MINIMUM PHONEME ERROR BASED FILTER BANK ANALYSIS FOR SPEECH RECOGNITION

Hao Huang, Jie Zhu

Department of Electronic Engineering of Shanghai Jiao Tong University, Shanghai, 200240, P.R. China
{haohuang,zhujie}@sjtu.edu.cn

ABSTRACT

In this paper the optimal filter-bank design method based on the Minimum Phone Error (MPE) criteria is investigated. We use Gaussian type filter bank for optimization and various parameters of the filters such as gain, bandwidth and center frequency are trained aiming at maximize the MPE objective function to reduce word error. Preliminary experimental results on a large vocabulary continuous Mandarin speech recognition task given in this paper showed that, compared with both the untrained Gaussian type filters and traditional triangle shaped filter bank, cepstral coefficients derived from the optimized filter bank parameters result in a superior performance for word accuracy. The filters consistent with the MPE criteria are also illustrated.

1. INTRODUCTION

Mel frequency cepstral coefficient (MFCC) is one of the most widely used speech feature representation in state-of-the-art automatic speech recognition systems. In such implementation scheme, speech signal is first pre-emphasized, windowed, transformed to the frequency domain (usually with FFT), and correlated with a bank of triangular filters which are equally spaced on a linear-log frequency axis, and the sum of magnitude coefficients scaled by each filter is log-compressed and transformed via the Discrete Cosine Transform (DCT) to cepstral coefficients. The purpose of applying filter bank analysis it to obtain a desired non-linear frequency resolution [1]. Usually the shapes of the filters are chosen based on a priori knowledge of the speech signal without taking into account the back-end classifier and may not result in a optimal recognition accuracy [2]. And a problem with such an implementation, as pointed out in [4], is that the bandwidth and center frequency of each triangle is determined by the frequency range of the filter bank as well as the number of filters in the bank. Hence [3, 2] originally proposed a overall design of filter bank and recognizer and provide an recognition-oriented filter bank analysis. This work optimized the filter-bank parameters as well as the prototype based distance classifier structure using the Minimum Classification Error/Generalized Probabilistic Descent (MCE/GPD) method.

On the other hand, recent research showed that discriminative training methods based on Minimum Mutual Information (MMI) [5] and Minimum Phoneme Error (MPE) [6] criterion to optimize Gaussian parameters achieved significant improvement for recognition accuracy. In [7] the MPE based Heteroscedastic Linear Discriminant Analysis (MPE-HLDA) was introduced and the proposed linear transformation directly relate to the objective of reducing recognition error rate showed performance improvement.

Inspired by these works, we applied the MPE optimization method to the filter-bank analysis in the MFCC and HMM based framework. Optimal cepstral coefficients are obtained from a overall designed filter bank directly aiming at reducing word error rate. Experimental results on a Mandarin LVCSR system show that recognition accuracy is improved with the MFCCs calculated with MPE trained filter bank.

The rest of the paper is organized as follows: In section 2, the filter bank based MFCC and MPE based criteria and its derivatives are briefly reviewed. Then optimization with respect to filter bank parameters are formalized. Section 4 describes the experimental tasks and depicts the implementation. Section 5 and reports the results and gives some analysis. And finally in section 6 the conclusions from the proposed work are given.

2. MPE BASED FITLER BANK DESIGN

2.1. Filter-bank based cepstrum

When Gaussian type filter bank is applied [2, 3], the weighting function is defined as :

$$w_{l,f} = \alpha_l \exp\left[-\beta_l \{p(\gamma_l) - p(f)\}^2\right]$$

where α_l, β_l and γ are gain,band width and center frequency factors of the l-th channel respectively. Note that a larger value of β_l denotes a narrower filter and vice-versa. f denotes the frequency. $p(f)$ is the frequency scaling function. For Mel frequency scaling:

