A Comparitive Analysis of Channel-Robust Features and Channel Equalization Methods for Speech Recognition

Saeed Vaseghi Ben Milner*

School of Electrical Engineering and Computer Science, Queen's University of Belfast, Belfast, UK.
*British Telecom Research Laboratories.

Abstract

The use of a speech recognition system with telephone channel environments, or different microphones, requires channel equalisation. In speech recognition, the speech models provide a bank of statistical information that can be used in the channel identification and equalisation process. In this paper we consider HMM-based channel equalisation, and present results demonstrating that substantial improvement can be obtained through the equalisation process.
An alternative method is to use a set of features which is more robust to channel distortion. Channel distortions result in an amplitude-tilt of the speech cepstrum, and so differential cepstral features should provide a measure of immunity to channel distortions. In particular the cepstral-time feature matrix, in addition to providing a framework for representing speech dynamics, can be made robust to channel distortions. We present results demonstrating that a major advantage of cepstral-time matrices is their channel insensitive character.

1 Introduction

In this paper we consider the problem of recognition of speech distorted in transmission through a communication channel or by a microphone. The channel distortion is modelled as an unknown convolutional operation. From the talker to the recognition system, three sources of convolutional distortion can be identified; namely the acoustic environment in which the microphone is placed, the microphone, and the communication channel.
In the time domain, the channel output $y(m)$, which is the input to the speech recogniser, is modelled as

$$y(m) = \big(x(m) + n_1(m)\big) * h(m) + n_2(m) \qquad (1)$$

where $x(m)$ is the clean speech, $h(m)$ is the channel response, $n_1(m)$ is the acoustic noise, and $n_2(m)$ is the channel noise. It is assumed that the channel distortion is dominant, and the channel noise and the acoustic noise can be ignored. An ideal channel equaliser, $H^{inv}(f)$, has a frequency response equal to the inverse of the channel. Real communication channels may not be invertible, hence $H^{inv}(f)$, may not be well defined for all frequencies. Such a situation occurs in channels with bands of frequencies in which the signals are heavily attenuated and immersed in noise. A second form of non-invertible channel occurs when the channel is non-minimum phase. However, speech recognition systems use features derived from the power spectrum, and therefore the channel magnitude response and not the channel phase is of primary interest.
In speech recognition systems where the speech features are based on log-spectra, such as mel-frequency cepstral coefficients (MFCCs), the convolutional distortion becomes an additive distortion as

$$\boldsymbol{y}(m) = \boldsymbol{x}(m) + \boldsymbol{h} \qquad (2)$$

Where $\boldsymbol{y}(m)$, $\boldsymbol{x}(m)$ and $\boldsymbol{h}$ are the channel-distorted speech, the channel input speech and the channel cepstral vectors respectively. Note that in the logarithmic domain the effect of a channel distortion is the addition of a tilt to the channel input signal cepstrum.
There are two broad approaches to robust speech recognition. One approach attempts to use an equaliser to suppress the distortion $\boldsymbol{h}$. The second approach uses speech features which are inherently robust to channel distortion. Mokbel *etal* (1993) developed a method for on-line adaptation of a speech recognition system to variations in telephone line conditions. A similar approach is also described in Wittmann (1993). Hermansky and Morgan (1992) proposed the use of bandpass filters on the time-variations of the log spectral speech bands. Hanson and Applebaum (1993) compare the effect of bandpass filtering of log spectral bands with a highpass filtering.
In the remainder of this paper Bayesian equalisation based on hidden Markov models, and channel-robust features are considered.

2 Bayesian Equalisation Based on HMMs

This section considers blind equalisation in applications where the statistics of the channel input can be modelled by a set of hidden Markov models as in recognition of speech distorted by a communication channel or a microphone figure(1). In speech recognition, it is assumed that the channel inputs are acoustic realisations of words selected from a vocabulary of size V. HMM-based equalisation can be stated as follows : Given a channel output sequence $\boldsymbol{Y}$ and that the channel input is drawn from a set of V HMMs $\mathcal{M}=\{\mathcal{M}_i\, i=1, ..., V\}$ estimate the channel response and the input.
The likelihood of an HMM $\mathcal{M}_i$ and a sequence of channel input vectors $\boldsymbol{X} = [\boldsymbol{x}(0), ..., \boldsymbol{x}(N-1)]$ can be expressed as

