J. Indian Inst. Sci., May–June 1999, **79**, 185–196.

Matched filtering approach to robust speech recognition

J. V. AVADHANULU* AND T. V. SREENIVAS
Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore 560 012.
tvsree@ece.iisc ernet.in

Abstract

Robustness of the performance of the automatic speech-recognition (ASR) systems has become important because of the widespread deployment of ASR in various information technology applications. This paper addresses robustness to environment noise in the speech signal. The speech pattern matching is recast as a sequence of sub-pattern matching problems in the time-frequency domain. Each sub-pattern matching is formulated as a 2D matched filter, which is known to be an optimum detector. This sequential detection is shown to provide robust recognition of the overall pattern. The new approach to ASR is evaluated on a limited vocabulary, speaker-dependent, isolated word-recognition task in an automobile acoustic environment and the results are promising.

Keywords: Short-time Fourier transform, 2D-matched filter, noisy speech, speaker-dependent ASR.

1. Introduction

Speech recognition has evolved from science fiction to applied research and commercial reality in the last four decades. While research has contributed to the understanding of the human speech communication process, digital signal processing (DSP) has provided the technology to translate the research knowledge into practical systems. A simple block diagram of a speech-recognition system is shown in Fig. 1. The short-time amplitude spectrum (or some transformation of the amplitude spectrum) represents the essential information in the speech signal. A test speech pattern is compared with the stored 'reference' patterns in such a spectral domain and the pattern classifier outputs the *best* match.

Much of the understanding of speech signal properties is gained through the spectrogram.[1] A more compact and effective representation that has spurred significant advances in automatic speech recognition (ASR) is the source/filter model of the vocal tract using linear prediction. A different approach to feature analysis involves modeling human auditory system using a set of non-uniformly spaced, overlapping bandpass filters followed by other nonlinear processing. Perceptual linear prediction (PLP)[2] has combined these two principles by applying a lower-order LPC analysis to the speech processed by a perceptually motivated filter bank. The other important block in Fig. 1 is the pattern classifier. The pattern classifier optimally aligns the sequence of test pattern vectors with the reference pattern vectors, taking into account the variability of speaking rate. Two important pattern classifiers in ASR are template matching and hidden Markov modeling (HMM) techniques. In template matching, optimally chosen reference speech pattern (template) consisting of a sequence of feature vectors is

*Currently with M/s Sigmatech, Bangalore 560 064.

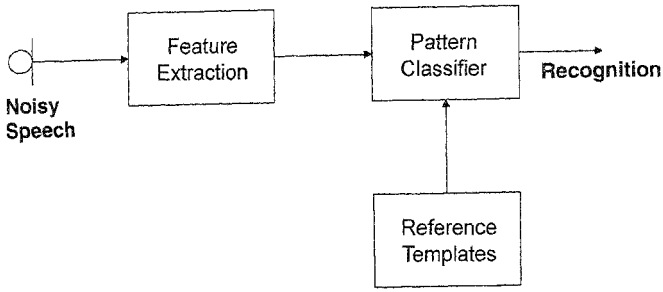


FIG. 1. Block diagram of speech-recognition system.

stored for each word of the vocabulary which is estimated during the training process. An unknown (test) speech pattern is compared with all the reference patterns to find the best match. Because of the variability in speaking rates, a dynamic time warping (DTW) technique (attributed to Itakura[5]) is used to stretch or shrink the time axis to minimize the distortion between templates. In contrast, an HMM[1] incorporates more information about variability of the statistical speech patterns which is a particularly attractive approach to speaker-independent ASR.

In this paper, we propose a novel approach to ASR using two-dimensional matched filter (MF). Section 2 deals with the formulation of MF approach using a time-frequency representation (TFR) of speech signal. In Section 3, we identify the issues in the application of MF to noisy speech and address them at some length. In Section 4, we discuss two alternative time-alignment techniques using matched filtering approach. We present the results of the new approach applied to a limited vocabulary, speaker-dependent, isolated word-recognition (SD-IWR) task in Section 5, followed by concluding remarks.

