ERROR CORRECTION FOR SPEAKER-INDEPENDENT ISOLATED WORD RECOGNITION THROUGH LIKELIHOOD COMPENSATION USING PHONETIC BIGRAM

H. Matsuo

Faculty of Engineering and
Resource Science, Akita University
Akita-shi, Akita 010-8502, Japan
matsuo@ie.akita-u.ac.jp

M. Ishigame

Faculty of Software and Information
Science, Iwate prefectural University
Takizawa-mura, Iwate 020-0173, Japan
ishigame@soft.iwate-pu.ac.jp

ABSTRACT

We propose an error correction technique for speaker-independent isolated word recognition by compensating for a word's likelihood. Likelihood is compensated for by likelihood calculated by a phonetic bigram. The phonetic bigram is a phoneme model expressing frame correlation within an utterance. A speaker-independent isolated word recognition experiment showed that our proposed technique reduces recognition error compared to conventional techniques. The proposed technique achieves performance almost equal that without speaker adaptation compared to the conventional phoneme model adapted using several words.

1. INTRODUCTION

Frame correlation for hidden Markov model (HMM) is often used to increase phoneme model accuracy[1][2]. Techniques using frame correlation model the relationship between state transitions and a pair of consecutive frames to express spectral transition.

We developed a phoneme model, the phonetic bigram[3], that models the relationship between pairs of separate frames in phonemes within an utterance. Local features such as spectral transition are expressed by the feature extracted from several frames of spectra and global features expressed by the phonetic bigram. A phoneme recognition experiment demonstrated the phonetic bigram's effectiveness.

We also propose error correction for speaker-independent isolated word recognition by compensating for a word's likelihood[4]. Likelihood is compensated for by the likelihood calculated by the phonetic bigram. A speaker-independent isolated word recognition experiment showed that the proposed technique reduces recognition error by about 19% compared to conventional techniques. A comparison of the proposed technique to speaker adaptation showed that the potential of the proposed technique is comparable to the phoneme model adapted by several words using maximum a posteriori probability estimation(MAP)[5]. The proposed technique does not require speaker adaptation.

2. PHONETIC BIGRAM

The probability of phoneme sequence is shown in Equation 1, which has a correlation between a spectrum and a phoneme and between separate frames of spectra.

$$\begin{aligned} &P\left(\mathbf{y}_1\mathbf{y}_2\cdots\mathbf{y}_n\middle|\, x_1x_2\cdots x_n\right) \\ &\quad= P\left(\mathbf{y}_1\middle|\, x_1\right) P\left(\mathbf{y}_2\middle|\mathbf{y}_1x_1x_2\right) \\ &\qquad\cdots P\left(\mathbf{y}_n\middle|\mathbf{y}_1\mathbf{y}_2\cdots\mathbf{y}_{n-1}x_1x_2\cdots x_n\right) \qquad (1)\end{aligned}$$

where x_n is the n'th phoneme in a phoneme sequence and $\mathbf{y}_n$ is its observation vector. A phoneme's probability is shown in Equation 2 assuming that the spectrum of a certain phoneme is correlated with spectra of all preceding phonemes. This model is called a phoneme bigram.

$$\begin{aligned} &P\left(\mathbf{y}_m\middle|\mathbf{y}_1\mathbf{y}_2\cdots\mathbf{y}_{m-1}x_1x_2\cdots x_m\right) \\ &\quad= \left\{\prod_{k=1}^{m-1}\frac{P\left(\mathbf{y}_k\mathbf{y}_m\middle|\, x_kx_m\right)}{P\left(\mathbf{y}_k\middle|\, x_kx_m\right)}\right\}^{\frac{1}{m-1}}, \\ &\qquad\qquad (m>1) \qquad (2)\end{aligned}$$

3. LIKELIHOOD COMPENSATION FOR ISOLATED WORD RECOGNITION

Equations 1 and 2 cannot be used as for word recognition due to enormous computational cost and the fact that the reliability of preceding phonemes is not considered. Computational cost is directly proportional to the square of the number of phonemes (Equation 2). Frames of preceding phonemes are used regardless of reliability, leading to use of incorrect frames.

