ROBUST ISOLATED WORD RECOGNITION USING WSP-PMC COMBINATION

Tzur Vaich and Arnon Cohen
Electrical and Computer Engineering Department
Ben Gurion University, P.O.Box 653
Beer- Sheva, 84105 Israel
Tel.: +972-7-6461545; FAX: +972-7-6472949; E Mail: Arnon@Newton.bgu.ac.il

ABSTRACT

A new robust algorithm for isolated word recognition in low SNR environments is suggested . The algorithm, called WSP, is described here for left to right models with no skips. It is shown that the algorithm outperforms the conventional HMM in the SNR range of 5 to 20db, and the PMC algorithm in the range 0 to -9db.

1. INTRODUCTION

Speech recognition in extreme noisy environments, is of importance in many applications. Conventional HMM recognition algorithms, trained in noiseless conditions may be used only in high SNR conditions. Compensation schemes, such as the PMC[1] must be used when the SNR is low.

The Weighted States Probabilities (WSP) is a novel algorithm used for robust isolated word recognition and spotting. It was first presented in [2]. Here we provide a short description of the WSP basic idea and in the next paragraph the WSP word recognition algorithm is described in details, for the case of left to right models with no skips. The algorithm can be modified to deal with the general HMM. In conventional HMM recognition, the recognition decision is based on the probability of the particular HMM to generate the given sequence of observations. The internal involvement of the model's states, is only indirectly employed. The WSP directly uses the "pattern" of participation of the states. The algorithm is based on the basic assumptions, that each state of the model represents a distinct "sound" of the word, and each word is represented by means of N sounds. The sequence of states, represents the sequence of sounds.

Sounds with high energy (mainly vowels) will retain part of the states pattern even under noisy conditions. The WSP algorithm (not unlike the basic mechanism of human recognition of noisy speech) uses this fact to perform robust recognition.

The WSP algorithm uses the forward variable, $\alpha_t(i)$: The probability of partial observation sequence, $O_1O_2...O_t$, and state S_i at time t, given HMM λ. The scaled coefficients set $\hat{\alpha}_t(i)$ [2] is defined as:

$$\hat{\alpha}_t(i) = \frac{\alpha_t(i)}{\sum_{j=1}^{N} \alpha_t(j)} \qquad (1)$$

$$t = 1,2,..,T\,; \qquad i = 1,2,\ldots,N$$

where N is the total number of states in the model.

The scaled coefficient $\hat{\alpha}_t(i)$ is thus the relative probability of having the partial observation sequence, $O_1O_2...O_t$ (until time t), while being in state S_i, of model λ, at time t. One can describe $\hat{\alpha}_t(i)$ as the relative probability of sound i at time t, given the partial observation sequence, $O_1O_2...O_t$ (until time t), and given the model λ.

The scaled coefficients have detailed information on pattern of the word, i.e. the speech sounds sequence. When introducing a word, W_i, to the (left-to-right) HMM λ_i, the scaled coefficients describe the word as passing from the first state (sound) at the beginning to the last state (sound) at the end of the word. The scaled coefficient is a two dimensional function, which may be presented in gray scale, in the plane (i(state number), t(frame number)). When introducing a word , W_k, to HMM λ_i, where i=k, we get a staircase like state pattern as depicted in fig. 1.

When introducing a word , W_k, to HMM λ_i, where $i \neq k$, the state pattern loses the staircase like structure since the

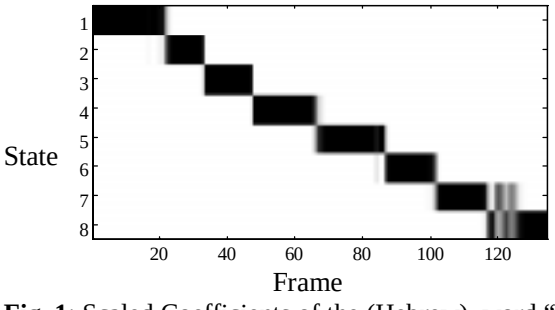

Fig. 1: Scaled Coefficients of the (Hebrew) word "**Five"** introduced to model of word "**Five".** (8 states, no skips, left-to-right HMM.)

sounds sequence is different from one word to another. The WSP does not show the model optimal states sequence but the hypothetical scaled probability of being in each one of the states at each time t.
It turns out that WSP is robust. Even under noisy conditions the pattern retains part of its staircase shape as will be shown in the results.