2. Matched filtering for ASR

Figure 2a shows the time-domain plot of a speech signal, $x(n)$, of the proper noun 'John Smith' spoken by a male adult, sampled at $Fs = 8$ kHz. The short-time Fourier transform (STFT) of the signal is given by

$$X_{STFT}(k,n) = X_{STFT}\left(e^{i\omega},n\right)\Big|_{\omega=2\pi k/N} = \sum_{m=0}^{L-1} x(n-m)w(m)e^{-j2\pi mk/N}, 0 \le k \le N-1, \forall n. \quad (1)$$

The amplitude spectrogram of the speech signal obtained through STFT (with $N = 256$, $L = 128$ and successive window overlap of 87.5%) is shown in Fig. 2c. Figure 2b shows the thresholded STFT plotted as an equilevel contour. The islands of high energy in the spectrogram are easily noticeable. Due to the large amplitude dynamic range of speech signal components, many components of the signal in the time-frequency domain have high SNR even at

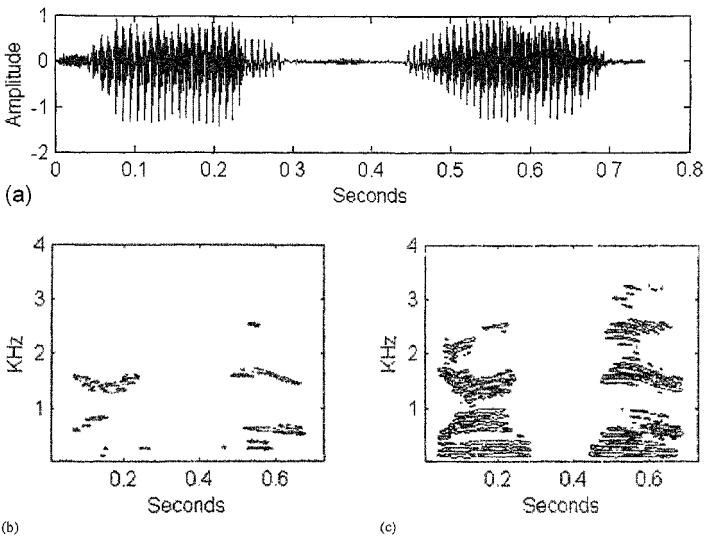


FIG. 2. Estimation of matched filter; a. clean speech signal, b. thresholded spectrogram, and c spectrogram.

a low overall SNR and are likely to be less affected by the additive noise. The STFT being a linear transform, there are no cross terms, unlike other transforms such as Wigner distribution. Hence, the high-energy regions in the time-frequency domain are robust signal components, albeit of limited resolution because of the STFT window function.

Considering an additive noise model of speech, given by $x(n) = s(n) + \gamma(n)$, its STFT spectrum can be viewed as an additive noisy pattern: $\mathcal{E}|X_{STFT}(k, n)|^2 = |S(k, n)|^2 + \mathcal{E}|\Gamma(n, k)|^2$, under the assumption of uncorrelated noise. While the time signal is highly fluctuating and of low SNR, local regions of $|X_{STFT}|$ would be slow varying and of high SNR. Thus, for the purpose of noisy speech recognition, it would be advantageous to focus on these local regions than on the whole pattern. Also, for known signals in additive noise, matched filtering is an optimum detector. Hence, we formulate an ordered set of matched filters, each of which is an optimum detector of the local region of the whole speech pattern.

Let $S(k, n)$ be a two-dimensional complex-valued signal (estimated speech pattern). Let $S(k, n)$ be corrupted by additive colored noise $\gamma(k, n)$ (of power spectral density $S_\gamma(\omega_k, \omega_n)$); thus,

$$X(k, n) = S(k, n) + \gamma(k, n). \tag{2}$$

The matched filter for optimum detection of $S(k,n)$ from $X(k,n)$ is a 2D linear filter $H(k,n)$ that maximizes the output SNR.[4] The output of the matched filter is given by

$$Y(k, n) = H(k, n) * X(k, n) \tag{3}$$

where * denotes convolution. The optimum matched filter is given by

$$H(k, n) = \Gamma(k, n) * X(-k, -n), \text{ where } \Gamma(k, n) = F^{-1}\{1/S_f(\omega_k, \omega_n)\}. \tag{4}$$