We propose the phonetic bigram as postprocessing rather than direct phonetic bigram use. Isolated words are first recognized using a conventional phoneme model. A posteriori probability is calculated simultaneously for each phoneme frame by frame. Spectra weighted by a posteriori probability are summed, then a phonetic image of the word is calculated by normalizing them (Equation 3).

$$\mathbf{z}_W(\omega) = \frac{\sum_t \{PP_W(t,\omega)\,\mathbf{y}(t)\}}{\sum_t PP_W(t,\omega)} \tag{3}$$

$$PP_W(t,\omega) = \frac{P(t,\omega)}{\sum_{\xi\in W} P(t,\xi)} \tag{4}$$

where W is a word, t is the frame number, ω is a phoneme in W, $\mathbf{z}_W(\omega)$ is the phonetic image of ω in W, $P(t,\omega)$ is the probability of ω in W at t, and $PP_W(t,\omega)$ is a posteriori probability of ω in W at t. The phonetic image is used instead of preceding phoneme spectra in Equation 2. Likelihood is recalculated by phonetic bigram using the phonetic image (Equation 5) for each upper candidate of first word recognition results.

$$\begin{aligned} &P(\mathbf{y}|\, x, \mathbf{z}_W) \\ &= \left\{\prod_{\xi\in W} \frac{P(\mathbf{y},\mathbf{z}_W(\xi)|\, x,\xi)}{P(\mathrm{z}_W(\xi)|\, x,\xi)}\right\}^{\frac{1}{n}} \end{aligned} \tag{5}$$

where n is the number of types of phonemes included in W. Phoneme boundaries determined by a Viterbi algorithm are also used.

The computational cost of this phonetic bigram is directly proportional to the number of types of phonemes, and is reduced below the computational cost of the original phonetic bigram directly proportional to the square of the number of phonemes. The entire computational cost of word recognition is reduced because only a few candidates are calculated. The phonetic image becomes increasingly reliable because reliable spectra are focused over the utterance using the a posteriori probability of phonemes.

The phonetic bigram using a phonetic image is applied to observation probability distribution of phoneme HMM. The phonetic image is extended to calculate each phoneme HMM state (Equation 6).

$$\begin{aligned} &b_x(i,\mathbf{y},\mathbf{z}_W) \\ &= \left[\prod_{\xi\in W}\left\{\sum_k \frac{P(\mathbf{y},\mathbf{z}_W(\xi[k])|\, x[i],\xi)}{P(\mathbf{z}_W(\xi[k])|\, x[i],\xi)} \right.\right. \\ &\qquad \left.\left. \times\, \lambda_{x[i],\xi}(k)\right\}\right]^{\frac{1}{n}} \end{aligned} \tag{6}$$

$$\sum_k \lambda_{x[i],\xi}(k) = 1 \tag{7}$$

where x,ξ are phonemes in W, $x[i]$ is the i'th state of x, $\xi[k]$ is the k'th state of ξ, $\mathbf{z}_W(\xi[k])$ is the phonetic image of $\xi[k]$, $\lambda_{x[i],\xi}(k)$ is the mixture weight of $x[i]$ for $\xi[k]$ and $b_x(i,\mathbf{y},\mathbf{z}_W)$ is the observation probability distribution of $x[i]$.

The result of word recognition is determined by the logarithmic likelihood calculated from the first logarithmic likelihood by the original phoneme HMM and the logarithmic likelihood by the phonetic bigram (Equation 8). The logarithmic likelihood of the original phoneme HMM is compensated for by the phonetic bigram.

$$\begin{aligned} L(W) &= \rho\log(P(\mathbf{y}_1\mathbf{y}_2\cdots|\, W)) \\ &\quad + (1-\rho)\log(P(\mathbf{y}_1\mathbf{y}_2\cdots|\, W,\mathbf{z}_W)) \end{aligned} \tag{8}$$

where ρ is the mixing ratio of logarithmic likelihood and $L(W)$ is the compensated logarithmic likelihood of W.

4. ISOLATED WORD RECOGNITION EXPERIMENT

Experimental conditions were as follows:

- Training/test samples: isolated spoken words (212 Japanese words) uttered by ten men and women each. Words totaled 4174.
- Speech is passed through a 29-channel band pass filter at 10 ms per frame, and the dimension of the spectrum pattern is reduced from 145 (29 channels by 5 frames) to 15 using the K-L expansion twice. The dimension of phonetic image is 10.
- Type of HMM: 16-state non left-right HMM[6] with duration probability distribution. It has phoneme context-dependent state transition probabilities. Other parameters are phoneme context-independent.
- Continuous observation probability distribution (full covariance matrix).
- Each speaker is tested using parameters trained by 19 other speakers (open speaker/closed vocabulary).

