AN HYPOTHESIZED WIENER FILTERING APPROACH TO NOISY SPEECH RECOGNITION

Alberto D. Berstein and Ilan D. Shallom

DSP Group, Inc., 4050 Moorpark Avenue, San Jose, CA 95117 U.S.A.

ABSTRACT

The problem of Speech Recognition in a noisy environment is addressed. Particularly the mismatch problem originated when training a system in a "clean" environment and operating it in a noisy one. When measuring the similarity between a noisy test utterance and a list of clean templates a correction process, based on a series of Wiener filters built using the hypothesized clean template, is applied to the feature vectors of the noisy word .The filtering process is optimized as a by product of the Dynamic Programing algorithm of the scoring step. Tests were conducted on two data bases, one in Hebrew and the second in Japanese, using additive white and car noise at different SNRs. The method shows a very good performance and compares well with other methods proposed in the literature.

1. INTRODUCTION

The performance of speech recognition systems designed to work in noise free conditions is strongly affected by the presence of noise. If the system has to be operated in different noise environments, training the system in one environment and operating it in a different one leads to a mismatch problem responsible for a poor performance.

In contrast with other methods which use a speech enhancement step in order to input to the recognition system with noise reduced utterances, the key feature of the proposed method is doing the filtering at the scoring step using the information present in the clean templates

The proposed method performs a feature correction on the noisy tests utterances in order to eliminate noise effects. The correcting mechanism is based on optimal filtering and on Dynamic Programming. The optimal filter at each state of the Dynamic programming is based on information present in each state. This correction is performed when computing the local distance between two feature vectors, one pertaining to a clean template and the other to a noisy test word. An estimate of the background noise and different template hypothesis are used to built the correcting filters. In this way the decisions are made using all the useful information present in the recognizer. The method was implemented using the system depicted in Figure 1. The system is trained only in a quiet environment and the background noise present in the operating environment (assumed to be stationary or slowly time variant) is learned in the neighborhoods of the words to be recognized. The Hypothesized Wiener Filtering (HWF) mechanism is performed during the DTW scoring step.

2. THE METHOD

In the following discussion we assume a speaker dependent isolated word recognition system , but the method may be extended to other kinds of application. The system is trained in a "clean" environment using well established techniques [1]. A set of Q templates is assumed to be available. Each template is built as a series of feature vectors $[\underline{C}_\alpha(1),\underline{C}_\alpha(2),\ldots,\underline{C}_\alpha(J)]$ where J is the number of frames in the template.

The following assumptions are made:
- A smoothed spectral representation of each frame can be obtained from the feature set.
- An estimate of the power spectral density (PSD) of the background noise is available.

2.1. THE LOCAL DISTANCE

When measuring the similarity between a noisy frame of the test and a clean frame of one of the templates the following steps are performed:
1. A PSD estimate ($P_S(w)$) of the clean

CH2977-7/91/0000-0913 $1.00 © 1991 IEEE

frame is computed.
2. A PSD estimate ($P_Y(w)$) of the noisy frame is computed.
3. A PSD estimate ($P_d(w)$) of the background noise is computed.
4. A Wiener filter is evaluated according to

$$W(w) = \frac{G \cdot P_S(w)}{G \cdot P_S(w) + P_d(w)} \qquad (1)$$

where G is a SNR matching gain defined later
5. A modified spectral representation for the test frame is obtained multiplying its PSD ($P_Y(w)$) by the Wiener filter transfer function. When the spectral shape of $P_S(w)$ is similar to that of the underlying speech spectrum present in the noisy frame this filter will "clean" otherwise will change the spectral shape of the noisy frame according to an unmatched filter. This modified spectral shape is used to calculate a corrected feature vector.
6. A similarity measure is calculated between the clean feature vector and the corrected test vector.

The basic assumption of the approach is that the correction leads to a minimum global distance between a template and a noisy word only when the hypothesis is correct, i.e. when the feature vectors used to build the filters pertained to a template which is similar to the patterns present in the noisy test. The SNR matching gain G is calculated according to an estimate of the SNR of the test frame. This gain does not modifies the spectral shape of the clean template. A shape modification can be introduced as a function of this SNR estimate in order to alleviate the Lombard effect [2] . This issue was not addressed in this paper because the noisy words were built adding noise to noise-free words so the mentioned effect is absent in our data bases.