The optimum matched filter theory provides for maximizing the power of the output signal with respect to the power of stationary noise in the signal. In the context of the TFR of speech, each of the local regions occupies a fraction of the signal bandwidth. If the local region is defined over $k_1 < k < k_2$ and $[(k_2-k_1)/(F_s/2)]$ is small, we can assume $S_f(\omega_k, \omega_n)$ to be uniform over the local region. This assumption permits us to neglect the $\Gamma(k, n)$ term in (4) and use simple $H_i(k,n)$ obtained from the local regions of the TFR of clean speech as matched filters to process the test utterance. With this simplification, we get

$$Y(k, n) = S^*(-k, -n)*X(k, n) = R_{XS}(k, n). \tag{5}$$

For the case when there is a translational shift in the observed signal, the optimum matched filter is modified to $S^*(-k -k_0, -n -n_0)$, leading to the measure of $R_{XS}(k + k_0, n + n_0)$. Often we do not know (k_0, n_0); hence, the optimum matched filter corresponds to $\max_{(k, n)} \{R_{XS}(k, n)\}$, for best detection of the sub-pattern.

As indicated earlier, for each speech pattern we can extract several local regions of H_i from the clean signal, (see Fig. 2c)

$$H_i(k, n) = X_{STFT}(k, n).W_i(k, n) \tag{6}$$

where . denotes element by element multiplication and $W_i(k, n)$ is a binary-valued function that defines the local region. It may be noted that $R_{XS}(k, n)$ can in general be complex valued and we need to combine the detection measures of several matched filters cohesively. Hence, without much loss of generality, we can define

$$y_i = \max_{s} \left| R_{XS_i}(\Delta k, \Delta n) \right| = \max_{(\Delta k, \Delta n)} \left| \sum_k \sum_n \left(X(k, n).\, W_i(k, n) \right) H_i(k - \Delta k, n - \Delta n) \right|$$

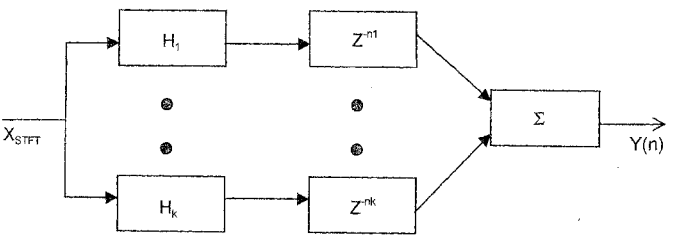


FIG. 3. Matched filtering of speech.

$$= \max_{(\Delta k, \Delta n)} \left| \sum_k \sum_n \left(X(k,n).\, W_i(k,n) \right) S_i^*(k - \Delta k, n - \Delta n) \right| \tag{7}$$

For speaker-dependent ASR pattern matching, we can simplify the above search to only the n-axis because the frequency variations are minimal within the same speaker patterns. Thus, we can write

$$y_i(n) = y_i(0, n) = |H_i(k, n) * X(k, n)|, \forall n. \tag{8}$$

If we have K such local regions over the STFT of the reference signal, we can formulate a time-ordered set of matched filters, each of which is an optimum detector of the local region and can realize the optimum 'receiver' as

$$Y(n) = \sum_{i=1}^{K} y_i(n) z^{-n_i}. \tag{9}$$

Equation (9) indicates that the optimum receiver is realized by 'delay and sum' of the time-ordered output of matched filters. When $y_i(n)$ are delayed by n_i, the peaks are aligned and