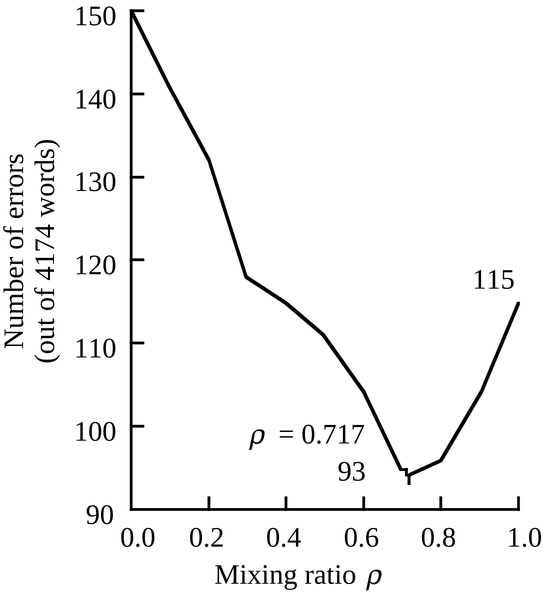

Figure 1: Recognition result

- Training HMM: Training sections are determined by hand-label exactly.

The mean vector of the observation probability distribution is initialized by small random numbers. Word recognition uses a beam search with a bandwidth of 50. The word's phonetic image is calculated from 50 candidates for each frame.

Recognition error is shown in Figure 1, where ρ is the mixing ratio demonstrated in Equation 8. Logarithmic likelihood mixing (Equation 8) was effective experimentally. At $\rho = 0.717$, error was a minimum and reduced about 19% compared to that at $\rho = 1.0$ (conventional model only). The number of errors at $\rho = 0.0$ (phonetic bigram only) increases compared to that at $\rho = 1.0$, indicating that conventional recognition error may sometimes be expanded by the phonetic bigram. Often, the proposed technique reduces error. The reasons are (1) the difference in likelihood between correct and incorrect answers widens when the phonetic bigram operates correctly and (2) tendencies of error differ between the conventional phoneme model and phonetic bigram.

5. COMPARISON WITH SPEAKER ADAPTATION

Speaker adaptation is widely used to make phoneme models more accurate. In an experiment comparing supervised speaker adaptation with the proposed technique, the original phoneme model was adapted by MAP using words from samples of one person tested. Remaining samples were tested by the adapted model. Word recognition error rate is used for evaluation because the number of samples differs from the previous experiment.

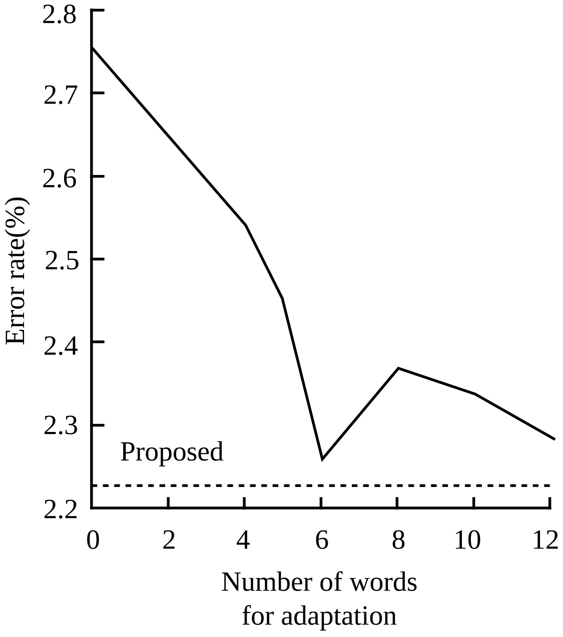

Figure 2: Comparison with speaker adaptation

Recognition error rate becomes flat over 6 words for adaptation (Figure 2). The error rates for the proposed technique and speaker adaptation are almost equal.