2.2 THE MODIFIED DTW ALGORITHM

The fact that we use the clean template when building the Wiener filters leads to a series of filters which are time aligned with the template and not with the test which has to be filtered. We have here two related problems: in order to get the similarity measure through DTW we must first "filter" the feature vectors of the test word, but in order to build the filter series we need the warping function. The two problems can be solved by the Dynamic Programming algorithm present in the DTW step. A modification of the local distance definition in the standard algorithm will find the warping function which matches the Wiener series of filters giving the corrected test vectors while minimizing the accumulated distance between the clean template and the modified test word.

Well known, straight forward, transformations exists between the different features sets obtained from an LPC analysis. We chose here to use the cepstrum feature set in order to reduce the computational effort, but any other set can be used.

Before reaching the DTW algorithm, a series of parameters are estimated as follows.

Each template word is represented by a series of AR models. An LPC analysis is performed on each frame using the autocorrelation method and the associated cepstrum vector and PSD are calculated from the LPC vector $\underline{\alpha}$ according to:

$$C_i = \alpha_i - \sum_{k=1}^{i-1} \frac{k}{i} C_k \alpha_{i-k} \quad i = 1,p \qquad (2)$$

and

$$P_S(w) = \frac{g^2}{\left| 1 - \sum_{k=1}^{p} \alpha_k \cdot e^{-jkw} \right|^2} \qquad (3)$$

$P_S(w)$ is a smoothed power spectrum and the computation of (3) can be performed applying an FFT to the zero padded LPC vector. The power level of each frame $R_S(0)$ is used in order to set the model's gain g according to Parseval's Theorem.

The Background noise is modeled by an AR model which is estimated using an average autocorrelation vector $\underline{R}_b$. This vector is obtained by averaging the instantaneous autocorrelation vector estimates $\underline{R}_d(n)$. The averager is activated by a Voice Operated Switch (VOX) only in those frames were speech activity is not detected.

In an Isolated word application the noise present before and after each test word is learned. The correlation vector $\underline{R}_b$ is transformed into an LPC vector $\underline{\beta}$ and the PSD of the noise is estimated according to eq. 3 were $\underline{\alpha}$ is replaced by $\underline{\beta}$.

After marking the endpoints of the test word an LPC analysis is also conducted on the noise corrupted frames. The cepstral vectors are evaluated and the power level of each frame is saved.

The searching mechanism finding the warping function giving the minimum

accumulated distance is performed according to a DTW with boundaries constraint relaxation [3].

Now entering the DTW grid we place on the vertical axis one of the Q available templates. We have for each frame a spectral representation, its power level and the corresponding cepstral vector. On the horizontal axis we place the test utterance characterized, by the cepstral vector and the power level of each frame. Also the Background noise estimate $P_b(w)$ and power level are available. Let us assume that the reference is of length J and the test is of length I the situation is depicted in Fig. 2.

For example at point (i,j) the calculations needed to correct $\underline{C}_r(i)$ and measure the local distance are:

1. The Wiener filter $W_{i,j}(w)$ is evaluated using eq. 1 in the specific point of the grid

$$W_{i,j}(w) = \frac{G_{i,j} \cdot P_s(w,j)}{G_{i,j} \cdot P_s(w,j) + P_b(w)} \quad (4)$$

where $G_{i,j}$ is a SNR matching gain defined as

$$G_{i,j} = [R_y(0,i) - R_b(0)]/R_s(0,j) \quad (5)$$

negative gains are avoided by clipping G at a minimum value. This filter is built under the hypothesis that $P_s(w,j)$ is similar to the speech spectrum present in the i^{th} noisy test frame.

2. The Cepstral representation of $W_{i,j}(w)$ ($\underline{C}_w$) is calculated. This can be done by taking the natural logarithm of each component "w" of the transfer function, performing an Inverse Fourier Transform (IFFT) and truncating the resulting vector to its first "p" components using a raised cosine lifter.

3. Compute the local distance $d_{i,j}$ at the specific point of the grid as given by

$$d_{i,j} = \| \underline{C}_\alpha(j) - (\underline{C}_r(i) + \underline{C}_w) \|^2 \quad (6)$$

The addition of the component $\underline{C}_w$ causes homomorphic filtering which is equivalent to filtering the noisy speech by the Wiener filter. The correction component $\underline{C}_w$ is evaluated using all the useful information present at the scoring step : clean template, background noise estimate and noisy test SNR.

3. RESULTS

The method was tested at the simulation level using two data bases: one in Japanese and the other in Hebrew. The Japanese vocabulary included 20 common Japanese names and the Hebrew vocabulary consists of ten digits and 6 confusable words. The recordings were made in a quiet environment and the noisy words were created by adding either white noise or car noise recorded in a car running on a high way. The mixing process was done at different signal to noise ratios. Several processing methods were compared, recognition rates are given in the following Tables:

Table I:
Data Base 20 Japanese names (6 repetitions) + Car noise (High Way).

