ROBUST SPEECH RECOGNITION VIA MODELING SPECTRAL COEFFICIENTS WITH HMM'S WITH COMPLEX GAUSSIAN COMPONENTS

William J.J. Roberts[a] *and Sadaoki Furui*[b]

[a]Defence Science Technology Organisation,
Information Technology Division,
PO Box 1500, Salisbury, South Australia 5108.
bill.roberts@dsto.defence.gov.au

[b]Department of Computer Science,
Tokyo Institute of Technology,
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552, Japan.
furui@cs.titech.ac.jp

Abstract

Robust speech recognition via hidden Markov modeling of *spectral* vectors is studied in this paper. The hidden Markov model (HMM) mixture components are assumed complex Gaussian with zero mean, diagonal covariance, and with incorporating an unknown scalar gain term. The gain term is associated with each spectral vector and it models the varying energy of speech signals. It is estimated by applying the maximum likelihood (ML) criterion. On an isolated digit database, in clean conditions, the spectral modeling with ML gain estimation approach achieved similar performance to cepstral modeling of speech.

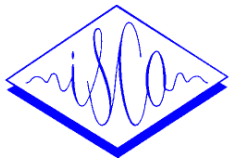
Two additive noise compensation approaches for the spectral modeling scheme are also considered. The first approach requires a full noise HMM. This HMM is combined with the clean speech HMM to yield a noisy speech HMM. The second approach requires only the spectral shape of the noise. A term dependent on the spectral shape, together with an unknown magnitude term, is incorporated into the clean speech HMM to yield a noisy speech HMM. The unknown magnitude of the noise is estimated via the ML criterion. The performance of these two approaches for isolated digit recognition in noise is demonstrated and compared to a robust cepstral modeling approach from the literature.

I. Introduction

As the effects of additive noise on cepstral coefficients are difficult to quantify, speech recognition in noise using this approach often yields poor performance. In this work we model spectral components rather than cepstral components. This has the advantage that the effects of noise on the speech spectral components may be quantified in some cases. For noise additive to and independent of the speech waveform, the spectral magnitude of the noisy speech is given by the addition of the spectral magnitude of the clean speech with that of the noise. The modeling of spectral components for speech recognition has a long history. Prior to the introduction of cepstral coefficients it was the dominant approach for speech recognition, see e.g. [1]. However we know of no previous speech recognition techniques based upon explicitly modeling spectral coefficients by HMM's with complex Gaussian probability density functions (pdfs), an intuitively appealing choice for the pdf and one justifiable under appropriate conditions [2]. Modeling spectral coefficients using complex Gaussians, without their incorporation within the HMM, has been performed previously for other applications. In [2] minimum mean squared error estimates of clean speech spectral components are obtained by modeling the noisy speech components with complex Gaussians pdfs. In [3] the same model is used to derive second order statistics of cepstral coefficients.

For speech recognition purposes, only the magnitude of the spectral components needs to be kept as the likelihood of zero mean complex Gaussian random variables depends only on the magnitude of the variables, and not on their phase. The approach is thus mathematically similar to the time-domain modeling using circulant covariances presented in [4]. The spectral vectors may be obtained using any appropriate spectral estimation techniques. In this study we used a smoothed periodogram spectral estimator. We found it desirable to use a subband of the spectral magnitude components. The choice of the subband was made on heuristic grounds and its optimal choice is outside the scope of this paper. Reference [5] investigates subband based speech recognition in greater detail.

In modeling spectral vectors we need to revisit two problems that are readily addressed in cepstral based systems: compensation for the unknown gain of speech signals and compensation for unknown channels. In cepstral based systems, the first problem may be addressed via exclusion of the zeroth cepstral coefficient and the second problem via, for example, cepstral mean subtraction. In this work we model the gain by incorporating an unknown scalar term in the mixture component's pdf. The scalar gain is then estimated by applying the ML criterion. A closed form solution results for the gain in the clean case and an expectation-maximization (EM) based iterative solution is required in the noisy case. The problem of compensation for unknown communication channels is not addressed in this work but it should be noted that this problem does not arise in all applications, e.g. it does not arise when



6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

the communication channel remains constant during testing and training.