$$p(f) = 1127 \log\left(1 + \frac{f}{700}\right)$$

where m_l represents the windowed log energy output of the $l-$ th channel:

$$m_l = \log_{10}\left(\sum_f^F w_{l,f} x_f\right) \quad (1)$$

where x_f is the FFT output corresponding to either energy or power at frequency f. Then MFCCs are calculated using DCT:

$$c_i = \sqrt{\frac{2}{L}} \sum_l^L m_l \cos\left(\frac{\pi i}{L}(l-0.5)\right)$$

where $i(i = 1, 2, \cdots, I)$ indicates the ith cepstral coefficient. Let α, β and γ denote $\{\alpha_l\}_{l=1}^L$, $\{\beta_l\}_{l=1}^L$ and $\{\gamma_l\}_{l=1}^L$ respectively and let Θ denotes all the filter bank parameters: $\Theta = \{\alpha, \gamma, \beta\}$, then each vector in the output feature space can be expressed as:

$$y = f(\alpha, \beta, \gamma, x) = f(\Theta, x)$$

2.2. MPE objective and derivatives

Given a set of R training utterances, the MPE objective function is defined in [6] as:

$$F_{MPE}(\Theta) = \sum_r^R \frac{\sum_s P_\Theta(O_r|s)^k P(s)^k \epsilon(s)}{\sum_s P_\Theta(O_r|s)^k P(s)^k}$$

where O_r is the observation vectors of utterance r and s is sentence hypothesis in the lattice. $P(s)$ is the probability of the hypothesis as determined by the language model. k is applied to acoustic scores in order to reduce their dynamic range as described in [7]. $\epsilon(s)$ is the a measure of the number of words accurately transcribed in hypothesis s, as described in [6, 7].

The derivative of MPE objective function with respect to certain model parameter was given in [7]. Differentiate the function with respect to Θ to get:

$$\frac{\partial F_{MPE}(\Theta)}{\partial \Theta} = \sum_r^R \sum_s \frac{\partial P_\Theta(s|O_r)}{\partial \Theta}\epsilon(s) = k\sum_r^R D(r)$$

where

$$D(r) = \sum_{i=1}^Q \left\{ \frac{\partial \log P_\Theta(q_i|O_r)}{\partial \Theta} P_\Theta(q_i|O_r) \cdot \left[\epsilon'(q_i) - f_{mpe}(r)\right]\right\} \quad (2)$$

Q is the number of arcs in the lattice. $\epsilon'(q_i)$ is the average of $\epsilon(s)$ for all the sentences s that contain arc q_i. f_{mpe} is the weighted average $\epsilon(s)$ for all sentences in the lattice, as defined in [6, 7]. It was also proposed that both $\epsilon'(q_i)$ and f_{mpe} can be computed efficiently by a forward-backward pass. The derivative of $\log P(O_{q_i}|q_i)$ with respect to model parameter Θ within each arc can be calculated as follows [7]:

$$\frac{\partial \log P_\Theta(O_{q_i}|q_i)}{\partial \Theta} = \sum_{t=S_{q_i}}^{E_{q_i}} \sum_m \gamma_{q,m}(t) \frac{\partial \log P_\Theta(o_t|m_t)}{\partial \Theta}$$

where s_{q_i} and e_{q_i} are the begin time and end time of arc q_i marked in the lattice, and $\gamma_{q,m}(t)$ is the posterior probability of Gaussian mixture component m in arc q_i.

