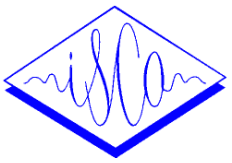

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

MAXIMUM LIKELIHOOD NOISE HMM ESTIMATION IN MODEL-BASED ROBUST SPEECH RECOGNITION

Martin Graciarena

Instituto de Ingenieria Biomedica, Facultad de Ingenieria - UBA,
Paseo Colon 850 piso 5 (CP 1063), Buenos Aires, Argentina
mgracia@fi.uba.ar

ABSTRACT

This paper presents a generalization of Rose's Integrated Parametric Model to the gaussian mixture hidden Markov model (HMM), formulation. Observations from clean speech HMM and noise HMM models are combined in the log spectra domain, through a corruption function, to generate noisy speech observations. In order to recognize noisy speech with the proposed model, when only the clean speech HMM and noisy speech adaptation data are available, a maximum likelihood (ML) estimation algorithm for the noise HMM parameters is provided. This algorithm uses the "max" approximation as the corruption function. Noisy digit recognition experiments, with NOISEX-92, show that the same performance is achieved between the proposed model using either a noise model calculated from silent sections of several utterances or the estimated noise model from a single noisy utterance.

1. INTRODUCTION

The performance of speech recognition systems in real world applications may suffer a severe degradation in unknown noisy situations. The source of this performance degradation is an environmental mismatch between training and testing conditions. The goal of robust speech recognition systems is to reduce this mismatch in order to bring back the performance to matched conditions.

The mismatch reduction is done in model-based techniques by the construction of a noisy speech model, from a clean speech model and a noise model, for a particular environment. Among the most important references of this group are Gales' Parallel Model Combination (PMC) [2], Rose *et. al.* Integrated Parametric Model (IPM) [1] and Logan's work [3]. Rose *et. al.* IPM model combines observations, from speech and noise gaussian mixture models, through a corruption function. If a speaker is modeled as a gaussian mixture, the formulae for ML estimation of the speaker model parameters were presented and applied to the problem of speaker recognition in noise, with an available noise model.

Most model-based techniques assume that the noise model is available *a priori* from a particular environment. The noise model is estimated from an available noise signal or from silent parts of noisy speech. However it is desirable that the estimation is carried out only with noisy speech adaptation data. Logan's work is the first to propose an estimation technique for the noise model. The main extension presented in that work, which is based in Rose *et. al.* work, forms ML estimates of the unknown autoregressive hidden Markov noise model parameters using the whole noisy speech utterance. It is an extension of a previous enhancement system in order to make it adaptive

In this paper, first a novel noisy speech model is introduced, which is a generalization of Rose's Integrated Parametric Model, from gaussian mixture models to the gaussian mixture HMM formulation. Observations from clean speech HMM and noise HMM are combined, through a corruption function in the log spectrum domain, to generate noisy speech observations. The noise corruption function, for additive noise in the linear spectrum domain with log spectrum domain observations, is the logarithm of the sum of the exponential of speech and noise observations. The proposed generalization enables the modeling of non-stationary signals, such as speech, in noisy (possibly non-stationary) environments. Therefore, it extends the application of the IPM model from noisy speaker recognition to noisy speech recognition.

Also in this paper a ML estimation algorithm for the gaussian mixture noise HMM parameters, is provided within the framework of the proposed noisy speech model. For parameter estimation, in order to produce closed form expressions, the "max" approximation is used as the corruption function. The adaptation data can be provided either in supervised mode (with transcriptions) or unsupervised mode (without transcriptions).

2. HMM-IPM MODEL

The goal of this section is to present an extension of the IPM model to the HMM formulation, which will be called HMM-IPM. The HMM-IPM model is a multidimensional model, where for each one of the original *M* clean speech HMM model states at time *t*, *N* new noise model states are added. The total number of states of the HMM-IPM is *MxN*. A state in the HMM-IPM model, at each time t can be accessed only from the previous state, as opposed to the IPM model where it can be accessed from any other state. The probability of accessing a certain state i_t, j_t from the previous state i_{t-1}, j_{t-1} is determined by $p(i_t, j_t / i_{t-1}, j_{t-1}) = T_{i_t, j_t}$. Therefore the state transition process is a Markov process. A diagram of this model is presented in figure 1.

