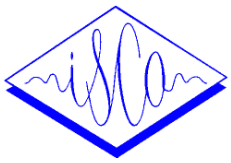


3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

A FREQUENCY-WEIGHTED CONTINUOUS DENSITY HMM FOR NOISY SPEECH RECOGNITION

Hiroshi Matsumoto and Hiroyuki Imose

Faculity of Engineering, Shinshu University
vspace12mm 500 Wakasato, Nagano-shi, Nagano 380, Japan

ABSTRACT

This paper presents a frequency-weighted Hidden Markov Model (HMM) for noisy speech recognition. In this HMM, the covariance matrices of Gaussian probability density functions are fixed to the inverse of frequency weithing matrices in order to utilize the robustness of quefrency weighted cepstrum and also to incorporate their relative perceptual importance in frequency domain into HMM. Two types of frequency weighting functions and the scaling methods of frequency weighting matrices are examined Using NOISEX-92 data base. As a result of ten digit word recognition tests, the 0.3 to 0.5th power of the smoothed power spectrum derived from each mean vector with a normalization factor are found to give the most robust HMM. A comparative experiments showed that the frequency-weighted HMM attained SNR gains of 12 dB, 6 dB, and 3 dB, 2 dB over a standard diagonal HMM for white, pink, car, and Linx noises. Furthermore, it was found that a duration control is important in the frequency-weighted HMM.

1. INTRODUCTION

Hidden Markov Modelings (HMM) have been successfully applied to a variety of speech recognition systems. While HMMs can effectively deal with the statistical variations in the training data, the performance inevitably degrades under mismatch conditions between training and testing, especially in noisy environment. In order to cope with this difficulty, it is important to make HMM itself robust to interfering noise, in addition to suppress noise components prior to recognition phase. Robust HMMs proposed so far have been based on a sort of adaptation technique. These include cepstral compensation [1],[2],[3], HMM composition [4], and Parallel Model Compensation [5]. Instead of the adaptation approach, this paper concerns with perceptual basis and robust parameters.

In DTW-based speech recognition, robust parameters such as quefrency weighted cepstral coefficients have been used in distance measures such as RPS [6] and SGD [7], and also frequency-weighting techniques has been applied to incorporate human auditory characteristics into distance measures such as WLR [8] and WGD [9]. In HMM-based speech recognition, however, the liftering has no effect on robustness, since the probability density function of the liftered cepstral coefficients is not different from that of the original cepstral coefficients except a multiplicative constant. Furthermore, since the relative contribution of spectral parameters to likelihood scores depends only on the variances in the training samples, the frequency-weighting technique have never been incorporated into HMM.

In order to utilize the frequency-weighting technique and the robustness of the liftered cepstrum in HMM, this paper proposes a frequency-weighted HMM of which covariance matrices are fixed to the inverse of frequency-weighting matrices. The fallowing sections describe the two types of frequency-weighting functions and the two scaling methods for covariance matrices. After examining the optimum weighting parameters, we compare the performance of the frequency-weighted HMM and conventional HMMs using the NOISEX-92 speech data base.

2. Frequency-Weighted Continuous Density HMM

2.1 Frequency-weighted HMM

In this study, we use the first p terms of quefrency-weighted cepstral coefficients as the components of an observation vector:

$$\boldsymbol{x} = \left[c_1, 2c_2, \cdots, pc_p\right]^T. \tag{1}$$

These parameters associate with the frequency derivative of log power spectrum, or equivalently smoothed group delay spectrum,

$$T(\lambda) = 2\sum_{n=1}^{p} nC_n \cos(n\lambda). \tag{2}$$

These spectrum are robust to global spectral variation due to wideband additive noise or channel distortion. Therefore, the Euclidean distance of these parameters, for instance, RPS, is well known as a robust distance measure.