2.3. MPE based Filter Bank parameter optimization

When diagonal covariance matrix are used and only the derivative of the cepstral coefficients c_i are considered, to make the filter bank parameters for optimization, the derivative of $\log P_\Theta(o_t|m_t)$ with respect to α_l, β_l and γ_l can be express as follows:

$$\frac{\partial \log P_\Theta(o_t|m_t)}{\partial \alpha_l} = -\sum_{i=1}^I \frac{c_i - \mu_i}{\sigma_i^2}\frac{\partial c_i}{\partial \alpha_l}$$

$$\frac{\partial \log P_\Theta(o_t|m_t)}{\partial \beta_l} = -\sum_{i=1}^I \frac{c_i - \mu_i}{\sigma_i^2}\frac{\partial c_i}{\partial \beta_l}$$

$$\frac{\partial \log P_\Theta(o_t|m_t)}{\partial \gamma_l} = -\sum_{i=1}^I \frac{c_i - \mu_i}{\sigma_i^2}\frac{\partial c_i}{\partial \gamma_l}$$

where

$$\frac{\partial c_i}{\partial \alpha_l} = \frac{1}{M_l} c(i,l) \sum_{f=1}^F x_f \exp\left(-\beta_l \delta_{l,f}^2\right)$$

$$\frac{\partial c_i}{\partial \beta_l} = \frac{-1}{M_l} c(i,l) \sum_{f=1}^F x_f \alpha_l \exp\left(-\beta_l \delta_{l,f}^2\right) \delta_{l,f}^2$$

$$\frac{\partial c_i}{\partial \gamma_l} = \frac{-2254}{M_l} c(i,l) \sum_{f=1}^F \alpha_l \beta_l \delta_{l,f} \frac{1}{700+\gamma_l} x_f \exp\left(-\beta_l \delta_{l,f}^2\right)$$

M_l denotes the inner summation within parentheses in (1). And $\delta_{l,f} = p(\gamma_l) - p(f)$, $c(i,l) = \sqrt{\frac{2}{L}}\cos\left(\frac{\pi i}{L}(l-0.5)\right)$. μ_i and σ_i are the ith element of mean and variance vector respectively.

3. EXPERIMENTAL SETUP

We evaluated the presented filter bank analysis scheme on a Mandarin LVCSR task. The corpus Er-Wai [8] from Microsoft Research Asia is used for training. The database contains read speech of about 31.5 hours from 100 male students, for a total 19,688 utterances. In the testing phase, the MSR [8] test set uses additional 0.74 hour 500 utterances from another 25 male speakers. Speech wave forms are sampled at

16kHz and 16bits. The acoustic feature vector has 39 elements. Besides the 12 MFCC coefficients, normalized energy plus first and second time derivative are used. The toolbox default options are slightly changed (with cepstral coefficients liftering and cepstral means normalization removed) for seeing a clear effect of the filter bank parameter optimization. The system uses context dependent triphone units for modeling Mandarin tonal syllables, which use a set of 185 phones proposed. Decision-tree based state tying are used and tied states with 8 Gaussian components per mixture are used in acoustic modeling. About 1/6 out of all the R =19688 train utterance are used to for MPE calculation.

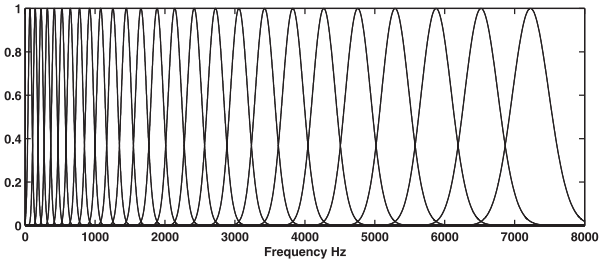


Fig. 1. Gaussian type filter bank (baseline I).

The word level lattices are generated using *HVite* command in HTK [1] based on traditional MLE trained HMMs. Three tokens in each state (*HVite* command option n=3) are used to create word lattices. No language model is applied during both lattice-generating and recognition process. Then word lattices are expanded into triphone based networks according the hypothesis contexts in the lattice. For each utterance an extra correct path is added into the lattice in case that the correct hypothesis is missing. After the triphone based networks are constructed, phone errors are calculated as proposed in [5]. The initial filter bank parameter are chosen to have same gain, bandwidth and center frequency as the HTK default triangular shape filter bank.

