Connected Word Recognition in Extreme Noisy Environment using Weighted State Probabilities (WSP).

T. Vaich and A. Cohen
Electrical and Computer Engineering Department
Ben-Gurion University
Beer-Sheva, Israel
arnon@bguee.bgu.ac.il

Abstract

Recognition of continuous speech in extreme noisy environments is a difficult task. A novel algorithm is suggested to enhance the performance of recognition in very low SNRs. The left to right HMM Weighted State Probabilities (WSP) method considers not only the probability of getting the given observation sequence, but also the pattern of states probabilities. On a ten digits (Hebrew) recognition task, with SNR of 10 db, the WSP has improved recognition results from 0% to 50%. It is suggested to apply the method, in conjunction with PMC enhancement algorithm, to very low SNR word spotting systems.

1. Introduction

The problem of automatic recognition of robust, speaker independent, continuous, spontaneous speech, is of great importance in many applications. In practical applications the speech is very often distorted by relatively large additive noise, providing signal to noise ratios of much lower than 20db. The general problem is very difficult to solve. The common approach is thus not to try and recognize each word in the utterance, but rather “spot” important keywords. The thesis is that having a sequence of identified keywords, the meaning of the uttered sentence may be estimated accurately even though all unimportant words were not identified. Such keyword spotting systems, require an a priori set of keywords, which is determined according to the application. The system is trained to spot the appearance of these keywords in continuous, noisy, speech. Word spotting systems automatically segment the noisy speech utterance into a sequence of keywords and “garbage” . “Garbage” is defined as all words which are not included in the keywords list. The training of such systems consists of determining models for all keywords and for the “garbage”.

Most recognition systems are based on hidden Markov models, discrete or continuous, which currently yield best performance.

It has been shown that the performance of HMM (and other) recognition systems, drastically decreases when the signal to noise drops below 20dB.

In this work we propose a new algorithm for robust speech recognition. The algorithm is termed the Weighted States Probability (WSP) algorithm. It is applied here initially to the problem of isolated word recognition, but is general and may be used for word spotting and continuous speech recognition. The method may be incorporated with any conventional enhancement algorithm. In this work we tested the algorithm with Parallel Model Combination (PMC) [1] enhancement algorithm.

2. Noise Enhancement Methods

Probably the best way to handle noisy speech is to train the HMMs under the same noisy conditions that are present at the recognition stage. This has been known to achieve the best recognition results. However, training HMM at the exact test conditions is practically impossible. We very seldom have the exact information on the noise (and speech) characteristics at the training stage. It was found that the method is very sensitive to mismatch in train and recognition conditions and thus is impractical. It is often used, however, in simulations for comparison with other noise reduction methods.

There are several approaches to preprocess the noisy speech and thus enhance recognition. Spectral subtraction is one such method. Here the power spectral density function of the noise is estimated and subtracted from that of the noisy speech. The enhanced speech is reconstructed from the subtracted PSD. Various versions of spectral subtractions have been suggested in the literature, all fail to operate in low SNR.

Another way to deal with this problem was introduced by Verga at 1990, and modified by Gales at 1993[1]. The main idea is to compensate the HMMs trained in clean conditions and to adapt them to represent the corrupted speech. The algorithm is known as Parallel Model Combination (PMC). Given the clean speech HMMs and an HMM for the noise (estimated from the noisy speech) the PDFs of the clean speech HMMs are modified, without changing the transition probabilities matrix. Using PMC algorithm Gales has shown improved recognition results. The use of PMC in our recognition system, improved recognition results of isolated words from 0% to 100% accuracy, and of connected digits from 10% to 68%, at SNR (defined on complete sentence) of 10 db. (tested on the ten Hebrew digits data base, one speaker, with synthetic colored noise).

3. Weighted State Probabilities (WSP)

In conventional HMM recognition, the probability of the particular HMM to generate the given sequence of observations is used. 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 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:

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

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

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. If one assumes that distinct sounds (e.g., phonemes, syllables) of the word being modeled can be associated with model states, one can describe $\hat{\alpha}_t(i)$ as the relative probability of the partial observation sequence, $O_1O_2...O_t$ (until time t), and sound i at time t, 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 at the beginning to the last state at the end of the word. Figure 1 depicts, in gray scale, the scaled coefficients estimated from an utterance of the word "Five" (in Hebrew) and an HMM of the word "Five". The model was a left-to-right model of order 8, with no state jumps. The staircase like state pattern depicted in fig. 1 exhibits the internal structure of the scaled coefficients of the given model when the correct utterance is used.