6. CONCLUSION

We proposed error correction for speaker-independent isolated word recognition by compensating for the likelihood of a word. Likelihood is compensated for by the likelihood calculated using the phonetic bigram. Two types of experiments were carried out. The number of word recognition errors is reduced about 19% compared to that of conventional techniques. Though speaker adaptation is not adopted in the proposed technique, the error rate of the proposed technique is almost equal to that for speaker adaptation, confirming the proposed technique's effectiveness. Computational cost of the proposed technique exceeds that for the conventional phoneme model, but it has an advantage when speaker adaptation is not available. Theoretically, the proposed technique is applicable to continuous speech recognition, and speaker adaptation can be applied to the phonetic bigram.

7. REFERENCES

[1] Wellekens C.J. "Explicit correlation in hidden Markov model for speech recognition". Proc. ICASSP-87, pp.384-386, 1987.

[2] Takahashi S., Matsuoka T., Minami Y. and Shikano K. "Phoneme HMMs constrained by frame correlations". Proc. ICASSP-93, 2, pp.219-222, 1993.

[3] Matsuo H. and Ishigame M. "A phoneme model considering correlation between phonemes within a word". Proc. Autumn Meet. Acoust. Soc. Jpn., 1-1-8, pp.15-16, 1997(in Japanese).

[4] Matsuo H. "An error correction method for isolated word recognition by using phonetic bigram". Proc. Autumn Meet. Acoust. Soc. Jpn., 1-1-9, 1998(in Japanese).

[5] Stern R.M. and Lasry M.J. "Dynamic speaker adaptation for feature-based isolated word recognition". IEEE Trans. Acoust., Speech & Signal Process., ASSP-36, 6, pp.751-763, 1987.

[6] Matsuo H. and Ishigame M. "Phoneme Recognition Based on HMM Taking Account of Learning Topology and Interaction between Phoneme Models". Trans. IEICE, vol. J76-D-II, No.9, pp.1835-1842, 1993(in Japanese).

APPENDIX

Reestimation for phonetic bigram: The following have been tested.

$$\alpha\left(t+1,j\right)=\sum_{i}\alpha\left(t,i\right)a\left(i,j\right)b\left(i,\mathbf{y}_t,\mathbf{z}_W\right)$$

$$\beta\left(t,i\right)=\sum_{j}a\left(i,j\right)b\left(i,\mathbf{y}_t,\mathbf{Z}_W\right)\beta\left(t+1,j\right)$$

$$t=1\cdots T,\alpha\left(1,i\right)=\pi\left(i\right),\beta(T+1,.)=1$$

where $\alpha\left(t,i\right)$ is forward probability of i'th state at t, $\beta\left(t,i\right)$ is backward probability of i'th state at t, $a\left(i,j\right)$ is state transition probability from i'th state to j'th state, and $\pi\left(i\right)$ is initial probability of i'th state.

$$b_x\left(i,\mathbf{y}_t,\mathbf{z}_W\right)=\left[\prod_{\xi\in w}\left\{\sum_{k}b'_{x[i],\xi}\left(k,\mathbf{y}_t\right)\right\}\right]^{\frac{1}{n}}$$

$$b'_{x[i],\xi}\left(k,\mathbf{y}_t\right)=\frac{P\left(\mathbf{y}_t,\mathbf{z}_W\left(\xi\left[k\right]\right)\middle|\,x\left[i\right],\xi\right)}{P\left(\mathbf{z}_W\left(\xi\left[k\right]\right)\middle|\,x\left[i\right],\xi\right)}\lambda_{x[i],\xi}\left(k\right)$$

$$P\left(\mathbf{y}_t,\mathbf{z}_W\left(\xi\left[k\right]\right)\middle|\,x\left[i\right],\xi\right)$$

$$=\frac{1}{\left(2\pi\right)^{\frac{\mathrm{D}_1+\mathrm{D}_2}{2}}\left|\mathbf{\Sigma}_{x[i],\xi}\right|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}\left(\mathbf{y}^*_t\left(\mathbf{z}_W\left(\xi\left[k\right]\right)\right)-\mu_{x[i],\xi}\right)^{\mathrm{T}}\mathbf{\Sigma}^{-1}_{x[i],\xi}\left(\mathbf{y}^*_t\left(\mathbf{z}_W\left(\xi\left[k\right]\right)\right)-\mu_{x[i],\xi}\right)\right)$$

