A FREQUENCY-WEIGHTED HMM BASED ON MINIMUM ERROR CLASSIFICATION FOR NOISY SPEECH RECOGNITION

Hiroshi Matsumoto and Masanori Ono

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

ABSTRACT

As a noise robust HMM, we previously proposed a frequency-weighted HMM (HMM-FW) whose covariance matrices are replaced by the inverse of frequency-weighting matrices. In this HMM, the frequency-weighting parameters were common to all classes and states, and were experimentally adjusted. In order to achieve further noise robustness, this paper examines the class- and state-dependent weighting parameters and their minimum error classification training (MCE) of their weighting characteristics. Using the NOISEX-92 database, the MCE-trained HMM-FWs are shown to be more robust even under untrained noise conditions than both the previous HMM-FW and conventional HMM.

1. INTRODUCTION

The approaches to noise robust speech recognition are broadly classified into speech enhancement [1],[2] in the front end and robust parameter and/or pattern matching in the recognition phase [3],[4],[5],[6], [7]. In HMM-based speech recognition, adaptation methods such as PMC [4] are very successful at moderate noise levels. However, these approaches may be difficult to adapt to rapidly varying environmental characteristics. In order to cope with this difficulty, it is important to make HMMs themselves robust to such variations.

As a noise robust HMM, we previously proposed a frequency-weighted HMM (HMM-FW) [8] similar to robust distance measures [5], [6], citematsu90. This HMM has been proved to be robust to additive noise over a wide range of SNR due to the use of both the group-delay spectra and the fixed covariances derived from frequency-weighting coefficients. Although HMM-FW can not deal with severe noise conditions, unlike adaptation approaches, it can be combined with speech enhancement techniques to achieve further robustness [9].

In the previous frequency-weighted HMM, the frequency-weighting parameters were common to all classes and states, and were experimentally adjusted based on recognition tests [8]. In order to achieve further noise robustness, this paper examines the class- and state-dependent weighting parameters and their minimum error classification training (MCE) [10],[11] in optimizing the weighting characteristics.

In the next section, the frequency-weighted HMM is described, followed by section 3 in which the MCE training procedure is presented. In section 4, the performance of the MCE-trained frequency-weighted HMMs is compared with those of the previous HMM-FW and conventional HMM under large variety of noise conditions, using the NOISEX-92 database.

2. FREQUENCY-WEIGHTED HMM

In frequency-weighted HMMs, we use a p-dimensional discrete group delay spectrum $\boldsymbol{y}$ as an observation vector to utilize its robustness to noise [6]. The $\boldsymbol{y}$ is given by the inverse cosine transform of quefrency-weighted cepstral coefficients:

$$\boldsymbol{y} = \boldsymbol{C}\boldsymbol{diag}[1, 2, \cdots, p]\boldsymbol{y}^c, \qquad (1)$$

where $\boldsymbol{C}$ represents the $(p \times p)$ cosine transform matrix, and $\boldsymbol{y}^c$ is a cepstral vector $[c_1, c_2, \cdots, c_p]$. Furthermore, in a frequency-weighted HMM, in order to utilize the robustness of group-delay spectrum and also to incorporate the human auditory characteristics into HMM, the uncorrelated covariance matrix of a single Gaussian HMM for the sth state of cth class is replaced by the inverse of frequency-weighting matrix derived from the mean

vector as follows:

$$\Sigma_{cs} = \rho_{cs}^2 \,\mathrm{diag}\,[w_{cs1}, \cdots, w_{csp}]^{-1}, \qquad (2)$$

where w_{csj} is a frequency weighting coefficient, and ρ_{cs} a scale factor. The w_{csj} is the smoothed and compressed power spectrum derived from the mean vector μ_{cs} as follows:

$$w_{csj} = \exp\{\beta_{cs} l_{csj}\}, \qquad (3)$$

$$[l_{cs1}, \cdots, l_{csp}]^T = C\,\mathrm{diag}\left[\frac{1}{1}, \cdots, \frac{1}{q}, 0, \cdots, 0\right] C^{-1} \cdot \mu_{cs}, \qquad (4)$$

and also, the w_{csj} was normalized so as to give

$$\sum_{j=1}^{p} w_{csj} = 1. \qquad (5)$$

Thus, the weighting characteristics of each state is controlled by (1) a scale factor ρ_{cs}, (2) a compression/expansion factor β_{cs}, and (3) a truncation order q for smoothing.

In the previous HMM-FW, the values of ρ_{cs} and β_{cs} for all classes and states were tied together. As a result, the previous HMM-FW did not take the spectral differences among mean vectors into account.