When introducing a word , W_k, to HMM λ_i, where $i \neq k$, the state pattern loses the staircase like structure since the sounds sequence is different from one word to another. Figure 2 depicts the state pattern of the word "Five" with the model of "Seven". The data of both figures is a very high signal to noise ratio data. For the WSP to be effective, the states patterns must be insensitive to noise.

If the states patterns are to be used for recognition, some kind of distortion measure has to be defined in order to be able to detect the desired pattern. In this work we have decided to perform the pattern classification by an artificial neural network. Several NN structures were tested, a 3 layers network with 21 input nodes, trained with back-propagation was chosen.

The suggested system consists of word models and classification NN, trained under no noise conditions, without any compensation for the (unknown) noise. The basic idea that motivates the

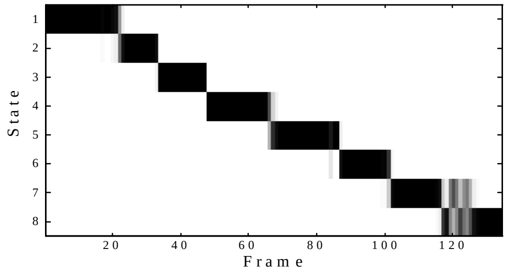

Fig. 1: Scaled Coefficients of the word "**Five"** introduced to model of word "**Five".**
(gray scale; black - high probability) .

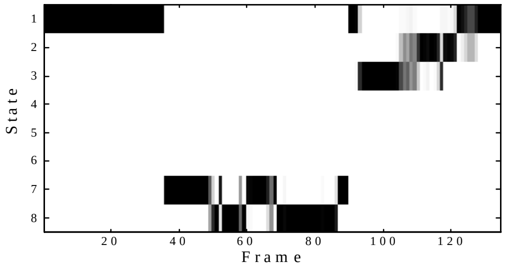

Fig. 2: Scaled Coefficients of the word "**Five"** introduced to model of word "**Seven".**

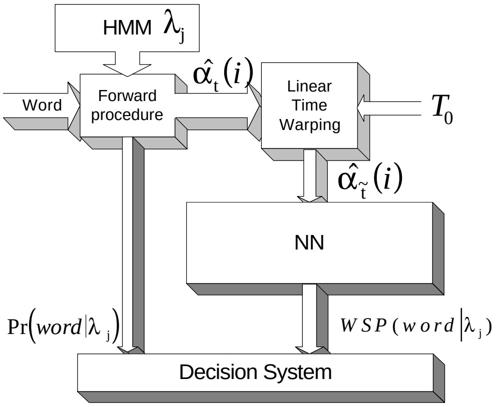

Fig. 3: Block diagram of the proposed system.

algorithm is that sounds with high energy (mainly vowels) will retain part of the states pattern even under noisy conditions, allowing the classification NN to recognize the existence of the structured pattern. This is not unlike the basic mechanism of human recognition of noisy speech. Fig. 3 shows the block diagram of the suggested system.

4. 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 several 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 and some don't, to train garbage models. The HCCD consists of high quality speech, sampled in an acoustic room at 16kHz.

5. Experiments

The speech analysis was performed on 16 msec frames with frame rate of 250 frames per second. From each frame 20 Mel Frequency Cepstral Coefficients - MFCC [3] were extracted, estimated from 256-point FFT power spectrum with 20 mel-scale band filter bank. The HMM word models were trained using noise-free speech. 8 states, continuous, left-to-right whole word HMMs were used, with single mode Gaussian with full covariance matrix. In this preliminary test we used one set of HCCD, 100 repetitions of the 10 digits, one speaker. The HMM models were trained with 90 repetitions of each one of the ten digits. 10 Repetitions were used as test group.To simulate the noisy speech, additive synthetic pink noise was employed by passing white noise through a band pass filter H(z).

$$H(z)=[1-1.4z^{-1}+1.14z^{-2} \\ -0.896z^{-3}+0.32z^{-4}]^{-1} \qquad (2)$$

The noise was added digitally to the speech to generate the required SNR. The SNR was determined for each segmented/isolated word.