First, we define a frequency-weighted Euclidean distance using smoothed group delay spectra sampled at the frequencies,

$$\lambda_i = \frac{2\pi i}{2p+1}, (i = 0, 1, \cdots, p), \tag{3}$$

as fallow:

$$d(f,g) = \sum_{i=-p}^{p} \left\{T_g(\lambda_i) - T_f(\lambda_i)\right\}^2 W(\lambda_i), \tag{4}$$

where $W(\lambda_i)$ is a frequency-weighting function, and f and g signify reference and test patterns respectively. This distance measure is also written in a matrix form:

$$d(f,g) = (\boldsymbol{T}_g - \boldsymbol{T}_f)^T \boldsymbol{W} (\boldsymbol{T}_g - \boldsymbol{T}_f) \tag{5}$$

where the matrix $\boldsymbol{W}$ and the vector $\boldsymbol{T}$ are,

$$\boldsymbol{W} = diag\big[W(\lambda_0), 2W(\lambda_1), \cdots, 2W(\lambda_p)\big] \quad (6)$$

and

$$\boldsymbol{T} = \big[T(\lambda_0), T(\lambda_1), \cdots, T(\lambda_p)\big]^T. \quad (7)$$

Furthermore, this distance measure in equation (5) is also rewritten using the vector $\boldsymbol{x}$ as

$$d(f,g) = (\boldsymbol{x}_g - \boldsymbol{x}_f)^T \boldsymbol{U}^{-1} (\boldsymbol{x}_g - \boldsymbol{x}_f) \quad (8)$$

$$\boldsymbol{U}^{-1} = \boldsymbol{C}^T \boldsymbol{W} \boldsymbol{C}. \quad (9)$$

where $\boldsymbol{C}$ is the cosine transform matrix whose ijth element is given by

$$c_{ij} = 2\cos\frac{2\pi ij}{2p+1}. \quad (10)$$

Thus, the weighting Matrix $\boldsymbol{U}$ in the frequency-weighted Euclidean distance (8) corresponds to the covariance matrix in the Maharanobis distance except their magnitude. On the basis of this correspondence, we propose a novel continuous density HMM in which the probability of the observation $\boldsymbol{x}$ emitted from state i is given by

$$P_i(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^p |\alpha_i \boldsymbol{U}_i|}} \times \exp\Big\{-\frac{(\boldsymbol{x}-\boldsymbol{\mu}_i)^T \boldsymbol{U}_i^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_i)}{2\alpha_i}\Big\} \quad (11)$$

where $\boldsymbol{U}_i^{-1}$ is the frequency-weighting matrix given by equation (9), and the α_i is a scale factor to adjust the overall dispersion. For simplicity a single Gaussian model is assumed here. Therefore, the covariances in this model are not estimated statistically, but instead are derived from the weighting function $W(\lambda)$ based on a prior knowledge on the perceptual importance in frequency domain and/or expected variance due to degradation of speech. In probability calculation, the computational efficiency extremely decreases by using the feature vector $\boldsymbol{T}$ instead of $\boldsymbol{x}$.

In conventional HMMs, since the probability density function of a linearly transformed Gaussian variable $\boldsymbol{Ax}$ is equal to the product of the original probability density function $P(\boldsymbol{x})$ and the determinat of $\boldsymbol{A}$, the quefrency-weighting has no effect on HMM in training and testing. On the other hand, in the frequency-weighted HMM, the quefrency-weighting is expected to make the HMMs robust to noise as in the RPS distance measure.

2.2 Frequency-weighting function

The fallowing tow types of weighting functions were examined.

1. Mean-vector-dependent weighting function
   This weighting function takes account of the spectral power corresponding to the mean vector of each Gaussian probability density function. The weighting function for $N(\boldsymbol{\mu}, \boldsymbol{\Sigma}_i)$ is given by the βth power of smoothed power spectrum :

   $$W(\lambda_i) = \exp\big\{\beta\, l_i\big\} \quad (12)$$

   where the log-power spectrum $\{l_i\}$ is derived from the truncated mean vector itself in terms of a Fourier cosine transform:

   $$\big[l_0, l_1, \cdots, l_p\big]^T = \boldsymbol{C}\big[\frac{\mu_1}{1}, \frac{\mu_2}{2}, \cdots, \frac{\mu_q}{q}, 0, \cdots, 0\big]^T. \quad (13)$$

   The parameter β, which will be called a compression factor, controls the range of weight, and the truncation order $q(\leq p)$ adjusts the smoothness of frequency characteristics.

2. Fixed weighting function
   In order to suppress the variation in high frequency region, the fallowing low- frequency-weighting function is applied to all the Gaussian probability density functions of each model:

   $$W(\lambda_i) = \big|1 + a\exp(j\lambda_i)\big|^2 \quad (14)$$

   This function does not take formants into account. The frequency characteristics of weighting function is adjusted by the coefficient $a(0 < a \leq 1)$.

The optimum values of the above parameters will be experimentally examined later.

