A Model-based Voice Activity Detection Algorithm using probabilistic neural networks

M. Farsinejad ; M.Mohammadi ; B.Nasersharif ; A.Akbari
Department of Computer Engineering
Iran University of Science and Technology,Tehran, Iran
Email: mm_farsinejad@comp.iust.ac.ir ; Mh_mohammadi@iust.ac.ir; Nasser_s@iust.ac.ir ; Akbari@iust.ac.ir

***Abstract*- In this paper we introduce an efficient probabilistic neural networks(PNN) Model-based Voice Activity Detection (VAD) algorithm. The inputs for PNN are code excited linear prediction coder parameters, which are stable under background noise. The PNN network output is 1 or 0 to determine the nature of the period (speech or NonSpeech). Experimental results show that the proposed VAD algorithm achieves better performance than G.729 Annex B at any noise level .The performance compares very favorably with Adaptive MultiRate VAD, phase 2 (AMR2).**

Keywords: Voice activity detection, PNN-VAD, probabilistic neural networks.

I. INTRODUCTION

Voice activity detection (VAD) is defined as the process of detecting voice and silence periods during speech. VAD is required in automatic speech recognition applications such as discontinuous transmission (DTX). DTX transmits voice data only during the voice period of the conversation. If periods of silence are detected, the transmitter is turned off. Therefore, DTX increases system capacity and reduces the radiated power emission through discontinuous transmission. VAD algorithms are also useful in other applications such as speech coding, hands-free telephony, and echo cancellation. For these applications, VAD applications need robustness in the VAD decision to varying speech features and varying background noise.VAD does not only reduce the mean speech coding rate by suppressing transmission during the silence periods, but also enables recycling the use of bandwidth to improve the transmission efficiency.

Neural networks can accomplish the classification function for VAD algorithms. In this paper we propose a VAD algorithm based on a probability neural network (PNN) denoted PNN-VAD. The PNN network has been studied due to its simple architecture, avoidance of chaotic behaviour and fast learning [1]. The inputs for the PNN-VAD are three parameters, which are basically used by code excited linear prediction (CELP), which operate stably under high-level background noise [2][4]. By using the CELP coder parameters, the PNN-VAD will be able to adapt easily for any CELP based coder, and background noise does not heavily influence the output of the PNN.

This paper is structured as follows; In Section II we expose the basic structure of proposed voice activity detection algorithm and introduce features that we have applied in this approach and then we describe our smoothing and correction approach in the end of this section. In Section III we introduce our dataset and two measure for evaluation proposed VAD. These measures are false alarm rate (FAR) and false rejection rate (FRR). We simulate the proposed PNN-VAD with MATLAB software and then; In section IV we show the results and discuss about performance of proposed VAD. In this section we compare performance of the proposed VAD and AMR2 VAD against the performance of G.729B VAD.

II. BASIC STRUCTURE OF PROPOSED PNN-VAD

A. Framework

Probabilistic neural networks can be used for classification problems. When an input is presented, the first layer computes distances from the input vector to the training input vectors and produces a vector whose elements indicate how close the input is to a training input. The second layer sums these contributions for each class of inputs to produce as its net output a vector of probabilities. Finally, a compete transfer function on the output of the second layer picks the maximum of these probabilities, and produces a1 for that class and a0 for the other classes. Fig. 1 shows a basic block diagram of the PNN-VAD system.

Notice that the expression for the net input of a radbas neuron is different from that of other neurons. Here the net input to the radbas transfer function is the vector distance between its weight vector w and the input vector p, multiplied by the bias b. (The || dist || box in figure 1 accepts the input vector p and the single row input weight matrix, and produces the dot product of the two.)

The transfer function for a radial basis neuron is:

$$radbas(n) = e^{-n^2}$$

The radial basis function has a maximum of 1 when its input is 0. As the distance between w and p decreases, the output increases. Thus, a radial basis neuron acts as a detector that produces 1 whenever the input p is identical to its weight vector w. The bias b allows the sensitivity of the radbas neuron to be adjusted. For example, if a neuron had a bias of 0.1 it would output 0.5 for any input vector p at vector distance of 8.326 (0.8326/b) from its weight vector w. Fig. 2 shows a plot of the radbas transfer function.