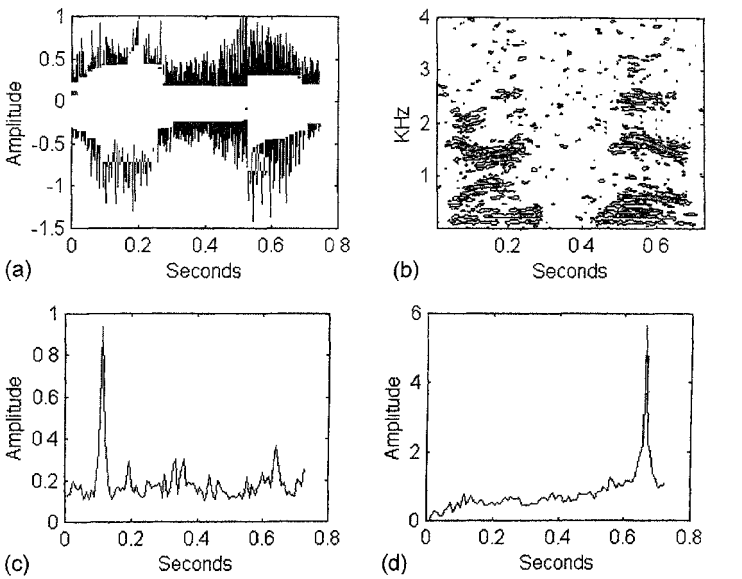


FIG. 4. Example of matched filtering in white noise; a. clean speech plus white noise, b. spectrogram, c. $y(n)$, d. $Y(n)$.

hence the summing operation yields a processing gain of $M^{(1/2)}$. The structure of this signal processor is shown in Fig. 3, which is reminiscent of the matched filter processor used in pulse-compression radar.[5]

We now show some experimental results to demonstrate the efficacy of the MF approach to ASR.

Figure 4a shows the waveform $x(n) + \gamma(n)$ (obtained by adding white Gaussian noise to the signal shown in Fig. 2a) at an SNR of 0 dB, over the entire word. The corresponding amplitude spectrogram is shown in Fig. 4b. The output $y_1(n)$ from matched filtering the STFT$[x(n) + \gamma(n)]$ using H_1 is plotted in Fig. 4c. Even though Fig. 4c shows a single peak, other peaks may occur if the word contains multiple instances of the TF pattern corresponding to H_1. If the time-aligned peak outputs from H_1 to H_6 are summed together, the resulting signal $Y(n)$ will have a peak amplitude close to six (as plotted in Fig. 4d). Figure 5 shows the results of the experiment with $x(n) + \gamma_c(n)$ (obtained by adding colored noise recorded in an automobile). The results are similar to the experiments conducted with the white noise case; this validates the assumption of uniform distribution of $S_\gamma(\omega_k, \omega_n)$ over H_i.

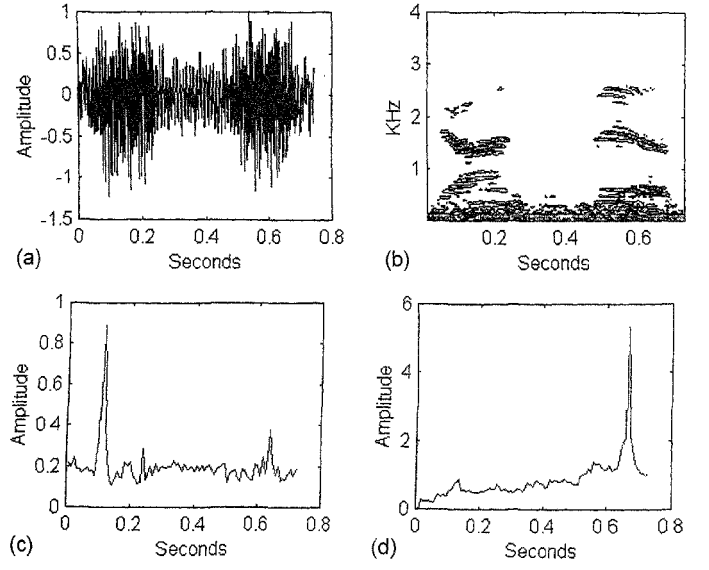


FIG. 5. Matched filtering in coloured noise; a. speech plus colored noise, b. spectrogram, c. $y(n)$, d. $Y(n)$.