For the training of the NN we used two groups of inputs the first one consisted of WSP of words, W_i, generated from matched HMM λ_i for the representation of the correct word pattern, and the second were those WSP of word , W_k, generated from HMM λ_i, where $i \neq k$ representing wrong patterns. Since the first layer of the NN has a constant number of inputs, linear time warping

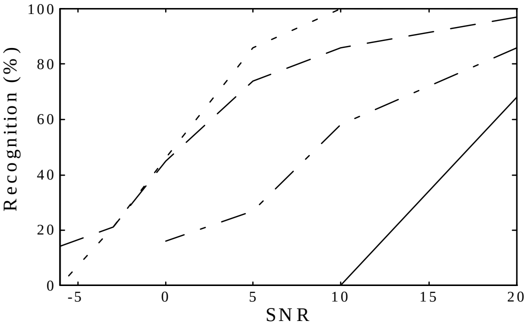

Fig. 4: Recognition results- Solid: Conventional uncompensated HMM; Dashdot: WSP; Dash: PMC with WSP; Dotted: PMC.

(LTW) was used to adjust the dimension of the scaled coefficients $\hat{\alpha}_t(i); t=1,\ldots,T$ to a constant size T_0 (T_0=100 was used here). The input to the NN is a combination of the LTW signal $\hat{\alpha}_{\tilde{t}}(i); \tilde{t}=1,\ldots,T_0, i=1,\ldots,N$.

The NN used had 21 inputs to the first layer. The first 8 inputs i=1,8 received $\hat{\alpha}_{\tilde{t}}(i)$; the next 7 inputs i=9,15 received $\hat{\alpha}_{\tilde{t}}(i-8)-\hat{\alpha}_{\tilde{t}}(i-7)$, while the last 6 inputs i=16,21 received $\hat{\alpha}_{\tilde{t}}(i-15)-\hat{\alpha}_{\tilde{t}}(i-13)$. This input combination allowed the NN to have both static and dynamic information on the scaled coefficients. The hidden layer consisted of 4 nodes, and the output layer had 1 node. The NN was trained with clean speech using 0.9 desired output for $\hat{\alpha}_{\tilde{t}}\left(i\middle|W_j,\lambda_j\right)$, and 0.1 desired output for $\hat{\alpha}_{\tilde{t}}\left(i\middle|W_j,\lambda_k; j\neq k\right)$. The NN was trained with 100 repetitions of correct patterns and 900 of incorrect patterns. Figure 4 describes the performance of WSP vs standard recognition using both uncompensated HMM and PMC compensated HMM.

6. Discussion and Conclusions

The two lower curves in figure 4 show recognition results for the uncompensated case. The WSP proves to be much superior to the uncompensated HMM. The upper two curves show compensated results. The addition of WSP to PMC compensated HMM somewhat decreases recognition rates in the higher SNR range and increases recognition rates at very low SNRs.

The results depicted in figure 4 are for the WSP alone without the use of the conventional observation probability (see fig. 3). Basing the final decision on both WSP and observation probability has the potential of improving recognition results to 93% at 5dB and 59% at 0dB. In order to get information of best recognition achievable under the given conditions two additional tests were performed. The recognition system was tested with HMMs that were trained with the same noise as that presented in the test. The second test was that of human recognition. The results are given in figure 5.

The work presented here is a preliminary work on WSP that proves its potential. We are currently working on incorporating the conventional observation probability into the decision strategy. We have strong indications that WSP will provide better recognition results than the PMC in SNRs below 0db.

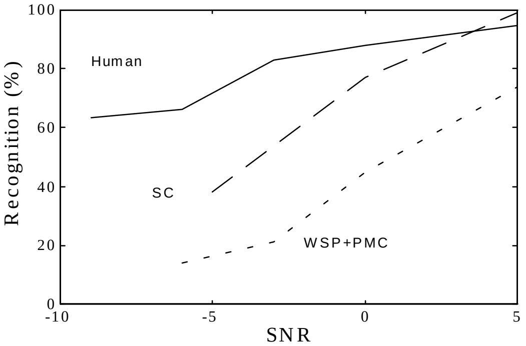

Fig. 5: Recognition results- Solid: Human; Dash: Training and test in same noisy conditions; Dotted: PMC with WSP.

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. L.R. Rabiner, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”, Proc. of the IEEE, Vol. 77, No. 2, February 1989.

3. J.W. Picone, “Signal Modeling Techniques in Speech Recognition”, Proc. IEEE, Vol. 81, pp. 1215-1247, 1993.