2. WSP RECOGNITION ALGORITHM

We assume a noisy, isolated words, recognition problem with vocabulary of L words. L left to right, N states CDHMMs, λ_ℓ ; $1 \le \ell \le L$ are trained in noiseless conditions.
The WSP word recognition consists of 3 stages: Preprocessing, Staircase similarity estimation and Recognition. The preprocessing stage provides a quantized states' pattern for each one of the vocabulary models. The second stage employs a staircase similarity measure and chooses the three models with the "best" staircase like patterns. The third stage employs a finer measure and performs the word recognition.

2.1 Preprocessing stage.

In this stage the two dimensional WSP function is quantized, smoothed and represented as a sequence of binary stair data. The preprocessing stages are:
2.a The two dimensional scaled coefficients function $\hat{\alpha}_t(i)$ is calculated for each one of the models λ_ℓ; $1 \le \ell \le L$. The data is then quantized into two bits as described in figure 2.
2.b Median smoothing is performed. For each given state, q_i ; $1 \le i \le N$, the sequence $\hat{\alpha}_t^q(i)$ is median filtered by a 5th order one dimensional median filter.
2.c All the quantized 2D WSP data is represented in terms of a sequence of K triplets $\{s_k, \ell_k, t_k\}$; $k = 1 \ldots K$, Where s_k is the k th state, ℓ_k is its duration and t_k is its start time. Each triplet describes a "stair".
2.d Smoothing is performed on the frame axis by joining consecutive triplets with the same state $(s_k = s_{k+1})$.
The processed WSP coefficients are denoted $\tilde{\alpha}_t(i)$; $1 \le i \le N$; $1 \le t \le T$ and the N*T WSP matrix By:

$$\tilde{\alpha}(i,t) = \{\tilde{\alpha}_t(i)\}\ ;\ 1 \le i \le N, 1 \le t \le T \qquad (2)$$

For example, the final WSP of figure 1 after applying steps 2.a to 2.d will be described as the triplet sequence:
$\{1,22,0\}\{2,8,22\}\{3,10,30\}\cdots\{7,24,100\}\{8,10,124\}; K = 8$
Over 90% compatibility between the optimal state, estimated in noiseless conditions, and the WSP optimal sequence, estimated in 20db SNR, was demonstrated.

```
for s = 1 to N
 for t = 1 to T
    if α̂_t(s) ≥ Th
       or
          (α̂_t(s) ≥ Th/2  and  α̂_t(s) + α̂_t(s+1) ≥ Th)
       then:   â_t^q(s) = 1.0
               â_t^q(s') = 0.0   ; s' = 1,...,N;   s' ≠ s
    else
       â_t^q(s) = 0.0
 end
end
```

Fig 2: Quantization of $\hat{\alpha}_t(i)$ (step 2.a)

2.2 The Staircase Similarity Measure (SSM)

The goal of this stage is to assign, to each one of the WSP functions, a measure (Staircase Similarity Measure -SSM), that describes how close it is to a staircase function. The SSM algorithm is described in figure 3.
The SSM is applied to models with number of states $K \ge 3$. In lower order models, it is difficult, if not impossible, to define staircase functions.
The SSM may be used as a measure for word recognition: Let SSM_T^ℓ be the Staircase Similarity Measure of the ℓ th vocabulary word, then the unknown word is classified as the j th vocabulary word where:

$$j = \underset{1 \le \ell \le L}{ArgMax} \{SSM_T^\ell\} \qquad (3)$$

3. WORD RECOGNITION

This stage serves as the fine recognition stage. It operates on the three best SSM candidates, namely the three vocabulary models with WSPs most similar to staircase functions, these are denoted Ω_3. The recognition algorithm is based on apriori information concerning the different states pdfs in each vocabulary word HMM. This information is acquired, in a non-noisy conditions, during the training stage. It is described by the Similarity Matrix (SM) and the states average duration.
The similarity matrix, SM={sm(i,j)}; $1 \le i, j \le N$, gives the probability that an observation vector that according to the optimal (Viterbi) state sequence was generated by

if $K \prec 3$,

$SSM_T = 0$

else,

initial conditions:

$SSM(1, s_1) = N + CS - s_1$

$SSM(1, j) = 0$; j = 1…N, $j \neq s_1$

loop:

for $2 \le i \le N$ *and* $1 \le j \le N$ *do*:

$$SSM(i,j) = \begin{cases} \text{for } j \neq s_i \quad SSM(i-1, j) \\ \text{for } j = s_i \\ \quad \max \begin{cases} SSM(i-1, s_i) \\ SSM(i-1, s_i - 1) + N + CS - 1 \\ . \\ . \\ SSM(i-1,1) + N + CS - 1 \end{cases} \end{cases}$$

$$SSM_T = \underset{1 \le j \le N}{Max} \{SSM(K, j)\}$$

Fig. 3: Calculation of the SSM

state i, is in fact generated by state j. Let us denote the optimal state sequence by $s^* = [q_1^*, q_2^*, ..., q_T^*]$ where q_i^* ; $1 \le i \le T$ is the state along the optimal path, at frame i, when the observation is O_i. The formal definition of the matrix is:

$$sm(i,j) = \Pr(O_t, q_t = j | q_t^* = i, \lambda) \; ; 1 \le t \le T \quad 1 \le i, j \le N \qquad (4)$$

The Similarity Matrix is estimated, from the training data, as follows: Given R training tokens of a vocabulary word ℓ, and the appropriate HMM λ_ℓ , the Viterbi algorithm is applied to all tokens r =1,2,...,R to estimate the optimal state sequence. The SM is then estimated by averaging over the tokens.

The mean duration of each one of the states is stored in a state duration sequence

$$D_\ell = [d_1^\ell, d_2^\ell, ..., d_{N_\ell}^\ell]; \quad 1 \le \ell \le L \qquad (5)$$

The scaled coefficients $\hat{\boldsymbol{\alpha}}(i,t)$ of the word to be identified are estimated under noisy conditions. As a result, the staircase like shape is disrupted. The SM contains information concerning the similarity between states, namely the states that may probably replace the "stairs" due to noise. We therefore weigh the scaled coefficients by the SM, to get the weighted scaled coefficients:

$$\hat{\boldsymbol{\alpha}}_w(i,t) = SM(i,j)\hat{\boldsymbol{\alpha}}(i,t) \qquad (6)$$

The state duration vector of the word to be identified is defined as $D = [d_1, d_2, ..., d_N]$; For each one of the candidates in Ω_3 with duration vector D_ℓ , all state sequences, $s \in S$, that comply with the following state duration constraints are computed:

$$d_j^\ell * 0.7 \le d_j \le d_j^\ell * 1.3; \quad 1 \le j \le N \qquad (7)$$

The WSP recognition measure for the ℓ th vocabulary word, $\ell \in \Omega_3$, is defined by

$$WSP^\ell = \underset{s \in S}{Max} \{ \sum_t \log(\hat{\boldsymbol{\alpha}}_w(q_t \in s, t)) \} \qquad (8)$$

The optimal WSP recognized word will be the i th vocabulary word:

$$i = \underset{\ell \in \Omega_3}{ArgMax} \{WSP^\ell\} \qquad (9)$$

The WSP recognition algorithm is used to recognize noisy words where the training is performed in a noiseless environment. The method does not require the estimation of noise statistics. As a rough check for the robustness of the method, the optimal state sequence was estimated, in noiseless condition. Applying this state sequence to the WSP yielded 99% correct recognition in SNR of 20db as compared to 91% without the WSP.

4. THE DATABASE

The WSP algorithm was evaluated with part of the Hebrew Car-Control Database (HCCD). The HCCD consists of 100 repetitions of 20 isolated words (including the ten digits) recorded from 20 speakers. These are used for the evaluation of isolated word recognition systems and word spotting systems. The HCCD also includes files of about 2 minutes duration of continuous speech. Some files include all or part of the 20 words, to evaluate word spotting systems the rest of the files include none of the words, to train garbage models. The HCCD consists of high quality speech, sampled in an acoustic room at 16kHz with 12 bits.