It is assumed that there are Q input vector/target vector

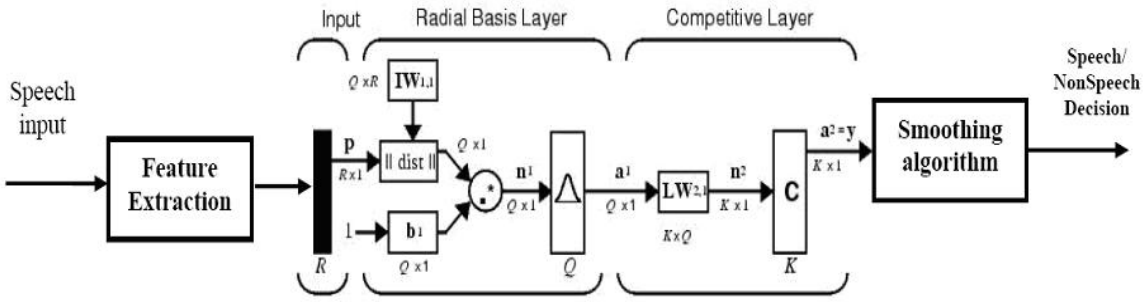

Fig. 1. Basic structure of proposed PNN-VAD.

pairs. Each target vector has K=2 elements (Speech or NonSpeach). One of these elements is 1 (Speech) and the other is 0 (NonSpeach). Thus, each input vector is associated with one of 2 classes.

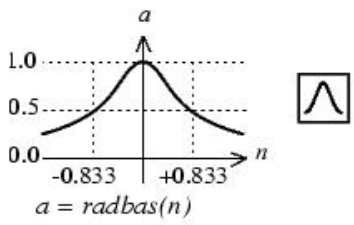

Fig. 2. plot of the radbas transfer function.

The first-layer input weights, IW1,1 (net.IW{1,1}), are set to the transpose of the matrix formed from the Q training pairs, P'. The PNN-VAD is trained on the clean (8kHz sampled) TIMIT core training set. When an input is presented, the || dist || box produces a vector whose elements indicate how close the input is to the vectors of the training set. These elements are multiplied, element by element, by the bias and sent to the radial basis transfer function. An input vector close to a training vector is represented by a number close to 1 in the output vector a1. If an input is close to several training vectors of a single class, it is represented by several elements of a1 that are close to 1.

The second-layer weights, LW1,2 (net.LW{2,1}), are set to the matrix T of target vectors. Each vector has a 1 only in the row associated with that Speech class of input, and 0's NonSpeech. The multiplication Ta1 sums the elements of a1 due to each of the 2 input classes. Finally, the second-layer transfer function, compete, produces a 1(Speech) corresponding to the largest element of n2, and 0's(NonSpeech) elsewhere. Thus, the network classifies the input vector into a specific 2 class because that class has the maximum probability of being correct.

B. Features for VAD

Three Features were used as the input variables for the PNN-VAD. These are the short-time power, the zero-order most likelihood parameter and the pitch period difference. These parameters are widely used by the CELP based coder, which works stably under a high background noise level [1][4]. These parameters are used in PNN-VAD training and the output of the PNN-VAD determine if the speech frame is either speech or NonSpeech.

The short-time power parameter is defined as the logarithm value of the zero-order autocorrelation function. The short-time power parameter is computed as follows:

$$E = \log\left[\sum_{n=m-N+1}^{m} S_w^2(n)\right] \qquad (1)$$

where the speech frame is Sw(n) with length N. The short-time power E is small at NonSpeech and is high during speech frames. The zero-order most likelihood parameter measure is generated as the logarithm sum of the squares of the linear prediction coefficients
(LPC):

$$F = 15 \log \sum_{i=0}^{p} \alpha_i^2 \qquad (2)$$

where p is the linear prediction order and αi is the LPC. This parameter is known as the distance between the voiced and the unvoiced flat spectral envelop. The pitch period difference parameter measure is generated as the normalized maximum autocorrelation. This is related to the pitch stability. The pitch period difference parameter is defined as :

$$P = \max\left[\frac{\log \sum_{t=0}^{N-1}\{r(t)r(t-\tau)\}}{\log \sum_{t=0}^{N-1}\{r(t)r(t)\}}\right] \qquad (3)$$

Where $20 \le \tau \le 160$ and r(t) is the linear prediction error signal in time t. Also, time t=0 means the beginning part of the analysis.