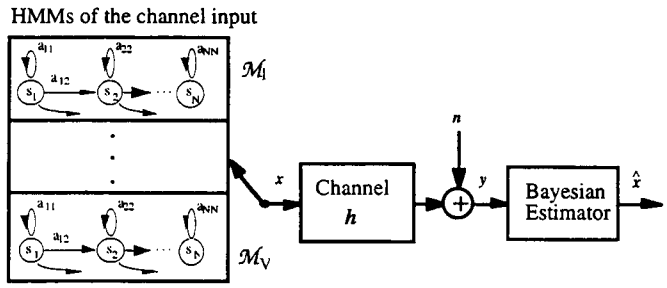


Figure 1 Illustration of a channel with the input alphabet modelled by a set of HMMs.

$$f_{X|\mathcal{M}}(X|\mathcal{M}_i) = \sum_{s} f_{X|\mathcal{M},s}(X|\mathcal{M}_i,s)P_{s|\mathcal{M}}(s|\mathcal{M}_i) \quad (3)$$

where $f_{X|\mathcal{M}S}(X|\mathcal{M}_i,s)$ is the likelihood that the sequence X was generated by the state sequence s of model $\mathcal{M}_i$, and $P_{S|\mathcal{M}}(s|\mathcal{M}_i)$ is the Markovian pmf of the state sequence s. The state observation is modelled by a Gaussian pdf as

$$f_{X|\mathcal{M},S}(x|\mathcal{M}_i,s) = \frac{1}{(2\pi)^{P/2}|\Sigma_{x,s}|^{1/2}}\exp\left(-\frac{1}{2}(x-\mu_{x,s})^T\Sigma_{x,s}^{-1}(x-\mu_{x,s})\right) \quad (4)$$

where $\mu_{x,s}$ and $\Sigma_{x,s}$ are the mean vector and the covariance matrix of the Gaussian pdf of the state s. For a given model $\mathcal{M}_i$ and state sequence $S = \{s(0), s(1), ..., s(N-1)\}$, the pdf of a sequence of N independent observation vectors $Y = \{y(0), y(1), ..., y(N-1)\}$ is

$$f_{Y|\mathcal{M},S,H}(Y|\mathcal{M}_i,s,h) = \prod_{m=0}^{N-1} f_{X|\mathcal{M}}(y(m)-h|\mathcal{M}_i,s(m))$$
$$= \prod_{m=0}^{N-1}\frac{1}{(2\pi)^{P/2}|\Sigma_{x,s(m)}|^{1/2}}\exp\left(-\frac{1}{2}(y(m)-h-\mu_{x,s(m)})^T\Sigma_{x,s(m)}^{-1}(y(m)-h-\mu_{x,s(m)})\right) \quad (5)$$

From Eq. (5) the maximum likelihood estimate of h is

$$\hat{h}^{ML}(Y,s) = \sum_{m=0}^{N-1}\left(\sum_{k=0}^{N-1}\Sigma_{xx,s(k)}^{-1}\right)^{-1}\Sigma_{xx,s(m)}^{-1}\left(y(m)-\mu_{x,s(m)}\right) \quad (6)$$

2.1 MAP Channel Estimate Based on HMMs

Given a sequence Y of N P-dimensional vectors, the a posteriori pdf of the channel h along a state sequence s of an HMM $\mathcal{M}_i$, is defined as

$$f_{H|Y,S,\mathcal{M}}(h|Y,s,\mathcal{M}_i) = \frac{1}{f_Y(Y)}f_{Y|H,S,\mathcal{M}}(Y|h,s,\mathcal{M}_i)f_H(h)$$
$$= \frac{1}{f_Y(Y)}\prod_{m=0}^{N-1}\frac{1}{(2\pi)^P|\Sigma_{xx,s(m)}|^{1/2}|\Sigma_{hh}|^{1/2}}\exp\left(-\frac{1}{2}(y(m)-h-\mu_{x,s(m)})^T\Sigma_{xx,s(m)}^{-1}(y(m)-h-\mu_{x,s(m)})\right)\times$$
$$\exp\left(-\frac{1}{2}(h-\mu_h)^T\Sigma_{hh}^{-1}(h-\mu_h)\right) \quad (7)$$

where it is assumed that each state is Gaussian with a mean vector $\mu_{x,s}(m)$ and a covariance matrix $\Sigma_{xx,s}(m)$, and that the channel h is also Gaussian with a mean vector μ_h and a covariance matrix Σ_{hh}. The MAP estimate along state s, can be obtained as

$$\hat{h}^{MAP}(Y,s,\mathcal{M}_i) = \sum_{m=0}^{N-1}\left(\sum_{k=0}^{N-1}\left(\Sigma_{xx,s(k)}^{-1}+\Sigma_{hh}^{-1}\right)\right)^{-1}\Sigma_{xx,s(m)}^{-1}\left(y(m)-\mu_{x,s(m)}\right)+$$
$$\left(\sum_{k=0}^{N-1}\left(\Sigma_{xx,s(k)}^{-1}+\Sigma_{hh}^{-1}\right)\right)^{-1}\Sigma_{hh}^{-1}\mu_h \quad (8)$$