We applied four set of experiments similar to those mentioned in [2], which we referred as:

1. MPE-Gg traing: Only gain factor α_l of the L channels are trained while other factors remain fixed.
2. MPE-Gb training: The bandwidth factors β_l are optimize while keep others constant.
3. MPE-Gc training: Only the center frequencies $\gamma_l, l = 1, 2, \cdots, L$ are adjusted.
4. MPE-GS training: The three types of parameters are trained simultaneously.

The iterative training process can be summarized as follows:

1. Initialize the Gaussian type filter bank parameters.
2. Train the models with the traditional Baum-Welch re-estimation.
3. Generate word lattice and expand the lattice into triphone based lattices.
4. Calculate derivertives with respect to α, β, γ in the lattices.
   (a) For each utterance r:
      i. Run forward-backward pass within each arc to compute $\gamma_{q,m}(t)$ and the derivative of $\log P(o_t|m_t)$ with respect to various filterbank parameters (α, β and γ).
      ii. Run forward-backward pass at the lattice level to calculate posterior probability of $P(q|O_r)$.
      iii. Using Viterbi alignment according to the transcription of each training utterance to give mark in time at the phoneme level.
      iv. Run another forward-backward pass at the lattice level to calculate $\epsilon'(q)$ and $f_{mpe}(r)$ in (2).
   (b) Sum all the derivatives for all utterances.
5. Update α, β, γ either selectively or simultaneously according to different training types (Gg training, Gb training, Gc training and GS training) using gradient-descent method.
6. Go to 4 iteratively until convergence or max number of iteration is reached.

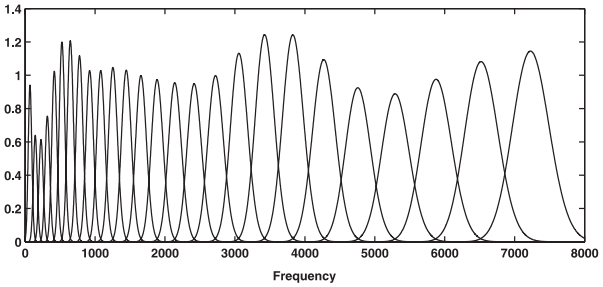
There are two baseline systems in the our test. Baseline I system is a initial Gaussian type filters with the same gain, and bandwidth and center frequency as Baseline II. Baseline II system uses HTK default triangular filter bank and number of channels was set to 26 ($L = 26$).

4. RESULTS

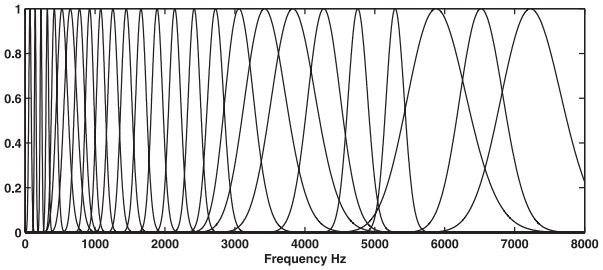
Fig.2 (a)-(d) illustrate the filters trained by MPE. We think differences from those obtained in [2] partly lies in that the parameters are optimized using different tasks and data. Since we use gradient descent method for optimizing, the solutions might not have reached global minima (for minimizing $-F_{MPE}(\Theta)$). However it can be shown some improvements have been achieved. The effect of MPE based filter bank optimization on Word Accuracy (WA) is shown in Table 1. We can see that among these results, MPE-GS training obtains a best Word Accuracy (WA) of 49.16%, an absolute gain of 0.70% compared with that of baseline I and 0.42% w.r.t baseline II.