$$\mathbf{y}^*_t\left(\mathbf{z}_W\left(\xi\left[k\right]\right)\right)=\begin{bmatrix}\mathbf{y}_t\\ \mathbf{z}_W\left(\xi\left[k\right]\right)\end{bmatrix}$$

$$P\left(\mathbf{z}_W\left(\xi\left[k\right]\right)\middle|\,x\left[i\right],\xi\right)$$

$$=\frac{1}{\left(2\pi\right)^{\frac{\mathrm{D}_2}{2}}\left|\mathbf{\Sigma}'_{x[i],\xi}\right|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}\left(\mathbf{z}_W\left(\xi\left[k\right]\right)-\mu'_{x[i],\xi}\right)^{\mathrm{T}}\mathbf{\Sigma}'^{-1}_{x[i],\xi}\left(\mathbf{z}_W\left(\xi\left[k\right]\right)-\mu'_{x[i],\xi}\right)\right)$$

$\mathrm{D}_1=15$ and $\mathrm{D}_2=10$ in this paper.

$$\gamma\left(t,i,j\right)=\frac{\alpha\left(t,i\right)a\left(i,j\right)b\left(\mathbf{y}_t,i,\mathbf{z}_W\right)\beta\left(t+1,j\right)}{\sum_{i}\alpha\left(\mathrm{T}+1,i\right)}$$

$$\gamma'\left(t,i\right)=\sum_{j}\gamma\left(t,i,j\right),\phi_{x[i],\xi}\left(k,\mathbf{y}_t\right)=\frac{b'_{x[i],\xi}\left(k,\mathbf{y}_t\right)}{\sum_{k}b'_{x[i],\xi}\left(k,\mathbf{y}_t\right)}$$

reestimation

$$\mu_{x[i],\xi}=\sum_{t}\left[n\gamma'(t,i)^{\frac{1}{n}}\sum_{k}\left\{\phi_{x[i],\xi}\left(k,\mathbf{y}_t\right)\times\mathbf{y}^*_t\left(\mathbf{z}_W\left(\xi\left[k\right]\right)\right)\right\}\right]/\sum_{t}n\gamma'(t,i)^{\frac{1}{n}}$$

$$\mathbf{\Sigma}_{x[i],\xi}=\sum_{t}\left[n\gamma'(t,i)^{\frac{1}{n}}\sum_{k}\left\{\phi_{x[i],\xi}\left(k,\mathbf{y}_t\right)\times\left(\mathbf{y}^*_t\left(\mathbf{z}_W\left(\xi\left[k\right]\right)\right)-\mu_{x[i],\xi}\right)^{\mathrm{T}}\times\left(\mathbf{y}^*_t\left(\mathbf{z}_W\left(\xi\left[k\right]\right)\right)-\mu_{x[i],\xi}\right)\right\}\right]/\sum_{t}n\gamma'(t,i)^{\frac{1}{n}}$$

$$\mu'_{x[i],\xi}=\sum_{t}\left[n\gamma'(t,i)^{\frac{1}{n}}\sum_{k}\left\{\phi_{x[i],\xi}\left(k,\mathbf{y}_t\right)\times\mathbf{z}_W\left(\xi\left[k\right]\right)\right\}\right]/\sum_{t}n\gamma'(t,i)^{\frac{1}{n}}$$

$$\mathbf{\Sigma}'_{x[i],\xi}=\sum_{t}\left[n\gamma'(t,i)^{\frac{1}{n}}\sum_{k}\left\{\phi_{x[i],\xi}\left(k,\mathbf{y}_t\right)\times\left(\mathbf{z}_W\left(\xi\left[k\right]\right)-\mu'_{x[i],\xi}\right)^{\mathrm{T}}\times\left(\mathbf{z}_W\left(\xi\left[k\right]\right)-\mu'_{x[i],\xi}\right)\right\}\right]/\sum_{t}n\gamma'(t,i)^{\frac{1}{n}}$$

$$\lambda_{x[i],\xi}=\sum_{t}n\gamma'(t,i)^{\frac{1}{n}}\phi_{x[i],\xi}\left(k,\mathbf{y}_t\right)/\sum_{t}n\gamma'(t,i)^{\frac{1}{n}}$$

a and π are reestimated from γ in the same way of conventional HMM.