The MAP estimate over all HMMs is given by

$$\hat{h}(Y) = \sum_{i=1}^{V}\sum_{S}\hat{h}^{MAP}(Y,s,\mathcal{M}_i)P_{S|\mathcal{M}}(s|\mathcal{M}_i)P_{\mathcal{M}}(\mathcal{M}_i) \quad (9)$$

2.2 Use of Statistical Averages Over All HMMs

A simple approach to blind equalisation, is to use the average mean vector μ_x and the covariance matrix Σ_{xx}, taken over all the states of all the HMMs. The ML estimate of the channel, $\hat{h}^{ML}$, is defined as

$$\hat{h}^{ML} = (\bar{y}-\mu_x) \quad (10)$$

where $\bar{y}$ is the time-averaged channel output. The channel input estimate is

$$\hat{x}_t = y_t - \hat{h}^{ML} \quad (11)$$

The MAP channel estimate becomes

$$\hat{h}^{MAP}(Y) = \sum_{m=0}^{N-1}(\Sigma_{xx}^{-1}+\Sigma_{hh}^{-1})^{-1}\Sigma_{xx}^{-1}(y(m)-\mu_x)+(\Sigma_{xx}^{-1}+\Sigma_{hh}^{-1})^{-1}\Sigma_{hh}^{-1}\mu_h \quad (12)$$

2.3 Hypothesised-Input HMM

For each HMM in the input vocabulary, a channel estimate is obtained and used to equalise the channel output. Thus a channel estimate $\hat{h}_w$ is based on the hypothesis that the input word is w. It is expected that a good channel estimate is obtained from the correctly hypothesised HMM, and a poorer estimate from an incorrectly hypothesised HMM. The hypothesised-input HMM algorithm is as follows :

For i =1 to number of words V {

step-1 Using HMM, $\mathcal{M}_i$, estimate the channel, $\hat{h}_i$,

step-2 Using, $\hat{h}_i$, estimate the input as $\hat{x}(\hat{h}_i) = y - \hat{h}_i$

step-3 Compute a score for model $\mathcal{M}_i$, given $\hat{x}(\hat{h}_i)$. }

Step-4 Select the most probable word.

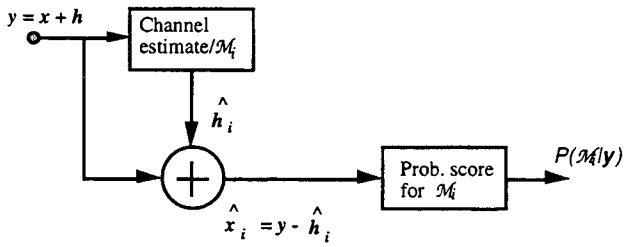


Figure 2 Hypothesised channel estimation procedure.

3 Cepstral-Time as Channel-Robust Features.

A cepstral-time matrix, $c_t(n,m)$, may be obtained from a 2-D DCT of a log spectral-time matrix, $X_t(f,k)$. [Milner 1994]. In transformation from a spectral-time to a cepstral-time matrix, via a 2-D DCT, the frequency axis f of the spectral-time matrix is converted to quefrency n and the time axis, k is converted to frequency, m. The lower index coefficients along the axis n represent the spectral envelope, whereas the higher coefficients represent the pitch and the excitation. Along the axis, m, the lower coefficients represent the long time variation of the cepstral coefficients, and the higher coefficients the short time variation. Figure 3 illustrates these regions.

The cepstral-time matrix contains information regarding the transitional dynamics of the speech. The zeroth column contains the time-averaged value of the cepstral coefficient. The effects of channel distortion is concentrated in this column and the exclusion of this column provides a channel robust feature set.

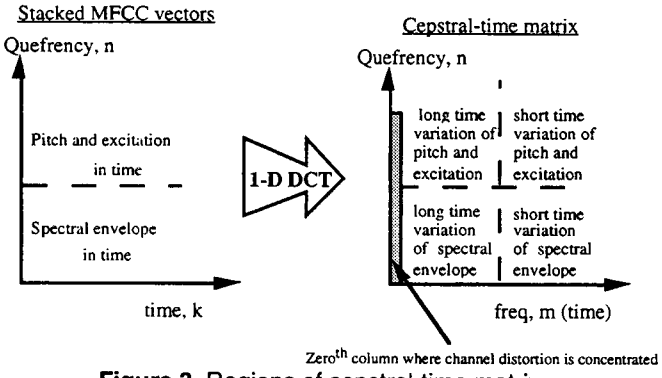


Figure 3 Regions of cepstral-time matrix.