SNR(dB)	HWF	IW	NP
10	96.6	92.5	91.6
5	94.1	89.1	88.3

Table II:
Data Base 16 Hebrew words (4 repetitions)

SNR(dB)	HWF	IW	SMC	NP
clean	93.7	91.6	87.5	93.7
White Noise				
10	89.5	77.1	91.6	35.4
5	77.1	33.3	72.9	16.6
0	52.0	18.7	50.0	12.5
Car Noise				
10	93.7	91.6	(*)	91.6
5	93.7	91.6	(*)	83.3
0	83.3	85.4	(*)	68.7

1. Hypothesized Wiener Filtering (HWF) (Proposed method)
2. Iterative Wiener Filtering (IWF) (According to Ref. 4)
3. No Processing (NP) (Approach ignoring noise presence)
4. Short-Time Modified Coherence) (According to Ref. 5)

(*) The SMC method was developed for additive white noise therefore results are given only for this type of noise.

5. CONCLUSIONS

The use of all the information present in the speech recognizer at the decision phase leads to a filtering approach which copes with the mismatch problem. The results, although based on

a small data base, show that the method is robust even in harmful conditions like car noise. Further experiments will include testing with larger data bases in order to reach statistically meaningful conclusions.

The application of the hypothesized wiener filtering (HWF) approach was depicted in a DTW based isolated word recognition system, but extensions can be made to other kind of systems, including HHM based recognizers.

REFERENCES:

[1] Rabiner, L.,R. and Wilpon, J.,G., "A Simplified, robust Training procedure for Speaker Trained, isolated word recognition systems", J. Acoust. Soc. Am. 68(5), Nov. 1980.

[2] Soong, F.,K. and Sondhi, M.,M., "A Frequency-Weighted Itakura Spectral Distortion Measure and Its Application to Speech Recognition in Noise", IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. 36, No.1, January 1988.

[3] Shallom, I.,D., Haimi-Cohen, R. and Golan, T.,"Dynamic Time Warping with Boundaries Constraint Relaxation", Proc. of the 16th Conference of IEEE in Israel, March 1989.

[4] Lim, J.,S., "All-Pole modelling of Degraded Speech", IEEE Trans. On Acoustics, Speech and Signal Processing, Vol. ASSP-26, No.3, June 1978.

[5] Mansour, D. and Juang, B.,H., "The Short Time Modified Coherence Representation and Noisy Speech Recognition".IEEE Trans. On Acoustics, Speech and Signal Processing, Vol. ASSP-37, No.6, June 1989.

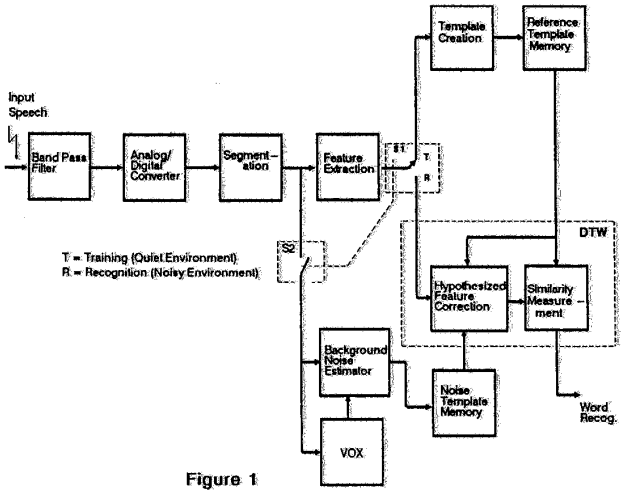

Figure 1

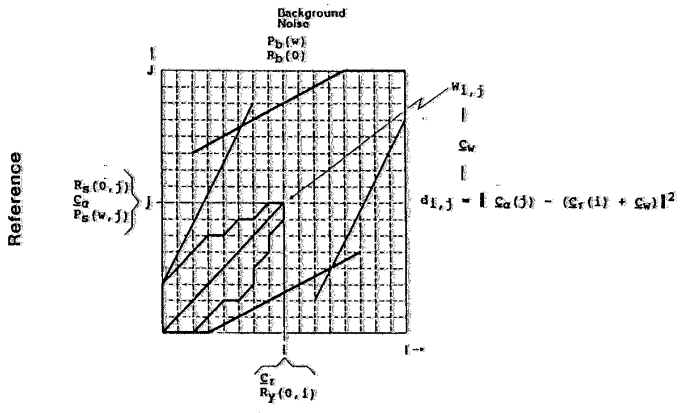

Figure 2