2.3 Scaling of weighting matrix

The frequency weighting-matrix reflects only the relative contribution on the frequency domain expected by a prior knowledge. Therefore, to convert it to a covariance matrix, it is required to scale the matrix. The fallowing two methods are examined.

1. Maximum Likelihood Scaling
   The scaling factor α_i for the ith state is given by

   $$\alpha_i = \frac{-p}{p}\mathrm{Trace}(\boldsymbol{\Sigma}_i \boldsymbol{U}_i^{-1}) \quad (15)$$

   where $\boldsymbol{\Sigma}_i$ is the sample covariance matrix in the state.

2. Normalization Scaling
   The scale factor α_i for the ith state is calculated by

   $$\alpha_i = \sum_{n=0}^{p} W(\lambda_n) \quad (16)$$

   This method equalizes the sum of frequency-weighting coefficients among states, but does not take statistical variations into account.

2.4 Training procedure

The frequency-weighted HMM is trained by the fallowing procedure:

1. Make an initial HMM based on the Baum's algorithm using a set of training data.
2. Calculate the matrix $\boldsymbol{U}_i$ for each state from equation(9) and then replace the covariance matrices $\boldsymbol{\Sigma}_i$ of the initial model with $\alpha_i \boldsymbol{U}_i$.
3. With the fixed $\boldsymbol{U}_i$, reestimate the mean vector $\boldsymbol{\mu}_i$ and transition probabilities $\{a_{ij}\}$ and α's in the maximum likelihood scaling by Baum's algorithm.
4. Stop if it converges, otherwise go to the step (3).

3. EVALUATION

3.1 Data base and speech analysis

The subsequent experiments use a subset of NOISEX-92, i.e., twenty repetitions of isolated 10 English digits uttered by a male speakers.

The speech signals are digitized at a sampling frequency of 16 kHz. The linear predictive auto-correlation (LPC) method with 26 poles was applied to the speech data with a Hamming window of 25 ms, a frame shift

of 10 ms, and a first-order adaptive preemphasis of $(1 - az^{-1})$. The first 16 mel-cepstral coefficients are calculated using Oppenheim's recursion with the coefficient of 0.41 in bilinear transform.

Four stationary back ground noises were used: car and Linx helicopter noises and white and pink noises generated in a computer. The pink noise has the frequency characteristics of $1/(1 - 0.9z^{-1})$, i.e., spectral slope of -6dB/oct. The degraded speech by car and Linx noises was used from NOISEX-92. On the other hand, the white and pink noises were added to clean speech in the fallowing manner. The noise level for each word was adjusted so that the global signal to noise ratio (SNR) was equal to a predetermined value. In calculating the average power for each word, the word length was defined as the total number of frames whose power was larger than a threshold of -40 dB below the peak power of all the frames of the input word.

Each word was represented by a left-to-right HMM with 8 to 26 states. The HMM models were trained using 10 repetitions of noise-free samples. Another set of ten utterances was used for testing. The Vitarbi algorithm was used for testing. The beginning and end points are fixed to those in the label files. Thus, only substitution errors were scored.

3.2 Effects of weighting characteristics

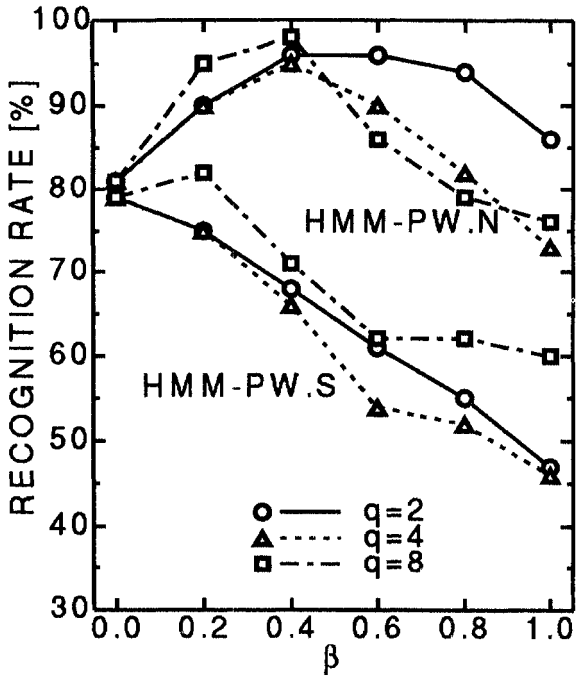
First, the effects of both the smoothing order q and the compression factor β in a mean-vector-dependent weighting function as well as the scaling methods were examined. Fig. 1 (a) and (b) show the results for the pink noise of 6 dB and car noises of 0 dB. It is clear that the normalized scaling method (.N) is much better than the maximum likelihood scaling (.S). This less effectiveness of the maximum likelihood scaling seems to be caused by low estimation accuracy of the covariance matrix in equation (15) due to insufficient amount of training data. The normalized scaling method with β of .3 to .5 significantly improves recognition accuracy. The lager q tends to slightly improves the scores but to reduce the effective range of β. The optimum values of β and q slightly variy depending on noises and SNRs.