C. PNN-VAD Decision Smoothing And Correction

The output of the PNN network produces a 1 (Speech) corresponding to the largest element of n2, and 0's (NonSpeech) elsewhere. It is named The primary VAD decision. This primary VAD decision is smoothed (hangover) to reflect the stationary nature of both the speech signal and the background noise.

A hangover is a number of frame transition delays for determining PNN-VAD. This is because the power level is usually low and pitch is not stable at transitions from speech to NonSpeech. A hangover compensates for the period determining error. This increases the error in determining speech to NonSpeech transitions.The

transition from Speech to NonSpeech period is controlled by the duration of the hangover and helps reduce the classification errors. When using hangover frames, the hangover duration must be predetermined. After examining the results, the most efficient hangover length is 5 to 6. In every experiment, the hangover length was set as 6.

III. SIMULATION

We first conducted experimental evaluation of the PNN-VAD in speech detection performance in various noisy environments. The frame-wise false alarm rate (FAR) and false rejection rate (FRR) were used as evaluation measures. FAR is the percentage of non-speech frames incorrectly classified as speech, and FRR is the percentage of speech frames incorrectly classified as non-speech[5].

The PNN-VAD algorithm are evaluated on the (8kHz sampled) TIMIT core test set. The PNN-VAD are training from data in the training set of TIMIT. The core test set of TIMIT is used for evaluation. Core test set is composed of 12 sentences contributed by 2 male and 2 female speakers. The testing data results in 10,700 frames to classify with a frame length of 25 msec and skip rate of 10 msec. Three types of noise sources are used for noisy simulations: white (WHN), car (CAR) and babble (BAB) at 20dB, 10dB, and 0dB SNRs. The performance is evaluated in terms of the false alarm rate (FAR) and false rejection rate (FRR).

IV. RESULTS AND DISCUSSION

The proposed VAD is applied to a set of 12 sentences (about 107 seconds) from 4 different speakers; two males and two females from TIMIT database. The G.729 encoder runs on 107 frame/sec (80 samples/frame) and provides the values of energy, low-band energy, zero crossing rate, and ten Line Spectral Frequencies (LSFs) for each frame. The voice streams are corrupted by three different types of background noise; white noise, babble noise and car noise at different average SNR levels between 20 dB and 0 dB.

The performance is evaluated in terms of the probability of false rejection (FR), P_{FR}, and the probability of false alarm (FA), P_{FA}, where:

- P_{FA} is the ratio of the number of noise frames that are mistakenly classified as speech to the total number of noise frames.
- P_{FR} is the ratio of the number of speech frames that are mistakenly classified as noise to the total number of speech frames.

Furthermore we compare the proposed technique with G.729B using improvement measures P_{FR} and P_{FA} , which is given by following relations :

$$\frac{-(P_{FR|_{PNN-VAD}} - P_{FR|_{G.729}})\times 100}{P_{FR|_{G.729}}} \quad \text{(improvement in } P_{FA}\text{)}$$

$$\frac{-(P_{FA|_{PNN-VAD}} - P_{FA|_{G.729}})\times 100}{P_{FA|_{G.729}}} \quad \text{(improvement in } P_{FA}\text{)}$$

Table 1 shows a comparison between the performance of the proposed PNN-VAD and Adaptive MultiRate VAD, phase 2 (AMR2) [3] against the performance of ITU G.729 B VAD.

In general, AMR2 VAD provides the lowest FRR over G.729B VAD and the proposed PNN-VAD (with 93.02% improvement over G.729B VAD). This happens at the cost of higher false alarm rate (42.37% average degradation), specially in the case of Babble noise. On contrary, the proposed PNN-VAD provides a balanced, yet significant, improvement to G.729B for false alarm rate(FAR) and false rejection rate (FRR) ; 90.41 and 87.50%, respectively.