The clean speech signal model l_s viewed as a generative process, goes through a sequence of hidden states $I = (i_1, i_2, ..., i_T)$ where $i_t = 1, \ldots, M$, and this sequence of states turnes into an independent sequence of D-dimensional signal vectors $X = (x_1, x_2, ..., x_T)$, through a set of state dependent continuous pdfs $b_{i_t}(x_t)$. Similarly the noise signal source l_b goes through a sequence of hidden states $J = (j_1, j_2, ..., j_T)$,

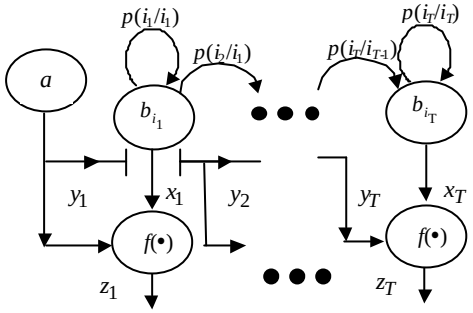

Figure 1: The HMM-IPM model for one state noise model, where $j_t = 1; \forall t$

where j_t =1, ..., *N*, generates an independent sequence of *D*-dimensional signal vectors $Y = (y_1, y_2, ..., y_T)$, through a set of state dependent continuous pdf $a_{j_t}(y_t)$. Both the sequence of states and the sequence of signal vectors are assumed independent in time. Additionally, the processes *X* and *Y* are assumed independent. The densities $b_i(x_t)$ and $a_j(y_t)$ are assumed here to be Gaussian with diagonal covariance matrices,

All the following derivation will be done for one gaussian per state HMM. The extension to the multi-gaussian per state HMM case is straightforward.

The likelihood of generating the sequence of observations *Z* given the speech l_s and noise l_b models, *P*(*Z*) is computed as follows

$$\begin{aligned} P(Z) &= \sum_I \sum_J P(Z, I, J) \\ &= \sum_I \sum_J P(Z / I, J) P(I, J) \end{aligned} \tag{1}$$

where the first term in the second line of (1) is the observations probability given a state sequence in the signal noise state space, and the second term is the state sequence probability. The first term can be calculated according to the independence hypothesis, as follows

$$\begin{aligned} P(Z / I, J) &= \prod_{t=1}^{T} p(z_t / i_t, j_t) \\ &= \prod_{t=1}^{T} \iint_{C_t} b_{i_t}(x_t) a_{j_t}(y_t) dx_t dy_t \end{aligned} \tag{2}$$

Were in (2) the observed signal density for each observation z_t is formed from a general function of speech and background $f(x_t, y_t)$ where C_t denotes the contour defined by $z_t = f(x_t, y_t)$. The contour chosen is for the "max" approximation, which assumes that a noisy observation is formed as the maximum of the signal and background observations. This contour is presented in figure 2.

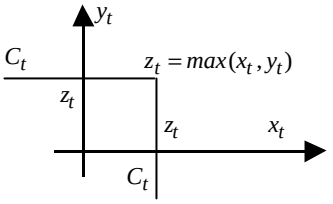

Figure 2: Integration contour C_t for the "max" approximation in equation (2).

The second term of the second line in (1) can be calculated according to the markovian hypothesis as follows

$$P(I, J) = \prod_{t=1}^{T} T_{i_t, j_t} \tag{3}$$

If we define the likelihood of the pairs *X, Y, I,* and *J* to be

$$P(X, Y, I, J) = \prod_{t=1}^{T} b_{i_t}(x_t) a_{j_t}(y_t) T_{i_t, j_t} \tag{4}$$

then (4) can be expressed as

$$P(Z / l_s, l_b) = \sum_I \sum_J \iint_C P(X, Y, I, J) dX dY \tag{5}$$

In the previous equation, the double summation is over all possible length T state sequences through the signal background state space lattice.