Considering a single mixture component for a particular state of the HMM, the gain compensated pdf of a K-dimensional complex spectral vector Y_t of unknown scalar gain g_t is given by

$$p(Y_t|g_t) = \frac{1}{\pi^K g_t^{2K} \prod_{k=1}^K \lambda(k)} \exp\left(-\sum_{k=1}^K \frac{|Y_t(k)|^2}{g_t^2 \lambda(k)}\right) \quad (1)$$

where $\lambda(k)$ is the variance of the kth speech spectral component. This pdf is derived by considering a complex Gaussian spectral vector that has been multiplied by the scalar g_t to produce Y_t. Note that the notation in (1) does not identify the particular state and mixture component. We continue to use this simplified notation and present the mathematics in terms of complex Gaussian pdf's only rather than presenting the mathematics of the HMM in its full generality. This is because estimation of HMM initial state probabilities, mixture weights, state transition probabilities, and state sequences may be performed in the standard manner using Baum or Viterbi decoding.

In the next section we present the algorithms for training the model based upon (1) in clean conditions. In section III we present equations for recognition in clean and noisy environments. In section IV we discuss the implementation and results on an isolated digit experiment.

II. Training

Training the HMM involves estimation of the $\{\lambda(k)\}$ for each state and mixture from the T vectors that have been assigned to that state and mixture. Given the state and mixture component, the spectral vectors are, as is usually done, assumed to be independent. Thus the pdf of the T vectors is given by $\prod_{t=1}^T p(Y_t|g_t)$. The end-product of the training procedure is the ML estimate of $\{\lambda(k)\}_{k=1}^K$. This estimate requires estimates of the gains $g_t, t = 1, \ldots, T$, and estimates of the gains can only be obtained given an estimate of $\{\lambda(k)\}$. Thus training consists of alternate $\{\lambda(k)\}$ and gain estimation. The estimates of g_t represent the gains of the individual spectral vectors rather than estimates of the underlying speech model and they are discarded once the estimation of $\{\lambda(k)\}$ has been accomplished. By taking the logarithm of the pdf of the T vectors that have been assigned to a particular state and mixture component, and setting the derivative with respect to g_t to zero, the ML estimate of g_t^2, for a given $\{\lambda(k)\}$ is found to be

$$g_t^2 = \frac{1}{K} \sum_{k=1}^K \frac{|Y_t(k)|^2}{\lambda(k)}. \quad (2)$$

For given values of $g_t, t = 1, \ldots, T$, the ML estimate of the $\{\lambda(k)\}$ may be obtained in a similar manner and is given by

$$\lambda(k) = \frac{1}{T} \sum_{t=1}^T \frac{|Y_t(k)|^2}{g_t^2}. \quad (3)$$

Thus for each state and mixture of the HMM, an iterative training process is required which yields the estimates of the $\{\lambda(k)\}_{k=1}^K$. These iterations are ceased once convergence of the estimates has occurred.

III. Recognition

Recognition involves calculating the likelihood for each spectral vector for each HMM Gaussian mixture component. In both clean and noisy cases the ML gain estimate is substituted into the likelihood equation. In the clean case, the ML gain estimate is available in closed form and likelihood calculation is thus a simple non iterative procedure. In the noisy case, the ML gain estimation must be performed iteratively. We emphasize again that the expression given in terms of single Gaussian components may be readily embedded in either Baum or Viterbi style decoding.

A. Recognition in clean conditions

In clean conditions, we must calculate a likelihood of the form in (1) for each spectral vector Y_t for each Gaussian mixture component. In contrast to the training case, the $\{\lambda(k)\}$ are now all known. We use the ML estimate of g_t given by (2) in (1). Thus the likelihood calculation, in clean conditions, is not iterative. Substitution of the ML gain estimate (2) into the mixture state pdf (1), yields the following expression for the log likelihood in clean conditions for a particular state and mixture component within the HMM

$$P(Y_t) = \frac{K^K \exp(-K)}{\pi^K \left(\sum_{k=1}^K \frac{|Y_t(k)|^2}{\lambda(k)}\right) \prod_{k=1}^K \lambda(k)}. \quad (4)$$

Generally the log likelihood is used and (4) may be simplified by dropping constant terms.