In order to improve this limitation, ρ_{cs} and β_{cs} are trained independently of each class and state based on a minimum error classification criterion. The other parameter q is set to 8 as in the previous HMM-FW [8].

3. MINIMUM ERROR CLASSIFICATION LEARNING

3.1. Error Criterion

First, we define the misclassification measure for the nth training token y_{kn} from the kth class as follows:

$$d(y_{kn}, \Theta) = -\log\left(P(y_{kn}, y_c \,|\, \theta_c)\right) + \log\left(\left[\frac{1}{h}\sum_{c, c\neq k}^{h} P(y_{kn}, y_c \,|\, \theta_c)^{\eta}\right]^{\frac{1}{\eta}}\right) \qquad (6)$$

where y_c is the optimum Viterbi state sequence of y_{kn} for the cth class, and θ_c denotes the weighting parameters of the cth model. In this study, the values of η and h are set to 2 and 5, respectively. Using $d(y_{kn}, \Theta)$, the total loss over N_k tokens from each of K classes is given by

$$L(y, \Theta) = \sum_{k=1}^{K}\sum_{n=1}^{N_k} l(d(y_{kn}, \Theta)), \qquad (7)$$

where $l(d(y_{kn}, \Theta))$ is the cost function defined here by

$$l(d(y_{kn}, \Theta)) = \begin{cases} d(y_{kn}, \Theta), & if \quad d(\cdot) > 0 \\ 0, & otherwise. \end{cases} \qquad (8)$$

3.2. Training Algorithm

First, we derive the previous HMM-FW from an initial HMM trined by the Baum's algorithm. We calculate the matrix Σ_{cs} for each state using equations (2) to (5) and then replace the covariance matrix of the initial model with Σ_{cs}. With the fixed Σ_{cs}, the mean vector μ_{cs} and transition probabilities $\{a_{ij}\}$ are reestimated by Baum's algorithm until it converges.

Second, with the fixed μ_{cs} and $\{a_{ij}\}$, the ρ_{cs} or β_{cs} is separately modefied by MCE. At the nth iteration of the gradient decent algorithm, the parameter Θ ($\{\rho_{cs}\}$ or $\{\beta_{cs}\}$) is updated by the following equation:

$$\Theta(n) = \Theta(n-1) - \varepsilon(n)\,\nabla L(y, \Theta). \qquad (9)$$

In calculating $\nabla L(y, \Theta(n))$, we use the parameter $\xi_{cs} = \log \rho_{cs}$ instead of ρ_{cs}, and the derivative of the loss function [11] defined by

$$\frac{\partial l(d(y_{k,n}, \Theta))}{\partial d(y_{k,n}.\Theta)} = \begin{cases} 1, & if \quad d(\cdot) > 0, \\ 0, & otherwise. \end{cases} \qquad (10)$$

The step size $\varepsilon(n)$ in equation (9) is controlled as fallows:

$$\varepsilon(n) = \varepsilon(n-1) \cdot 2^{\cos\phi_{n,n-1}}, \qquad (11)$$

where $\cos\phi_{n,n-1}$ is the directional cosine between $\nabla L(y, \Theta(n))$ and $\nabla L(y, \Theta(n-1))$. By equation (11), for $\cos\phi_{n,n-1}$ close to 1, the update proceeds in a similar direction and thus the step size is made larger. But, for $\cos\phi_{n,n-1}$ close to -1, the previous update might have passed an optimum point and thus the step size is made smaller.

Table 1. Data and experimental conditions

Training Testing	10 utterances 10 utterances
Noises	Car noise, White noise Speech babble noise, F16 noise
HMM Recognition Error	26 states Viterbi algorithm Substitution

Table 2. Analysis conditions

Sampling Frequency	16kHz
Window	25ms Hamming
Frame Period	10ms
Preemphasis	Adaptive : $(1\text{-}az^{-1})$
Order of LPC	26
Spectral Parameter	16 mel-Cepstral Coeff.

4. EVALUATION

4.1. Apeech Data and Analysis Conditions

In the following evaluation, the speech database from NOISEX-92 and experimental conditions in Table 1 were used. The analysis conditions are summarized in Table 2. The white noise was generated in a computer, and was added to clean speech so that the global SNR for each word is equal to a predetermined value. In testing, the Vitarbi algorithm was used with fixing the beginning and end points to those in the label files. Thus, only substitution errors were scored.

4.2. Robustness to closed noise conditions

First, the robustness of the following HMMs were compared under various levels of white and car noises:

(1) A conventional HMM with a single diagonal covariance trained by a maximum likelihood method using clean speech(HMM-M1).