4 Experimental Results

The experiments were performed using speech distorted by simulated channel responses. Six channel distortion, shown in Figure 4, were chosen for experimentation. The filters were designed specifically to attenuate parts of the frequency spectrum where much of the speech energy is found (up to 3kHz). The filters a to e are all invertible, where as filter f is non-invertible. Filter f is designed to simulate a bandpass telephone channel, with sharp band limiting between 300Hz and 3000Hz. The HMMs were trained on clean speech. The speech features were 15 dimensional mel-frequency cepstral coefficients (MFCCs), including the coefficient $c(0)$. The speech database used was the NOISEX database of noisy speaker-independent isolated spoken English digits [VARGA 1992].

4.1 Channel Equalisation

Table-1 shows experimental results for the channel distortions, a to f, and two equalisation methods. The first is based on ML estimation. The second technique uses the hypothesised maximum likelihood channel estimation technique of Section 2.3 denoted as *ML_HMM*. To compare performance, the row labelled *NCC*, shows the case where no channel compensation has been applied.

The channel distortions 4.c to 4.e severely attenuate the speech spectrum, and subsequently cause deterioration in recognition performance, as shown by *NCC*. The maximum likelihood channel estimate, *ML*, using an average of all HMM means, produces good improvement for all channel distortions. However, the maximum likelihood estimate based on the hypothesised-HMM approach gives high accuracy for all the channel distortions tested. The results of the maximum likelihood estimates in table-1 are reasonably constant for all the channel distortions.

Figure 5 shows the log filter bank representation of two channel distortion filters, namely 4.d and 4.e. Also shown are the two estimates, based on the maximum likelihood, averaged over all HMMs *ML*, and the maximum likelihood estimate computed using the most likely HMM *ML_HMM*. The figure shows that these techniques both produce good estimates of the channel distortion. The *ML_HMM* produces a slightly better estimate of the channel than the *ML* approach.

	Flat	*a*	*b*	*c*	*d*	*e*	*f*
NCC	100	95	50	10	10	10	25
ML	100	94	94	94	94	93	89
ML_HMM	100	100	100	100	100	100	100

Table 1 % Recognition accuracy of blind deconvolution.

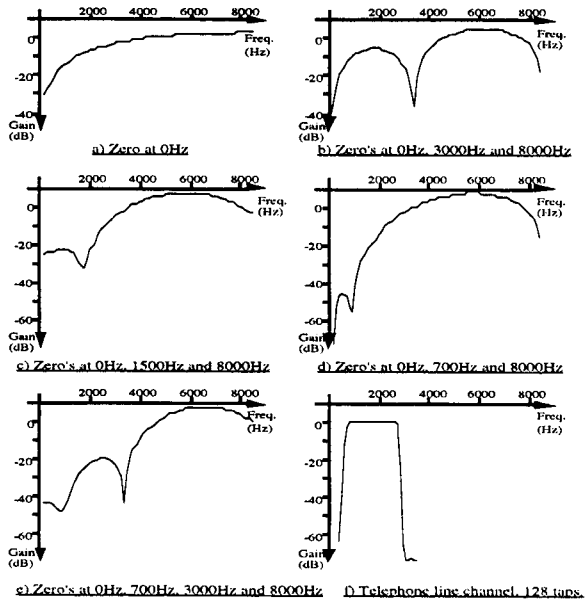


Figure 4 Synthesised channel distortions.

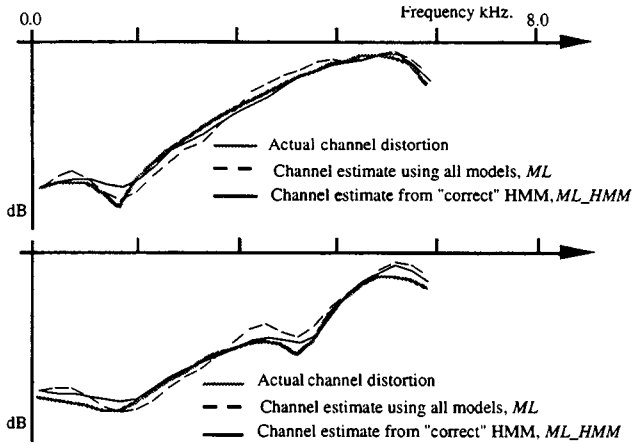


Figure 5 Illustration of actual and estimated channel response for two channels.