For recognition in noise we model the noise spectral components in that same manner as the speech spectral components, i.e. as uncorrelated complex Gaussian random variables. We consider two noise compensation schemes. The first requires a trained HMM for the noise, where as the second requires only the gross spectral shape of the noise.

B. Recognition in additive noise given a noise HMM

Here we consider the additive noise to be an HMM whose parameters have been obtained from training data. The model for the noisy speech spectral components is obtained by the convolution of the noise HMM with the clean speech HMM. The result is also an a HMM [6] with a number of states equal to the product of the number of states of the two HMM's. The number of mixtures per state of the noisy HMM is equal to the product of the number of mixtures per state of the two HMM's [6]. For the noisy spectral components Z_t, each Gaussian mixture components of the noisy speech HMM is thus of the form

$$p(Z_t|g_t) = \frac{1}{\pi^K \prod_{k=1}^K (g_t^2 \lambda(k) + \sigma^2(k))} \cdot \exp\left(-\frac{|Z_t(k)|^2}{g_t^2 \lambda(k) + \sigma^2(k)}\right) \quad (5)$$

where $\{\sigma(k)^2\}$ represent spectral energies from a particular state and mixture of the noise HMM. Unlike in the clean case, there is no closed form expression for the maximizing gain of (5). We now develop an iterative solution

using the EM algorithm. In the EM approach we calculate a new estimate $\hat{g}_t$ given the observed data Z_t and the old estimate g'_t. This is done using the following standard expression of the auxiliary function of the EM:

$$\begin{aligned}\hat{g}_t &= \arg\max_{g_t} E\{\log p(Z_t, Y_t|g_t)|Z_t, g'_t\} \\ &= \arg\max_{g_t} E\{\log\left[p(Z_t|Y_t)p(Y_t|g_t)\right]|Z_t, g'_t\} \\ &= \arg\max_{g_t} E\{\log p(Y_t|g_t)|Z_t, g'_t\}. \end{aligned} \quad (6)$$

Substituting the expression for $p(Y_t|g_t)$ from (1)

$$\begin{aligned}\hat{g}_t &= \arg\max_{g_t} \Big\{-K\log(g_t^2) \\ &\quad - \frac{1}{g_t}\sum_{k=1}^{K}\frac{E\{|Y_t(k)|^2|Z_t, g'_t)\}}{\lambda(k)}\Big\} \end{aligned} \quad (7)$$

Differentiating and setting the result to zero yields

$$\begin{aligned}\hat{g}_t &= \frac{1}{K}\sum_{k=1}^{K}\frac{E\{|Y_t(k)|^2|Z_t(k), g'_t\}}{\lambda(k)} \\ &= \frac{1}{K}\sum_{k=1}^{K}\frac{\sigma^2(k)W_t(k) + W_t(k)^2|Z_t(k)|^2}{\lambda(k)} \end{aligned} \quad (8)$$

where $W_t(k) = g_t'^2\lambda(k)/(g_t'^2\lambda(k) + \sigma^2(k))$. Equation (8) was obtained using the complex normal ($\mathcal{CN}$) conditional distribution $p(Y_t(k)|Z_t(k), g_t) \sim \mathcal{CN}[W_t(k)Z_t(k), \sigma^2(k)W_t(k)]$. A similar equation was derived in [6] for noisy auto-regressive gain estimation. Once the gain term has been estimated via (8) it is used in (5) to calculate the likelihood for each state and mixture component.