3. NOISE MODEL PARAMETER ESTIMATION

3.1. Auxiliary Q function

The goal of this section is to find the equations for maximum likelihood parameter estimates of the noise HMM model.

It is not possible to obtain the ML estimates directly. However, it is possible to iteratively improve on an initial model l_b, and find a new model $\bar{l}_b$ such that $P(Z / \bar{l}_b) \geq P(Z / l_b)$. Baum *et al* [4] showed that the desired improvement in likelihood could be obtained by finding a new model $\bar{l}_b$ that maximizes an auxiliary function

$$\begin{aligned} Q(\bar{l}_b, l_b) &= \mathrm{E}\{P(X, Y, I, J / \bar{l}_b)\} \\ &= \sum_I \sum_J \iint_C P(X, Y, I, J / l_b) \log\{P(X, Y, I, J / \bar{l}_b)\} dX dY \end{aligned} \tag{6}$$

Strictly speaking in the previous equation the reference should be to both models l_s and l_b. If we define $g_t(k, l, I, J)$ as the probability of being in state (*k,l)* at time *t,* and where $h_t(k, l, I, J)$ is the counting function, as follows

$$g_t(k,l)=\sum_I\sum_J h_t(k,l,I,J)P(X,Y,I,J/\overline{l}_b)$$
$$h_t(k,l,I,J)=\begin{cases}1 & \text{if } i_t=k,\ j_t=l\\ 0 & \text{otherwise}\end{cases} \quad (7)$$

From (4) and (7), equation (6) can be expressed as

$$Q(\overline{l}_b,l_b)=\sum_{t=1}^{T}\sum_{k=1}^{M}\sum_{l=1}^{N}\iint_C \log\{b_k(x_t)\overline{a}_l(y_t)\times T_{i_t=k,j_t=l,\overline{l}_b}\}g_t(k,l)dXdY \quad (8)$$

3.2. General Expressions for Noise Model Parameters

Individually maximizing the expectation of $Q(\overline{l}_b,l_b)$ in (8) with respect to each of the noisy density parameters is straightforward. The same hypothesis applies from the IPM model, therefore we achieve the similar expressions as in [1] for the noisy mean $\overline{m}_j$ and variance $\overline{s}_j^2$.

$$\overline{m}_j=\frac{\sum_{t=1}^{T}\sum_{i=1}^{M}\iint_C g_t(i,j)y_t dXdY}{\sum_{t=1}^{T}\sum_{i=1}^{M}\iint_C g_t(i,j)dXdY} \quad (9)$$

$$\overline{s}_j^2=\frac{\sum_{t=1}^{T}\sum_{i=1}^{M}\iint_C g_t(i,j)(y_t-m_j)^2 dXdY}{\sum_{t=1}^{T}\sum_{i=1}^{M}\iint_C g_t(i,j)dXdY} \quad (10)$$

Therefore, we need to achieve closed form expressions for the noisy density parameters.

If we define $a_t(k,l)$ as the probability of the partial observation sequence $Z_1^t=\{z_1,z_2,...,z_t\}$, such that at time *t*, the i[th] state is in *k* and the j[th] state is in *l*. Similarly $b_t(k,l)$ is the probability of the partial observation sequence $Z_{t+1}^T=\{z_{t+1},z_{t+2},...,z_T\}$ given that the i[th] state is in *k* and the j[th] state is in *l*. Therefore,

$$a_t(k,l)=P(Z_1^t,i_t=k,j_t=l)$$
$$b_t(k,l)=P(Z_{t+1}^T/i_t=k,j_t=l) \quad (11)$$

It can be shown for the denominator of (9) and (10) that by developing $g_t(i,j)$ according to the previous definition of $a_t(k,l)$ and $b_t(k,l)$,

$$\iint_C g_t(k,l)dXdY=a_t(k,l)b_t(k,l) \quad (12)$$

For $a_t(k,l)$ a standard recursion can be found

$$a_t(k,l)=\sum_{i_{t-1}}\sum_{i_{t-1}} a_{t-1}(i_{t-1},j_{t-1})T_{i_t=k,j_t=l}\,p(z_t/i_t=k,j_t=l) \quad (13)$$