4.2 Channel Robust Features

Table-2 shows the recognition performance for the 6 channel distortions, *a* to *f*, and for no channel distortion, *flat* response. The results show the performance of the 15x4, truncated cepstral-time matrix, and the 14x3 truncated cepstral-time matrix. Additionally the performance of 15 dimensional cepstral vectors is also shown. The speech data base in this experiment is NOISEX.

Table-2 shows that the 14x3 cepstral-time matrix remains unaffected by the invertible channel distortions, 4.a to 4.e, in comparison to the non-truncated 15x4 cepstral-time matrix which suffers degradation as a result of the channel. However the 14x3 cepstral-time matrix has not been completely robust to the non-invertible channel distortion, 4.f, although it does outperform the 15x4 cepstral-time matrix.

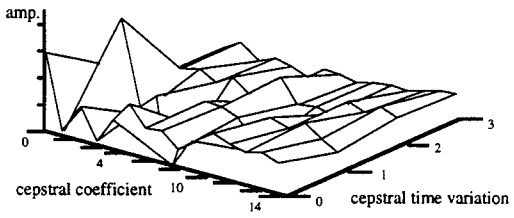
Figure 6 shows a plot of two 15x4 cepstral-time matrices. In Figure 6-a the cepstral-time matrix has been obtained from a distortion-free speech signal, and in Figure 6-b the same piece of speech has been corrupted by channel 4.e. It can be seen that the first column of the cepstral-time matrix has been contaminated by the channel, but the inner matrix has remained unaffected by the channel distortion.

	Flat	*a*	*b*	*c*	*d*	*e*	*f*
15-dim cep. vector	100	95	50	10	10	10	25
15x4 C-T matrix	100	91	74	45	10	12	58
14x3 C-T matrix	100	100	100	100	100	100	90

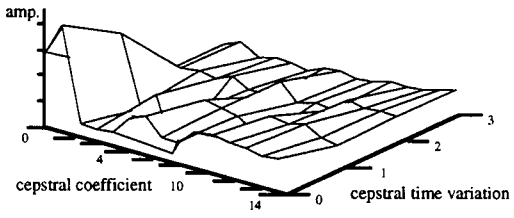
Table 2 % Recognition accuracy of cepstral-time matrices.

5 Conclusion.

Two approaches for the recognition of channel-distorted speech have been investigated. The first approach is based on blind deconvolution given the channel output signal and HMMs of the channel input signal. Maximum likelihood channel equalisation offer significant improvement in recognition performance. In particular a hypothesised-input deconvolution method, based on using each HMM to provide a different channel estimate, improves recognition accuracy considerably. The second part of this paper investigated the use of features which are robust to channel distortion. It has been shown that the cepstral-time matrix, with the first column omitted, is robust to channel distortions.



a) 15x4 cepstral-time matrix of un-distorted speech



b) 15x4 cepstral-time matrix of distorted speech

Figure 6 Cepstral-time matrix of a distorted and undistorted speech signal.

References

HANSON B.A., APPLEBAUM T. H. (1993), "Subband or Cepstral Domain Filtering for Recognition of Lombard and Channel-Distorted Speech," IEEE Int. Conf. Acoustics, Speech and Signal Processing Vol. II, Pages 79-82.

HERMANSKY H, MORGAN N (1992), "Towards Handling the Acoustic Environment in Spoken Language Processing", Int. Conf. on Spoken Language Processing Tu.fPM.1.1, Pages 85-88.

MILNER B. (1994), Speech Recognition in Adverse Environment, University of East Anglia, UK.

MOKBEL C., MONNE J. JOUVET D. (1993), "On-Line Adaptation of A Speech Recogniser to Variations in Telephone Line Conditions", Proc. 3rd European Conf. On Speech Communication and Technoplogy, EuroSpeech-93, Vol 2, Pages 1247-1250.

VARGA A. P., STEENKEN H. J. M., TOMLINSON M., JONES D.(1992), "The NOISEX-92 Study on the Effect of Additive Noise on Automatic Speech Recogintion", Tech. Report., DRA Speech Research Unit.

VASEGHI S. V., (1996), "Advanced Signal Processing and Digital Noise Reduction", John Wiley.

WITTMANN M., SCHMIDBAUER O. , AKTAS A., "Online channel compensation for robust speech recognition", EUROSPEECH-93, pp 1251-1254, 1993.