost function $c(\cdot)$ used to find initial weights for the subsequent training are illustrated in Fig. 2.

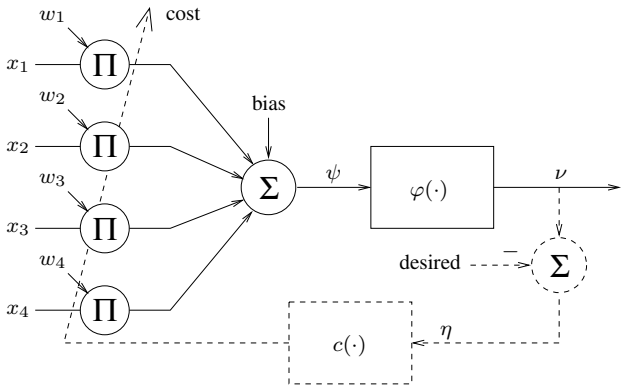

Fig. 1. The artificial neural network (solid) that is trained to find the weights of the generalized binary decision rule. The feedback loop with the cost function is dashed. $\nu = 0$ for $D_0(\lambda, k)$ and $\nu = 1$ for $D_1(\lambda, k)$, where $D_i(\lambda, k)$ means deciding on $H_i(\lambda, k)$ for $i \in \{0, 1\}$.

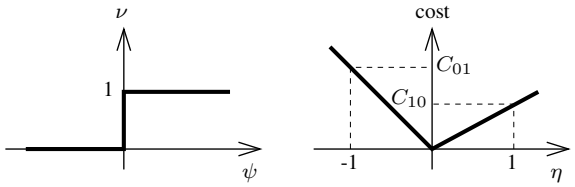

Fig. 2. The neuron function (left) and cost function (right) used in the search for initial weights.

3.3.2. Training using a Back-Propagation Error Algorithm

The initial weights are used to initialize the artificial neural network that is trained using a back-propagation error algorithm (BPEA) [4, 5], which requires differentiable neuron and cost functions. We therefore approximate the step neuron function by a differentiable sigmoid function, i.e.

$$\varphi(\psi) = \frac{1}{1 + \exp(-\beta \cdot \psi)}, \tag{11}$$

with $\beta = 1000$ and we choose the cost function to be

$$c(\eta) = k_1 \cdot \eta^2 \cdot \exp(-k_2 \cdot \eta), \tag{12}$$

with $k_1 = 4.4721$ and $k_2 = 1.4979$ which has a monotonically increasing derivative within the range of η and it satisfies $c(-1) \approx 20$, $c(0) = 0$, and $c(1) \approx 1$ (thus matching the cost of errors chosen in the experiments). Their derivatives, which are necessary in order to derive the BPEA are given by

$$\varphi'(\psi) = \beta \cdot \varphi(\psi)(1 - \varphi(\psi)), \quad \text{and} \tag{13}$$

$$c'(\eta) = k_1 \cdot \eta \cdot \exp(-k_2 \cdot \eta)(2 - k_2 \cdot \eta). \tag{14}$$

To prevent the network from being skewed towards the most recent training patterns we use cumulative weight adjustment [5] where the weights are updated after each epoch (an iteration though the whole training set).

4. EXPERIMENTAL RESULTS

The training set is composed of four different male and four different female speakers from the TIMIT database [6]. The same composure constitutes the evaluation set and not two speakers nor sentences are the same. Each set is concatenated into speech sequences of approximately 30 seconds. To have a local changing signal-to-noise ratio non-stationary highway noise is added to the two speech sets in various overall signal-to-noise ratios. Different noise recordings are used for the training and evaluation set. The constants used as part of the minimum statistics approach are the same[1] as used by Martin [1] and the sample frequency is 8 kHz. In the experiments the costs of decisions are chosen as $C_{10} = 1$ and $C_{01} = 20$. We define the periodogram bins with no speech presence as the bins below a time-frequency noise floor selected such that 5% of the clean speech power is in bins with power below the floor. This leads to a posteriori probabilities for no speech presence $P(H_0) = 0.924$ (and speech presence $P(H_1) = 0.076$) for the training set and $P(H_0) = 0.921$ (and $P(H_1) = 0.079$) for the evaluation set. Informal listening tests have shown that removing these bins from the clean speech causes a slightly tonal but still pleasant character with full speech intelligibility.

4.1. Initial Weights

The initial weights in Table 1 are obtained as the weights with the lowest total cost in the training set found among all possible combinations of $a \in \{-2, -1.99, \ldots, 1.99, 2\}$, $b \in \{0, 1\}$, $c \in \{0, 1\}$, and $d \in \{0, 1\}$.

SNR	a	b	c	d	SNR	a	b	c	d
0 dB	-0.38	0	0	0	15 dB	-0.63	1	0	0
5 dB	-0.59	1	0	0	20 dB	-0.53	1	0	0
10 dB	-0.61	1	0	0	25 dB	-0.52	1	0	0

Table 1. The initial weights that gave the lowest total cost of the training set at different signal-to-noise ratios (SNR).