We note that using white noise, the improvement of the proposed technique for the false alarm rate is better than

TABLE 1. THE PERFORMANCE OF THE PROPOSED PNN-VAD AND AMR2 VAD AGAINST THE PERFORMANCE OF G.729B VAD.

NOISE TYPE	SNR (DB)	G729B		AMR2				THE PROPOSED PNN-VAD			
		P_{FR} (%)	P_{FA} (%)	P_{FR} (%)	P_{FA} (%)	Improvement in P_{FR}(%)	Improvement in P_{FA} (%)	P_{FR} (%)	P_{FA} (%)	Improvement in P_{FR} (%)	Improvement in P_{FA} (%)
Babble	20	14.49	28.14	0.28	61.08	98.07	-117.06	0.21	0.02	98.55	99.92
	10	25.92	27.21	0.08	66.60	99.69	-144.76	0.03	0.62	99.88	97.72
	0	42.12	27.51	0.08	65.12	99.81	-136.71	0.01	26.90	99.97	2.21
Car	20	16.16	10.49	0.49	14.48	96.97	-38.04	2.40	0.05	85.14	99.52
	10	27.62	10.42	0.91	12.40	96.71	-19.00	2.80	0.01	89.86	99.9
	0	39.14	10.23	14.42	4.27	63.16	58.26	2.50	0.06	93.61	99.41
White	20	17.99	10.30	0.49	11.25	97.28	-9.22	0.02	0.05	99.88	99.51
	10	30.35	10.42	1.08	11.00	96.44	-5.57	3.5	1.10	88.46	89.44
	0	48.30	10.51	5.27	7.28	89.09	30.73	20.10	0.01	58.38	99.9
Average improvement over G.729B						93.02	-42.37			90.41	87.50

that for the false rejection rate. This is because the noise is more stationary and thus easier to track. On the other hand, in the case of babble noise the improvement in the FRR of the proposed technique is better compared to the improvement of the FAR because the noise is less stationary.

V. SUMMARY

In this paper, we propose an efficient VAD algorithm to work with CELP compliant encoders in their parameter domain with minimal additional computational load for feature extraction. The proposed VAD is a probabilistic neural networks PNN-based with a two-layered processing structure.

the output of the first layer of PNN produces a vector whose elements indicate how close the input is to a training input. The second layer sums these contributions for each class of inputs to produce as its net output a vector of probabilities. Finally, a compete transfer function on the output of the second layer picks the maximum of these probabilities, and produces a 1 for that class and a 0 for the other classes. it is named The primary VAD decision. This primary VAD decision is smoothed (hangover) to reflect the stationary nature of both the speech signal and the background noise.

The proposed VAD provides a significant improvement to G.729B with a good balance between the drop in FRR and in the FAR compared with that of the G.729 B VAD.

The performance of PNN-VAD is compared to AMR2 VAD and G.729 Annex B. The results show that PNN-VAD achieves better performance than G.729 Annex B at all noise levels and also demonstrates stable performance at all noise levels.

REFERENCES

[1] Chen, S., Cowan, C.F.N., and Grant, P.M.: 'Orthogonal least squares learning algorithm for radial basis function networks', IEEE Trans. Neural Netw., 1991, 2, pp. 302–309

[2] Ikedo, J.: 'Voice activity detection using neural network', IEICE Trans Commun., 1998, E81-B, (12), pp. 2509–2513

[3] ETSI EN 301 708 V7.1.1 (1999-12), European Standard (Telecommunications series), Digital cellular telecommunications system (Phase 2+); Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) speech traffic channels; General description, (GSM 06.94 version 7.1.1 Release 1998).

[4] Kim, H.-I. Park, S.-K. Voice activity detection algorithm using radial basis function network, Electronics Letters, Vol: 40, 28 Oct. 2004D.

[5] H.Othman, T.Abdulnasr "A semi-continuos state transition propability HMM-based voice activity detection", IEEE Proceeding ,Vol. 139,No. 4, pp.821-824, 2004".