Second, the effect of the fixed-weighting functions was examined under the same conditions as in the above experiments. The recognition scores as a function of the coefficient a saturated beyond 0.6. The effect of the scaling methods is similar to that in the mean-vector-dependent weighting function. In the pink noise condition, the recognition performance is improved from 81% without frequency-weighting to 90% with $a = 1.0$, but, in car noise condition, the improvement is very little. The best score obtained by the fixed weighting function is inferior to that by the mean-vector-dependent weighting function.

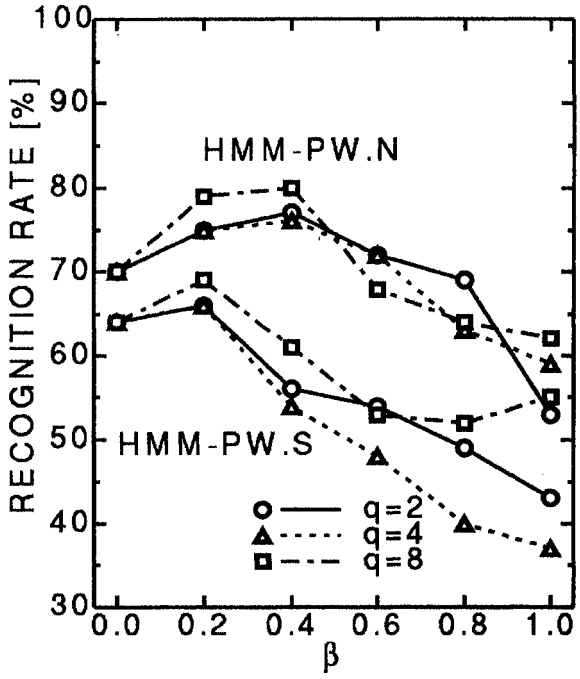
Consequently, using the normalization scaling and the optimum values of β and q, the mean-vector-dependent weighting function attained the recognition gains of 18, 10% over no weighting cases (i.e., $\beta = 0$) for the pink and car noise conditions, respectively.

3.3 Comparison with conventional HMMs

The performance of the frequency-weighted HMM was compared with two standard diagonal HMMs: (1) with individually trained variance for each state (HMM), and (2) with a fixed variance equal to the 'grand variance' over all the training speech (HMM-GT) [10]. The values of β and q are fixed to the globally optimum values for each noise. Table 1 shows the recognition scores for the HMM and the HMM-GT and the frequency-weighted HMM (HMM-PW.N) at various SNRs for white, pink, car, and Linx noises. The recognition scores for the HMMs with fixed variances, i.e., HMM-PW.N and HMM-GT, were significantly improved as compared with the standard HMM at low SNR conditions. Although this seems to be partly caused by the insufficient amount of trining data, this is mostly contributed by the quefrency-weighted cepstral coefficients as a robust parameter due to the fixed covariance. Furthermore, the frequency-



(a) Pink noise at 6 dB SNR.



(b) Car noise at 0 dB SNR.

Fig.1 Effects of the ccmpression factor β, q, and the scaling methods in frequency-weighting.

Table 1 Comparison of HMM, HMM-GT, and HMM-PW.N in various noise conditions

(a) White noise

SNR[dB]	0	6	12	18	24
HMM	10%	20%	36%	69%	97%
HMM-GT	37%	61%	80%	100%	100%
HMM-PW.N	42%	82%	100%	100%	100%

(b) pink noise

SNR[dB]	-6	0	6	12	18
HMM	11%	20%	57%	96%	100%
HMM-GT	28%	63%	81%	99%	100%
HMM-PW.N	32%	80%	98%	100%	100%

(c) Car noise

SNR[dB]	-6	0	6	12	18
HMM	24%	43%	94%	100%	100%
HMM-GT	27%	65%	98%	100%	100%
HMM-PW.N	45%	80%	98%	100%	100%

(d) Linx noise

SNR[dB]	-6	0	6	12	18
HMM	18%	34%	59%	94	100%
HMM-GT	26%	50%	84%	98	100%
HMM-PW.N	25%	61%	90%	100	100%

weighted HMM achieved a 10 to 18 % higher recognition accuracy for various noses of 6 dB SNR than the HMM-GT.