5. RESULTS

Two sets of experiments were conducted for the evaluation of the WSP word recognition algorithm. The first set used SNRs in the range of 20dB to 5dB, in which the recognition was performed by the WSP as described above. The second test set used data with low SNR (5dB to -9dB). In this test set the recognition was performed by WSP in conjunction with the PMC[1] algorithm. The basic idea of PMC algorithm. is to modify the noise free HMM according to the estimated noise. The PMC requires the estimation of the noise from the test data. This is usually done by locating a non speech segment from which noise characteristics are estimated. To simulate inaccuracies in SNR estimation, the PMC algorithm was evaluated here with noise which was ±3dB of the real SNR.

All experiments were evaluated using factory noise from NOISEX92 database and synthetic pink noise[1]. The word recognition results presented here are the mean over all 20 words in the database.

Table 1 shows a comparison between a conventional HMM recognition system , the SSM and the WSP. Results are shown in the SNR range of 5 to 20db.

Table 1: Comparison of conventional SSM and WSP recognition (in percent)

		SNR (db)		
NOISE		20	10	5
factory	HMM	97	54	15
	SSM	95	69	45
	WSP	96	80	61
synthetic	HMM	90	19	11
	SSM	90	51	24
	WSP	93	65	33

It is clearly seen that in the lower SNR range of 5 to 10 db, the SSM and WSP perform much better than the conventional HMM. The WSP performs better than the SSM but requires considerable extra computation. The reason for selecting the best three SSM candidates for the WSP, is demonstrated in table 2. The probability of the correct word to be in Ω_3 was found to be sufficiently high.

Table 2: Probability of the correct word to be in Ω_3

	SNR (db)		
Noise	20	10	5
Factory	99	89	69
Synthetic	98	75	40

In very low SNR both conventional and WSP methods collapse. In this range compensated HMM, such as the PMC, are used. The WSP algorithm was applied to PMC compensated HMM. Table 3 demonstrates recognition results with the PMC. The PMC requires the estimation of the SNR from the given noisy signal. Recognition results were checked with the correct SNR estimation and with mismatch of 3db.

Table 3: Recognition with PMC (factory noise) under various SNR estimation.

	Estimated SNR (db)				
True SNR (dB)	-3	-1	correct	+1	+3
0	42	71.4	80	86.8	96.4
-3	27	52.5	58.6	64.3	86.4

Note that the PMC is sensitive to SNR estimation and it is biased towards higher SNR.

The recognition results are shown in fig 3. It is seen that in the range of low SNR (-3 to -9 dB) the use of WSP with PMC has improved the recognition results. In the higher SNR range the WSP has improved the recognition results only when the estimation of the SNR was correct or lower than correct.

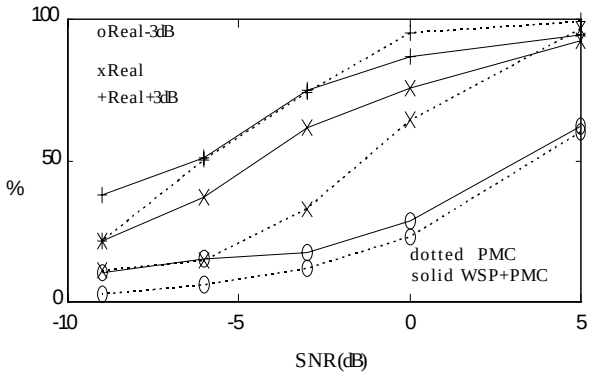

Fig. 4: Comparison of PMC and WSP+PMC methods in the low SNR range.

6. CONCLUTIONS

Two robust isolated word recognition algorithms (SSM and WSP) have been presented. For 20<SNR < 0db, the two algorithms outperform the conventional HMM. In the lower SNR range of -10<SNR<0db, the WSP applied to PMC, is superior to the conventional PMC. The results have been demonstrated for no skips, left to right CDHMM, it may however be generalized to include skips and ergodic models.

7. REFERENCES

[1] M.J.F. Gales and S. Young, "Cepstrum parameter compensation for HMM recognition in noise", *Speech Communication*, Vol. 12, pp. 321-239, 1993.

[2] T. Vaich and A. Cohen , *"Connected Word Recognition in Extreme Noisy Environment using Weighted State Probabilities (WSP)*", Proc. EUSIPCO-96, pp. 129-132, 1996.