3. Matched filtering of noisy speech

In Section 2, we have presented experimental results of adding noise to clean speech, where the underlying speech signal remains identical. In real noisy speech recognition, difficulties arise due to the variability in speaking rate and articulation effects.[1] Due to these factors, no two speech utterances of the same word (even by the same speaker) are identical. In particular, the matched filtering formulation is based on identifying high SNR 'islands' in the speech pattern. At low overall SNRs, the 'islands' of noisy speech can be quite different from that of clean speech.

For example, Fig. 6a shows the pattern of noisy speech ('John Smith') recorded in a moving car and Fig. 6b the corresponding spectrogram. The similarity of the high-energy regions with those of Fig. 2b is evident to the eye, not withstanding corruption by noise and variability of speech. However, Fig. 6c shows $y_j(n)$ and Fig. 6d $Y(n)$, which are much less promising than those obtained when noise was added to clean speech. We now analyse the reasons for this poor performance and put forward some strategies for MF of noisy speech.

From the TFR of speech, we can see that speech can be represented as a sum of sinusoids with complex time-varying envelopes:

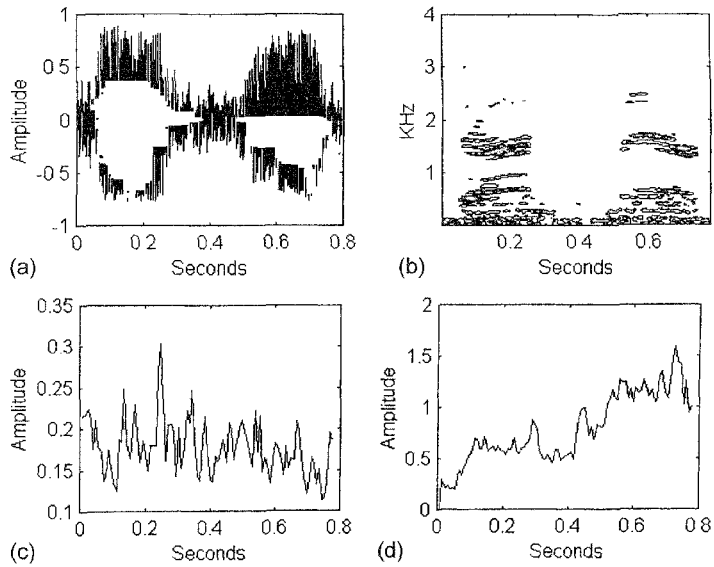


FIG. 6. Matched filtering of real noisy speech, a. noisy speech, b spectrogram, c $y(n)$ using clean speech filters, and d. $Y(n)$.

$$S(t) = \sum_{i=1}^{p} A_i(t)\left[\cos \omega_i(t) + a_i(t) g_i(t) + \Phi_i\right] \tag{10}$$

where $A_i(t)$ is the amplitude of *ith* sinusoid, $\omega_i(t)$ the dominant frequency of the *ith* component, $a_i(t)g_i(t)$ the frequency modulation term, Φ_i the constant phase shift and p the number of components in the signal. We can also rewrite eqn (10) in analytical form for reference and test signals, respectively:

$$S_{ref}(t) = \sum_{i=1}^{n} A_i(t) \exp\left(\omega_i(t) + a_i(t) g_i(t) + \Phi_i\right)$$

$$S_{test}(t) = \sum_{i=1}^{n} A_i'(t) \exp\left(\omega_i'(t) + a_i'(t) g_i'(t) + \Phi_i\right) + n(t) \tag{11}$$

where $n(t)$ is the additive noise and the primes indicate distortion in test with respect to reference signal. As a first-order approximation, the distortions may be modeled as additive errors given by

$$z(t) = z(t) + \Delta(t). \tag{12}$$

In the context of speaker-dependent ASR, the main sources of error are variability in the utterance and Lombard effect.[1] The relative modulation function[6] of each component is given by setting $n(t)$ to zero in (11) and taking the ratios of the components:

$$RMF_i(t) = \Delta A_i(t) \exp(\Delta\omega_i(t) + \Delta a_i(t) \Delta g_i(t)). \tag{13}$$