(2) The previous frequency-weighted HMM (HMM-FW) derived from the HMM-M1, where the parameters β were set to 0.6 for white noise and 0.4 for car noise, respectively.

(3) An HMM-FW with the scales ρ_{cs} trained by MCE using noisy speech at 0dB SNR of white noise and at -6dB of car noise from the initial models of HMM-FW (MCE-ρ).

Table 3. Effects of MCE for white noise [%]

Type of HMMs	Test SNR [dB]					
	0	6	12	18	24	CL
HMM-M1	10	20	44	69	97	100
HMM-FW	52	86	100	100	100	100
MCE-ρ	90	100	100	100	100	100
MCE-β	88	99	100	100	100	100

Table 4. Effects of MCE for car noise [%]

Type of HMMs	Test SNR [dB]					
	-6	0	6	12	18	CL
HMM-M1	29	54	93	100	100	100
HMM-FW	50	84	100	100	100	100
MCE-ρ	63	94	100	100	100	100
MCE-β	61	92	100	100	100	100

(4) An HMM-FW with β_{cs} trained by MCE under the same conditions as in MCE-ρ (MCE-β).

The recognition accuracy for four HMMs are shown in Table 3 and 4. From these tables, the results are summarized as follows: MCE-ρ achieves 13 to 38% higher accuracy at 6 to 0dB of white noise, and 8 to 13% higher accuracy at 0 to -6dB SNR of car noise than the previous HMM-FW. MCE-β attains almost the same recognition accuracy as MCE-ρ.

4.3. Robustness to open noise conditions

Finally, the conventional HMM with three mixtures per state (HMM-M3) and MCE-ρ were trained using 30 speech samples per digit that contained 10 speech samples under each of three conditions: noise-free, white noise of 0 dB SNR, and car noise of -6dB SNR. The recognition performance of both HMMs were compared for the open noise conditions of speech and F16 noises as well as for the closed noise conditions of white and car noise.

Tables 5 to 8 show the results. The recognition accuracy for MCE-ρ is lower than that of HMM-M3 under the same noise condition that is included in the training, but for open conditions – especially under speech and F16 noise – MCE-ρ proved to be significantly superior to HMM-M3.

Table 5. Gaussian white noise [%]

Type of HMMs	Test SNR [dB]					
	0	6	12	18	24	CL
HMM-M3	100	99	95	97	100	100
MCE-ρ	89	100	100	100	100	100

Table 6. Car noise [%]

Type of HMM	Test SNR [dB]					
	-6	0	6	12	18	CL
HMM-M3	99	100	100	100	100	100
MCE-ρ	78	99	100	100	100	100

Table 7. Speech babble noise [%]

Type of HMMs	Test SNR [dB]					
	-6	0	6	12	18	CL
HMM-M3	29	35	76	100	100	100
MCE-ρ	35	68	97	100	100	100

Table 8. F16 aircraft noise [%]

Type of HMMs	Test SNR [dB]					
	-6	0	6	12	18	CL
HMM-M3	10	16	26	77	100	100
MCE-ρ	19	36	66	87	100	100

5. CONCLUSION

This paper has shown that the MCE-trained frequency-weighted HMM achieves high robustness for a wide variety of noise levels and noise spectra. It should be noted that the mean vectors of the frequency-weighted HMM are fixed to those of clean speech. Therefore, we can expect further robustness of MCE-ρ by combining some noise compensation or noise reduction technique in the front end.

In future work, this frequency-weighted HMM will be extended to HMMs with mixture components, and its simultaneous optimization of different weighting parameters will be examined using a larger database.

REFERENCES

[1] P.Lockwood and J.Boudy, "Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models, and the projection for robust speech recognition in cars," Speech Communication, pp.215-228, 1992.

[2] 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)

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

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

[5] 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)

[6] 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.

[7] 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.

[8] H.Matsumoto and H.Imose, "A frequency-weighted continuous density HMM for noisy speech recognition", in Proc. ICSLP-94, Yokohama, pp.1007-1010, Sep. 1994.

[9] H.Matsumoto and N.Naitoh, "Smoothed spectral subtraction for a frequency-weighted HMM in noisy speech recognition," in Proc. ICSLP-96, Philadelphia, pp.905-908, Sep. 1996.

[10] B.H.Juang and S.Katagiri,"Discriminative Learning for Minimum Error Classification", IEEE Trans. Signal Processing, vol.40, no.12, pp.3043 - 3054,(1992.12).

[11] Kazumi Ohkura, David Rainton and Masahide Sugiyama, "Noise-Robust HMMs Based on Minimum Error Classification", Proc. of ICASSP93, pp.II-75-II-77,(1993).