If we define $K_t(k,l)$ to be,

$$K_t(k,l)=b_t(k,l)\{\sum_{i_{t-1}}\sum_{i_{t-1}} a_{t-1}(i_{t-1},j_{t-1})T_{i_t=k,j_t=l}\} \quad (14)$$

Therefore, from (13) and (14) we can rewrite (12) as

$$\iint_C g_t(k,l)dXdY=K_t(k,l)\,p(z_t/i_t=k,j_t=l) \quad (15)$$

For the numerator of (9), from (13) and (14) we can deduce:

$$\iint_C y_t g_t(k,l)dXdY=K_t(k,l)\iint_{C_t} y_t\,p(x_t,y_t/i_t=k,j_t=l)dx_t dy_t \quad (16)$$

Multiplying the last expression by $p(z_t/i_t=k,j_t=l)$ in the numerator and denominator, results in

$$\iint_C y_t g_t(k,l)dXdY=$$
$$=K_t(k,l)\,p(z_t/i_t=k,j_t=l)E\{y_t/z_t,i_t=k,j_t=l\} \quad (17)$$

Therefore (9) can be transformed from (15) and (17) in

$$\overline{m}_j=\frac{\sum_{t=1}^{T}\sum_{i=1}^{M}K_t(k,l)\,p(z_t/i_t=k,j_t=l)E\{y_t/z_t,i_t=k,j_t=l\}}{\sum_{t=1}^{T}\sum_{i=1}^{M}K_t(k,l)p(z_t/i_t=k,j_t=l)} \quad (18)$$

For (10) a similar expression can be found

$$\overline{s}_j^2=\frac{\sum_{t=1}^{T}\sum_{i=1}^{M}K_t(k,l)\,p(z_t/i_t=k,j_t=l)N_t(k,l)}{\sum_{t=1}^{T}\sum_{i=1}^{M}K_t(k,l)p(z_t/i_t=k,j_t=l)}$$
$$N_t(k,l)=E\{(y_t-m_j)^2/z_t,i_t=k,j_t=l\} \quad (19)$$

3.3. Close Form Expressions

We need to find close form expressions for the noise model parameter ML estimation equations. From (18) and (19) we need to find full expressions for $p(z_t/i_t=k,j_t=l)$, $E\{y_t/z_t,i_t=k,j_t=l\}$, and for $E\{(y_t-m_j)^2/z_t,i_t=k,j_t=l\}$. Solving equation (2) along the contour defined in figure 2 for the "max" approximation results in :

$$p(z_t/i_t=k,j_t=l)=a_l(z_t)B_k(z_t)+b_k(z_t)A_l(z_t) \quad (20)$$

and

$$E\{y_t/z_t,i_t=k,j_t=l\}=$$
$$\frac{z_t a_l(z_t)B_k(z_t)+b_k(z_t)A_l(z_t)E\{y_t/y_t<z_t,i_t=k,j_t=l\}}{p(z_t/i_t=k,j_t=l)} \quad (21)$$

which includes the definition of a truncated gaussian. Its expression is presented next,

$$E\{y_t / y_t < z_t, i_t = k, j_t = l\} = u_j - s_j^2 a_l(z_t) / A_l(z_t) \quad (22)$$

And for equation (19),

$$E\{(y_t - m_j)^2 / z_t, i_t = k, j_t = l\} =$$
$$= \frac{(y_t - m_j)^2 a_l(z_t) B_k(z_t) + b_k(z_t) A_l(z_t) M_t(k,l)}{p(z_t / i_t = k, j_t = l)} \quad (23)$$
$$M_t(k,l) = E\{(y_t - m_j)^2 / y_t < z_t, i_t = k, j_t = l\}$$

were the expression of the truncated gaussian is,

$$E\{(y_t - m_j)^2 / y_t < z_t, i_t = k, j_t = l\} =$$
$$= s_j^2 - s_j^2 (z_t - u_j) a_l(z_t) / A_l(z_t) \quad (24)$$