RMF shows how close the test utterance is compared to the reference and tends to zero when the reference and test signals are identical. We can interpret *RMF* as a mismatch in term, i.e. if H is the matched filter extracted from clean speech, the true matched filter of the test utterance is $H + \zeta\,(RMF_i)$. Such mismatches give rise to reduction in the peak output of the matched filter much in the same way as Doppler-shifted signals suffer losses in radar.[5] In the case of man-made signals (as in radar waveform design, techniques are used to contain the losses within acceptable performance limits over the expected Doppler shift. In the case of speech, all the terms in (13) are, in general, non-zero. To alleviate the losses in $y(n)$, we need to either represent $H(n)$ or X_{STFT} in such a way that $y(n)$ is not sensitive to *RMF* found in speech signals. One such obvious choice is to use $|X_{STFT}|$ because the *RMF* gets simplified to $\Delta A_i(t)$ provided the frequency resolution is not very high and the *i*th component of the signal remains in the same frequency bin corresponding to ω_i.

Figure 7 shows the output $Y(n)$ obtained by matched filtering noisy speech and manually time aligning $y_i(n)$ to account for the variability in the speaking rate. This indicates that MF of speech is feasible if we choose TFR that is relatively insensitive to *RMF* and also take care of the issues in time alignment. Our approach to addressing the issue of insensitivity to *RMF* is motivated by the perceptual model of the human auditory system and we use EarLyzer.[7] EarLyzer is a signal-processing algorithm that computes perceptually weighted power spectrum of speech with the following attributes.

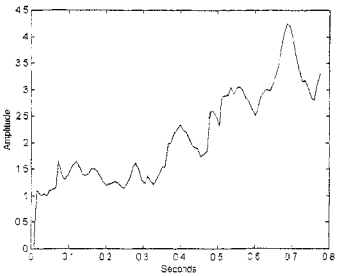


FIG. 7. Manually aligned $Y(n)$ with magnitude processing

(a) Overlapping Mel-spaced critical band integration using DTFT over window lengths matched to the bandwith. The overlap in the frequency reduces the sensitivity of the amplitude spectrum to RMF.

(b) Equal loudness compensation.

(c) Dynamic range compression to approximate intensity to loudness conversion.

In Section 5, we report the results with EarLyzer front-end. To address the time-alignment issues in matched filtering of speech, we present two novel ideas: one in the transform domain and the other based on DTW.

4. Aligned matched filters

4.1. *Transform domain approach*

Consider the feature vector $\mathbf{r}_n$ of size M given by

$$\mathbf{r}_n = (r_{n1}, r_{n2}, \ldots r_{nM})^t \tag{14}$$

where r_{nk} is the energy in the kth filter at time n. We can stack $\mathbf{r}_n$ to form an $M \times N$ matrix:

$$\mathbf{R} = (\mathbf{r}_1, \mathbf{r}_2, \ldots \mathbf{r}_N). \tag{15}$$

Now, consider a single row of $\mathbf{R}$, which is the output of a single filter k for the whole pattern:

$$\mathbf{r}_k = (\mathbf{r}_{1k}, \mathbf{r}_{2k}, \ldots \mathbf{r}_{Mk}). \tag{16}$$

We take the transform of $\mathbf{r}_k$ and select only a region of interest with the window W:

$$\underline{\rho} = F\{\mathbf{r}\}W \tag{17}$$

where $\underline{\rho}$ is the modulation spectrum of the output from the kth channel of the filter bank. The columns of $\mathbf{R}$ correspond to the power output from the filter channels 1... M. We select the region of modulation frequencies in the region 1 to 16 Hz as most of the information useful for ASR lies in this frequency band.[8] We take the transform of all the channels 1... M, and obtain a transformed $M \times L$ matrix $\mathbf{R}_f$. Note that L is independent of the duration of the utterance.

thereby eliminating the need for time alignment. For discrete transforms and efficient computation, L is chosen such that the signal vector $\mathbf{r}_k$ is appended with zeros for the required resolution. Defining

$$\mathbf{R}_{f(MF)} = \mathbf{R}_f \bullet \mathbf{W} \tag{18}$$

where . denotes element by element multiplication and W is a binary-valued matrix obtained by thresholding $\mathbf{R}_f$, i.e.