C. Recognition given the additive noise spectral shape

For the second noisy recognition scheme, we require knowledge of only the overall spectral shape of the noise, e.g. white, low-pass etc, rather than a full noise HMM. We assume that the noisy speech is modeled by the clean speech HMM with the addition of a noise variance term to the spectral variances associated with each state and mixture component. This noise term is given by the known noise spectral shape multiplied by an unknown noise gain term. Thus the pdf for a particular state and mixture of the the noisy speech HMM is of the following form

$$\begin{aligned}p(Z_t|n_t, g_t) &= \frac{1}{\pi^K\prod_{k=1}^{K}g_t^2\lambda(k) + n_tN(k)} \\ &\quad \cdot\exp\left(-\frac{|Z_t(k)|^2}{g_t^2\lambda(k) + n_tN(k)}\right) \end{aligned} \quad (9)$$

where n_t is the unknown noise gain and the $\{N(k)\}$ represent the known spectral shape of the noise. Both the noise and speech gains are now estimated by the EM. The EM iterations for the speech gain are performed using (8) where the $\sigma^2(k)$ terms in that equation are replaced by $n_tN(k)$. The n_t term is estimated using

$$\begin{aligned}\hat{n}_t &= \arg\max_{n_t} E\{\log p(Z_t, Y_t|g_t)|Z_t, n'_t\} \\ &= \arg\max_{n_t} E\{\log\left[p(Z_t|Y_t)p(Y_t|g_t)\right]|Z_t, n'_t\} \\ &= \arg\max_{n_t} E\{\log p(Z_t|Y_t)|Z_t, n'_t\} \end{aligned} \quad (10)$$

where n'_t is the current estimate of the noise energy. The distribution $p(Z_t|Y_t)$ is given by

$$\begin{aligned}p(Z_t|Y_t) &= \frac{1}{\pi^K n_t^K\prod_{k=1}^{K}N(k)} \\ &\quad \cdot\exp\left(-\sum_{k=1}^{K}\frac{|Z_t(k) - Y_t(k)|^2}{n_tN(k)}\right) \end{aligned} \quad (11)$$

Thus the $\hat{n}_t$ term is calculated by the following

$$\begin{aligned}\hat{n}_t &= \frac{1}{\sum_{k=1}^{K}N(k)}\sum_{k=1}^{K}\frac{1}{N(k)}\left(|Z_t(k)|^2(1 - 2W_t(k))\right. \\ &\quad \left.+\sigma_t^2W_t(k) + W_t(k)^2|Z_t(k)|^2\right) \end{aligned} \quad (12)$$

where $W_t(k) = g_t^2\lambda(k)/(g_t^2\lambda(k) + n'_tN(k))$. Estimation of the g_t and n_t proceeds in turn for each spectral vector for each mixture component density. The estimation is halted once the difference in likelihood between successive iterations is sufficiently small.

In this approach the noise gain is estimated *on-line* and this estimate is free to track a noise energy that is evolving with time. This could be an important feature in some applications.

IV. Implementation and Results

The performance of the spectral model was tested in clean and noisy conditions on an isolated digit recognition using the TI digits corpus. Full details of the corpus are available in [7]. We used the portion of the corpus consisting of two utterances per digit from 112 male speakers. The training set consisted of the two utterances from 66 of these speakers and the testing data consisted of the two utterances from the remaining 56 speakers. The speech signal, originally sampled at 20kHz, was down-sampled to 8kHz.

All experiments were conducted using left to right (LR) HMM's. Each digit HMM had 10 states and each state consisted of 2 complex Gaussian components. The vector dimension was $K = 100$. The initial estimate of parameter set was obtained from uniform segmentation of the training set into 10 intervals.

Pre-processing the speech signal consisted of estimating the spectral coefficients using the smoothed periodogram described in [3]. We did not keep all the spectral components form the spectral estimate. Rather we kept only those spectral components corresponding to the center frequencies of the MEL band, as given in [8]. This produced slightly superior performance to keeping all frequency components but we do not investigate this matter further in this work.

The techniques was tested using additive white Gaussian noise although the techniques may be applied to more complicated noise models. The use of white Gaussian noise allows comparison with noise robust cepstral based techniques from the literature that also used the TI corpus.