4. EXPERIMENTAL RESULTS

The proposed model and the estimation algorithm were tested on a speaker independent Spanish digit recognition task. Noise data was extracted from the NOISEX–92 database. Only stationary additive noise was considered. Speech was sampled at 16 kHz and 24 MEL log filter bank coefficients were extracted from Hamming windowed 25 ms frames at a rate of 10 ms. The clean speech models were 10 state HMM with 10-gaussian mixtures per state, for each of the 10 digits and 1 state HMM with 10-gaussian mixtures for the silence model. The noise model was a HMM with one state and 1 gaussian. The training database was composed of 1007 utterances, containing each 10 digits, and the recognition database was composed of 100 utterances, again each with 10 digits. The training database contained speech from 103 speakers of Caribbean and Latin American origin, and the recognition database contained speech from 10 independent speakers, of the same origin.

Recognition results for noisy speech, corrupted with white noise at different SNRs, for the clean and the HMM-IPM model are given in Table 1. Results are in "%accuracy" counting deletions, insertions and substitutions. The grammar used was a parallel model of all the digits and the silence model. First the results for the clean speech model results are presented. Second, results for the HMM-IPM model with a "noise only" model are presented. This "noise only" model was estimated from noisy speech of 50 utterances extracted from pause alignments of the original waveform. Additionally the results for the HMM-IPM model using the estimated noise model are presented. The noise model estimation was done with one sentence, in supervised mode. The recognition was done with the HMM-IPM model using the estimated noise model. The estimation and recognition steps were repeated four times; each time using a sentence from a different speaker. Recognition results averaged from these four results are presented in Table 1.

It is shown in Table 1 that the clean speech model degrades its performance in noisy speech as the SNR decreases. The HMM-IPM model using the "noise only" model increases the recognition of the clean speech model, particularly at 30 dB where the recognition result is still acceptable. At lower SNRs there is still an increase in recognition, however smaller. The HMM-IPM model using the estimated noise model achieves the same recognition as the HMM-IPM model using the "noise only" model for SNRs from infinity to 30dB. A small increase in recognition is achieved for lower SNRs. The main difference found between the estimated noise model and the "noise only" model for these conditions is an increase in the variances in the estimated noise model.

Model / SNR (dB)	Clean speech model	HMM-IPM: "noise only" model	HMM-IPM: estimated noise model
∞dB	92.60	92.70	92.60
30dB	72.00	90.70	90.70
15dB	31.70	43.80	45.75
10dB	14.90	24.80	34.10
5dB	12.30	13.70	17.55

Table 1: Recognition Results in White Noise Corrupted Speech for Clean Speech Model, HMM-IPM Model with "Noise only" model and HMM-IPM Model with Estimated Noise Model (average from 4 different results).

5. CONCLUSION

An extension of the IPM model is presented that enables the modeling of noisy speech. A ML algorithm is presented for the estimation of the noise model HMM parameters for the recognition of noisy speech when only the clean speech HMM and noisy speech adaptation data are available. Noisy digit recognition experiments show an increase in recognition compared to the clean speech model and similar performance to using a noise model calculated from silent sections.

6. ACKNOLEDGEMENT

The author would like to acknowledge the Speech Research and Technology Laboratory at SRI International for the use of the speech database.

7. BIBLIOGRAPHY

1 R. C. Rose, E. M. Hofsetter, and D. A. Reynolds, "Integrated Models for Signal and Background with Application to Speaker Identification in Noise", *IEEE Transactions on Speech and Audio Processing*, Vol. 2, pp. 245-257, April. 1994.

2 M. J. F. Gales, *Model-based techniques for noise robust speech recognition*. PhD. Thesis, Gonville and Caius College, Cambridge University, Sept. 1995.

3 B. T. Logan, *Adaptive Model-Based Speech Enhancement*, PhD. Thesis, Girton College, Cambridge University, 1998.

4 L. E. Baum, T. Petrie, G. Soules and N. Weiss, "A Maximization Technique Occuring In The Statistical Analysis Of Probabilistic Functions Of Markov Chains", *Ann. Math. Statist.*, vol. 41, pp. 164-171, 1970.