To find consistency in the best initial weights it was investigated how much the total cost would increase at 0 dB signal-to-noise ratio by limiting the search to all possible combinations of $a \in \{-2, -1.99, \ldots, 1.99, 2\}$ and $[b, c, d] = [1, 0, 0]$. The resulting best initial weights were $[a, b, c, d] = [-0.58, 1, 0, 0]$ and the total cost using these weights was only 0.64% larger than the best initial weights from Table 1. We chose the initial weights obtained at 5 dB signal-to-noise ratio to be representative for all signal-to-noise ratios and investigate the performance if these initial weights were used as final weights. The results are listed in Table 2.

[1]The only exception being that $Q_{eq}(\lambda, k)$ is lower limited by 10 and not 2 as used by Martin. This change improved our implementation of Martins method [1].

	SNR = 0 dB		SNR = 5 dB		SNR = 10 dB	
$P(D\|H)$	H_0	H_1	H_0	H_1	H_0	H_1
D_0	0.80	0.04	0.88	0.04	0.92	0.03
D_1	0.20	0.96	0.12	0.96	0.08	0.97

Table 2. Conditional a posteriori probabilities $P(D|H)$ for the evaluation set when the initial weights for 5 dB from Table 1 are used at all signal-to-noise ratios.

4.2. Trained Weights

The weights from the neural network BPEA training with stepsize $\mu = 10^{-9}$ and a total cost change between epochs (more than $2 \cdot 10^5$ training samples) less than 0.001 as convergence criterion are listed in Table 3. The performance of the trained weights are listed in Table 4 for both the training and the evaluation set.

SNR	a	b	c	d	Epochs
0 dB	-0.5626	0.9919	-0.0167	0.0002	4542
5 dB	-0.5904	1.0008	-0.0048	-0.0021	1149
10 dB	-0.5896	1.0000	-0.0001	-0.0019	152

Table 3. The resulting weights corresponding to a stationary point on the total cost 'surface' when the neural network is initialized to the initial weights for 5 dB from Table 1.

	SNR = 0 dB		SNR = 5 dB		SNR = 10 dB	
$P(D\|H)$	H_0	H_1	H_0	H_1	H_0	H_1
D_0	0.78	0.04	0.88	0.03	0.93	0.03
D_1	0.22	0.96	0.12	0.97	0.07	0.97

	SNR = 0 dB		SNR = 5 dB		SNR = 10 dB	
$P(D\|H)$	H_0	H_1	H_0	H_1	H_0	H_1
D_0	0.80	0.04	0.88	0.04	0.92	0.03
D_1	0.20	0.96	0.12	0.96	0.08	0.97

Table 4. Conditional a posteriori probabilities P(D|H) for the training set (top) and evaluation set (bottom) when the trained weights from Table 3 are used.

5. DISCUSSION

We have proposed a time-frequency domain binary classification method that classifies periodogram bins as 'speech present' or 'no speech present'. The proposed method is computationally efficient once the weights of the generalized decision rule have been trained in the artificial neural network and can be included in methods with minimum statistics based noise estimation [1] without any significant increase in computational cost. If the decision method is combined with other methods the bias compensation can be omitted at a very low decrease in detector performance.

By experiments we have shown that the speech presence detection method is efficient over a large range of signal-to-noise ratios using the same weights in the decision rule for all signal-to-noise ratios. Depending on the exact application of the decision method the ratio C_{01}/C_{10} between the cost of the two different types of wrong decisions could be chosen differently which would yield other weights and different classifier properties. The high cost for falsely classifying time-frequency bins with speech presence as without speech presence used in the experiments is most useful for accurate noise estimation because the 'no speech presence' classification is unlikely to be contaminated by undetected presence of speech.

Smoothing across frequencies should be included in the decision method to smooth, and therefore attenuate, very time-frequency localized high power contributions (which are considered unlikely to have been created by the human sound production system) and prevent false 'speech present' classifications. Depending on the attenuation rules used for each of the two hypotheses a useful consequence of frequency-smoothing would be larger solid time-frequency regions with the same classification, e.g. if signal-to-noise ratio based attenuation is used for $H_1(\lambda, k)$ and bins with $H_0(\lambda, k)$ is set to zero then having large solid regions of the same decision would make the speech estimate less tonal and improve perceptual quality since noise between pitch harmonic frequencies in the voiced regions would be attenuated instead of set to zero.

A. REFERENCES

[1] Rainer Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9(5), pp. 504–512, July 2001.

[2] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33(2), pp. 443–445, Apr. 1985.

[3] Israel Cohen, "Optimal Speech Enhancement Under Signal Presence Uncertainty Using Log-Spectral Amplitude Estimator," *IEEE Signal Processing Letters*, vol. 9(4), pp. 113–116, Apr. 2002.

[4] Simon Haykin, *Adaptive Filter Theory*, Prentice-Hall, 1996.

[5] Jacek M. Zurada, *Introduction to Artificial Neural Systems*, West Publishing Company, 1992.

[6] National Institute of Standards and Technology (NIST), Gaithersburg, MD 20899, USA, CD-ROM, *DARPA TIMIT Acoustic-Phonetic Speech Database*.