We tested both of the recognition in noise techniques. In the first method, an estimate of the noise variance was obtained from silence portions of the test utterance. In the second method the noise variance is estimated in a

Technique	SNR (dB)							
	Inf	30	25	20	15	10	5	0
Known noise HMM	98.6	98.4	98.1	97.4	96.2	94.7	92.6	89.9
Known noise spectral shape	98.4	97.5	96.5	94.6	91.1	85.6	-	-
Robust Cepstral [3]	98.8	98.0	97.2	96.4	94.4	88.9	-	-

TABLE I
Performance of systems in clean and noisy environments

ML manner for each speech vector thus a previous noise estimate energy estimate is not required. We did however assume a flat power spectral density for this method. As mentioned earlier this allows this approach to adapt to changing noise conditions although this feature is not being tested in this experiment.

The database and all of the experimental conditions outlined above except for the vector length K are the same as those used to test the robust cepstral based system presented in [3] and we compare our results to those presented in [3]. In [3], a spectral estimation scheme involving frames of length 256 together with the frames either side, was employed. Frame overlapping was also used.

For clean signals, the spectral modeling approach provided comparative performance to the cepstral technique of [3]. In noise, the spectral modeling technique using a noise HMM produced the best performance. Using an explicit noise model produced better performance than estimation the noise model on-line. Clearly the former technique is critically dependent on the quality of the noise HMM. As the noise for this experiment was relatively simple and easily estimated it would be expected that the noise model would be of good quality. In more difficult situations, or when it is impossible to estimate the full statistics of the noise model, the on-line estimation of the noise parameters would be expected to be a better technique. The robust cepstral model performed better than the on-line noise estimation scheme. However the former technique is only applicable to white noise and the noise energy must be estimate from silence portions. Thus it will not track evolving noise powers.

In the first technique using a noise HMM we found that the number of iterations required for convergence of the gain estimate varied with the SNR ratio. At 30dB an average of 5 iterations were required, at 10 dB an average of 30 iterations were required for effective gain estimation. This was also true for the second technique but the number of iterations required was approximately 5 times those required for the first technique. Intuitively extra iterations are required as more parameters are being estimated. However there are potentially fewer states in this noisy signal model.

V. Conclusion

We have demonstrated robust speech recognition using spectral vectors rather than using cepstral vectors as in common use. The approach was tested only on an isolated digit problem. On this task we compared the performance to a robust cepstral technique. The test was on a isolated digits and using only additive white Gaussian noise. The technique may be applied to noise modeled by a HMM or to noise whose spectral shape is known. In the latter case the system will estimate the energy of the noise in an on-line manner allowing it to track time evolving noise energy. The experiment in this paper did not test this feature. Possible extensions to this work could involve the incorporation of a constant multiplicative term in the model to handle channel conditions and testing on more difficult data and noise.

References

[1] B. A. Dautrich, L. R. Rabiner, and T. B. Martin, "On the effects of varying filter bank parameters on isolated word recognition," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 31, pp. 793–806, 1983.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short time spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[3] Y. Ephraim and M Rahim, "On second-order statistics and linear estimation of cepstral coefficients," *submitted for publication*, vol. 7, no. 2, pp. 162–176, Mar. 1999.

[4] W. J.J. Roberts and Y. Ephraim, "Robust speech recognition using HMM's with Toeplitz state covariance matrices," *Speech Communication*, vol. 31, pp. 1–14, 2000.

[5] H. Bourland and S. Dupont, "Subband-based speech recognition," in *Conference Proc. IEEE ICASSP*, 1997, pp. 1251–1245.

[6] Y. Ephraim, "Gain-adapted hidden Markov models for recognition of clean and noisy speech," *IEEE Trans. on Speech Processing*, vol. 40, no. 6, pp. 1303–1316, June 1992.

[7] R. G. Leonard, "A database for speaker-independent digit recognition," in *Conference Proc. IEEE ICASSP*, 1984, pp. 42.11.1–42.11.4.

[8] J. W. Picone, "Signal modeling techniques in speech recognition," *Proceedings of the IEEE*, vol. 81, no. 9, pp. 1215–1247, Sept. 1993.