$$\begin{aligned} w_{ij} &= 1 \text{ if } \mathbf{R}_f(i,j) \geq r_{f,th} \\ &= 0 \text{ otherwise} \end{aligned} \tag{19}$$

Thus, $\mathbf{R}_{f(MF)}$ forms the matched filter in the transform domain. We take the transform of the columns of the test pattern matrix $\mathbf{T}$ to obtain the $M \times L$ matrix $\mathbf{T}_f$. The recognizer output can be formulated as

$$i^* = \arg\max_v \left\{ \mathbf{T}_f^t \mathbf{R}_{f(MF)}^v \right\} \tag{20}$$

where v is the index of the vocabulary words. The surface plots for $\mathbf{R}_{f(MF)}$ and $\mathbf{T}_f$ are shown in Figs 8a and b, respectively, for the utterance 'John Smith'.

4.2. *Matched filtering with DTW (MF-DTW)*

The problem associated with spectral sequence comparison of speech arises from the fact that different acoustic renditions of the same speech utterance are seldom realized at the same speed over the entire utterance. In DTW, we define a dissimilarity measure $d_\Phi(\mathbf{R}, \mathbf{T})$ based on the optimum warping function $\Phi(\Phi_R, \Phi_T)$ as the accumulated distortion over the entire utterances:[1]

$$d_\Phi(\mathbf{R}, \mathbf{T}) = \frac{1}{M_\Phi} \sum_{k=1}^{K} d\big(\Phi_{\mathbf{R}}(k), \Phi_{\mathbf{T}}(k)\big) m(k) \tag{21}$$

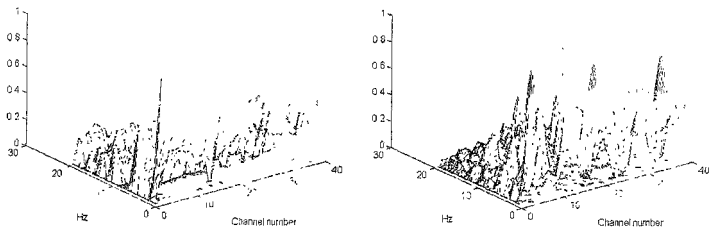
where $m(k)$ is a non-negative path weighting coefficient and M_Φ is a path-normalizing factor; $\mathbf{R}$ and $\mathbf{T}$ are the reference and test spectral sequences of *different* lengths. The goal of DTW is to minimize $d_\Phi(\mathbf{R}, \mathbf{T})$ over all possible paths, subject to some path constraints. It is possible to view DTW matching also as a form of matched filtering as follows. We used a 19-channel EarLyzer with 83-feature vectors as the front-end with DTW for time alignment and pattern matching. The MF-DTW algorithm incorporates the 2D-matched filtering into DTW by replacing each feature vector by a matrix as given below.

$$\mathbf{R}_n = [\cdots \mathbf{r}_{(n-l)}\ \mathbf{r}_{(n-l+1)} \cdots \mathbf{r}_n\ \mathbf{r}_{(n+l-1)}\ \mathbf{r}_{(n+l)} \cdots]^t \tag{22}$$

$$\mathbf{T}_n = [\cdots \mathbf{t}_{(m-l)}\ \mathbf{t}_{(m-l+1)} \cdots \mathbf{t}_m\ \mathbf{t}_{(m+l-1)}\ \mathbf{t}_{(m+l)} \cdots]^t. \tag{23}$$

This may be contrasted with the usual DTW which corresponds to $\ell = 0$; $\ell \neq 0$ corresponds to a feature matrix of consecutive feature vectors which is selected to be a matched filter. The matched filter $\mathbf{R}_n$ is defined by

$$\mathbf{R}_n^{mf} = \mathbf{R}_n \bullet \mathbf{W}_n \tag{24}$$



(a) (b)

FIG. 8a. Transform of R after thresholding, b. transform of T.

where • denotes element by element multiplication and $\mathbf{W}_n$ is a binary-valued matrix obtained by thresholding $\mathbf{R}_n$, i.e.

$$w_n(i, j) = \text{if } \mathbf{R}_n(i, j) \geq r_{threshold}$$
$$= 0 \text{ otherwise.} \tag{25}$$

The thresholding operation on the reference matrix results in high-energy regions to be used as matched filters. For the DTW we use a correlation distance measure, given by $d = 1 - \mathbf{t}'\mathbf{r}$, which effectively is similar to a matched filter, as shown earlier. The DTW also takes care of alignment along the time axis; along the frequency axis, the EarLyzer output is relatively insensitive to the articulation effects of speech.