Table 1: Word Accuracy (WA) using MPE trained filter bank

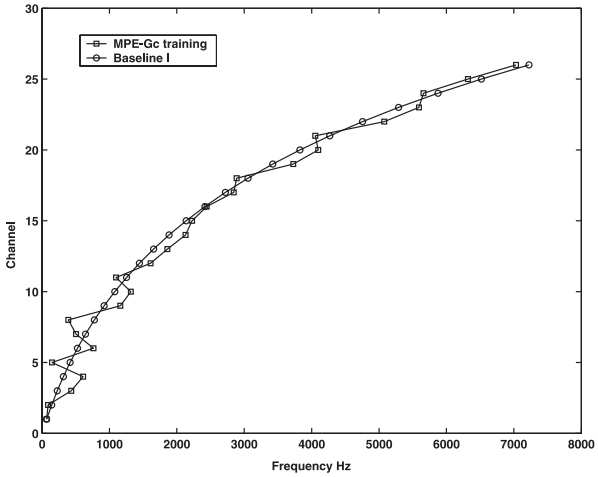
%	BL I	48.46	BL II	48.74
	MPE Gg	MPE Gb	MPE Gc	MPE GS
WA%	49.13	48.92	49.00	49.16



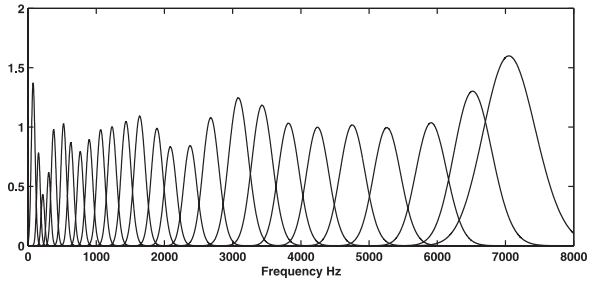
(a) MPE-Gg training



(b) MPE-Gb training



(c) MPE-Gc training



(d) MPE-GS training

Fig. 2. MPE trained filter bank.

5. CONCLUSION

In this paper, a novel scheme is proposed for obtaining optimal filter bank parameters. Gaussian type filter bank parameters were trained consistent with the MPE objective function for reducing word error. Results of experiment on LVCSR system showed word error reduction is achieved using cepstral coefficients obtained from the MPE trained filter bank. Future of our work includes adjusting other front-end feature extraction parameters using MPE method and also see the results of front-end parameters optimizing on top of MPE based model training for better performance.

Acknowledgement

We thank Bing Zhang [7] for helping us in understanding MPE method and derivatives.

6. REFERENCES

[1] S. J. Young et al., The HTK Book, http://htk.eng.cam.ac.uk .

[2] Alain Biem, Shigeru Katagiri, Erik McDermott, and Biing-Hwang Juang,"An application of discriminative feature extraction to filter-bank-based speech recognition," IEEE Transactions on Speech and Audio Processing, vol.9, no.2,pp. 96-110, Feb. 2001.

[3] Alain Biem and Shigeru Katagiri, "Cepstrum-based filter-bank design using discriminative feature extraction training at various levels," in Proceedings of ICASSP 1997.

[4] Mark D. Skowronskia and John G. Harrisb, "Exploiting independent filter bandwidth of human factor cepstral coefficients in automatic speech recognition," J. Acoust. Soc. Am. 116(3), pp.1774-1780,2004.

[5] V. Valchev et al., "MMIE training of large vocabulary recognition systems," Speech Communication 22,pp.303-314, June 1997.

[6] D. Povey and P. C. Woodland,"Minimum phone error and I-smoothing for improved discriminative training," in Proceedings of ICASSP, vol. 1, Page(s):I-105-I-108, 2002.

[7] Bing Zhang and Sypros Matsoukas, "Minimum phoneme error based heteroscedastic linear discriminant analysis for speech recognition," in Proceedings of ICASSP, Vol 1, pp 925-928, 2005.

[8] E. Chang, Y. Shi, J.L. Zhou, and C. Huang, "Speech lab in a box: a Mandarin speech toolbox to jumpstart speech related research," Proceedings of Eurospeech 2001, pp.2779-2782, 2001.