3.4 Effects of the number of states and adaptive preemphasis

In the above experiments, the number of states has been set to 26 and an adaptive preemphasis has been used in speech analysis. In this section, both effects are examined.

First, Table 2 (a) compares the recognition scores for HMM-PW.N with each of 8, 16, and 26 states for the pink noise of 12 dB and the car noise of 6 dB. Table 2 (b) also compares those for the conventional HMM. From these tables, it is seen that the larger number of states improves the recognition scores, especially in the HMM-PW.N. This is because the effective minimum duration is bounded by forcing a larger number of states on steady parts of speech.

Second, Table 3 compares the recognition scores by the fixed and adaptive preemphasises in several noise conditions. These results show that the adaptive preemphasis significantly improves the scores in low SNR conditions. Therefore, the adaptive preemphasis is effective to eliminate the spectral slope difference caused by additive noise.

4. CONCLUSION

This paper has presented the frequency-weighted HMM as a method to incorporate human auditory characteristics as well as to utilize the robustness of quefrency weighted cepstrum. As a result, it has been shown that the robustness of HMMs is significantly improved by fixing ' the covariances to frequency-weithing matrices based on a prior knowledge. The most effective frequency-weighting function was a smoothed and compressed power spectrum with a normalized scale derived from the mean vector of each Gaussian probability density function. The frequency-weighted HMM achieved high recognition scores in medium range of SNR.

Table 2 Effects of the number of states on the frequency-weithed HMM (HMM-PW.N) and a conventional HMM (HMM)

HMM model	HMM-PW.N			HMM		
No. of states	8	16	26	8	16	26
pink (6dB)	55	86	98	46	62	57
(0dB)	28	52	80	31	22	20
Car (6dB)	89	96	98	56	79	94
(0dB)	68	74	80	43	53	44

Table 3 Effects of adaptive preemphasis.

Noise	Colored			Car		
SNR (dB)	-6	0	6	-6	0	6
Fixed	15	56	96	41	65	100
Adaptive	32	80	98	45	80	98

References

[1] Y.Ephraim, "Gain-adapted Hidden Markov Models for recognition of clean and noisy speech", IEEE Transactions On Signal Processing, Vol 40, No.6, pp.1303-1316,(1992-6)

[2] B.A.Carlson and M.A.Clements, "Speech recognition in noise using a projection-based likelihood measure for mixture density HMM's", Proc. IEEE ICASSP, pp.237-240, (1992)

[3] B.H.Juang and K.K.Paliual, "Hidden Markov Models with first-order equalization for noisy speech recognition ", IEEE Transaction On Signal Processing, Vol 40,No.9,pp.2136-2143,(1992-9)

[4] A.P.Varga and R.K.Moore, "Hidden Markov Model decomposition of speech and noise", Proc. IEEE ICASSP, pp.845-848, (1990)

[5] M.J.F. Gale and S.J. Young, "Cepstral parameter compensation for HMM recognition in noise," Speech Communication, 12, pp.231-240, 1993.

[6] B.A.Hanson and H.Wakita, "Spectral slope distance measures with linear prediction analysis for word recognition in noise", IEEE Trans. Acoust.,Speech & Signal Process., ASSP-35,7, pp.968-973, (1987-7)

[7] F.Itakura and T.Umezaki, "Distance measure for speech recognition based on the smoothed group delay spectrum," in Proc. ICASSP, Dallas, pp.1257-1260, Apr. 1987.

[8] M. Sugiyama and K. Shikano, "LPC peak weighted spectral matching measures," IECE Trans., Vol.J64-A, No. 5, pp. 409-416, May 1981, (in Japanese).

[9] H.Matsumoto and H.Mitsui, "A robust distance measure based on group delay difference weighted by power spectra", in Proc. ICSLP-90, Kobe, pp.267-270, Nov. 1990.

[10] D.B.Paul, "A Speaker-stress resistant HMM isolated word recogniser," Proc. IEEE ICASSP, pp.713-716, Apr. 1987.