5. Experimental results

We have evaluated the algorithms presented in Section 4 for a limited vocabulary, speaker-independent task. Moving automobiles is selected as the acoustic environment for assessing the robustness of these algorithms. Reference patterns are obtained from clean speech recorded with the automobile in a parking place and test patterns correspond to noisy speech recorded in a moving vehicle with an average SNR of approximately 0 dB. The vocabulary size is 30, which is typical in speaker-dependent ASR applications. The isolated words are end-pointed manually. The results are summarized below.

Table I
Recognition scores for SD-ASR

Algorithm	Test data size (#words, #speakers)	Correct recognition rate
Transform domain	30 , 1	70.00%
MF-DTW, (ℓ =0)	300 , 10	84.66%
MF-DTW, (ℓ =1)	300 , 10	89.33%
MF-DTW, (ℓ =2)	300 , 10	91.33%
MF-DTW, (ℓ =3)	300 , 10	90.33%

The transform domain algorithm is evaluated with only one speaker as it is considered adequate for a preliminary assessment at the level of 70% correct recognition performance level. The MF-DTW algorithm yielded best recognition performance at $\ell = 2$. The poorer performance at $\ell = 3$ may be attributed to inclusion of non-stationary regions into the local regions obtained by thresholding. We are currently working on an algorithm to adaptively change the value of ℓ based on a measure of stationarity of the reference signal.

6. Conclusions

In this paper, we have formulated a new approach to speech recognition based on matched filtering. This approach holds the promise of robust performance. We have presented the theoretical basis for applying the matched filter approach to speech recognition and demonstrated its efficacy by adding noise to clean speech. The issues in matched filtering of speech in noisy acoustic environment have been discussed and two possible approaches to time alignment are suggested. Preliminary evaluation of these two approaches on limited-vocabulary SD-IWR application, with noisy speech recorded in an automobile environment, shows that the matched filtering formulation is promising. We are currently working on an algorithm to realize the time-ordered matched filters and use automatic time-alignment techniques to gate the outputs of each filter. We are also investigating the efficacy of matched filtering formulation to speaker-independent recognition tasks.

Acknowledgment

This research work is supported by a project sponsored by M/s Ericsson Inc., USA, who also provided the noisy speech database. We also thank our colleagues S. Sunil, M. Anand, and S. Dhiraj for providing the code for DTW and for many discussions during the course of this work.

References

1. RABINER, L. R. AND JUANG, B. H. — *Fundamentals of speech recognition*, Prentice-Hall, 1993.
2. HERMANSKY, H. — Perceptual linear prediction (PLP) analysis of speech, *J. Acoust. Soc. Am.*, 1990, **87**, 1738–1752.
3. PARSONS, T. W. — *Voice and speech processing*, McGraw-Hill, 1987.
4. JAIN, A. K. — *Fundamentals of digital image processing*, Prentice-Hall, 1989.
5. RIHACZEK, A. W. — *Principles of high resolution radar*, McGraw-Hill, 1969, pp. 133–144.
6. FINEBERG, A. B. — The recognition of multicomponent signals; in *Computational methods of signal recovery and recognition* (Mammone, R J., ed.), Wiley, 1992.
7. AVADHANULU, J. V., MATHEW, M. AND SREENIVAS, T. V. — EarLyzer: Perceptually motivated robust TFR of speech, *Proc. EuroSpeech-99*, Budapest, Hungary, Sept. 1999.
8. NOBORU, K., HERMANSKY, H. AND TAKAYUKI, A. — Desired characteristics of modulation spectrum for robust automatic speech recognition, *Proc. ICASSP*, 1998.