OPTIMIZATION OF SUB-BAND WEIGHTS USING SIMULATED NOISY SPEECH IN MULTI-BAND SPEECH RECOGNITION

Yik-Cheung Tam and Brian Mak

Department of Computer Science,
The Hong Kong University of Science and Technology,
Clear Water Bay, Hong Kong
{cswilson,mak}@cs.ust.hk

ABSTRACT

Recently multi-band speech recognition has been proposed to improve robustness under environmental noises. One important issue is how to combine decisions from individual sub-band recognizers to arrive at a final decision. Under the hidden Markov modeling (HMM) framework, one common approach is combining sub-band likelihoods linearly in an optimal manner so that the more reliable sub-bands are emphasized and the corrupted sub-bands are de-emphasized. In our experience, estimating the weights from clean speech is not effective as the weights are not optimal under noisy environments. In this paper, we derive the optimal weights from *simulated noisy speech* using discriminative training method with minimum classification errors (MCE) or maximum mutual information (MMI) as the cost function. The methods are evaluated on recognition of isolated TI digits. Compared with full-band recognition with noises at an SNR of 0dB, multi-band recognition with MCE-derived weights reduces word errors by 45.9% on a tone noise, and an average of 17.9% on three real noises. MCE-derived weights and MMI-derived weights have similar performance, and are much better than weights derived from other means.

1. INTRODUCTION

Recently multi-band speech recognition has been proposed by Bourlard *et al.* [2] and Hermansky *et al.* [5] to improve robustness under noisy environment. It is motivated by the empirical findings by Harvey Fletcher of Bell Labs [1] from a thorough study of human speech recognition. In their approach, the full frequency band is divided into sub-bands and a speech recognizer is built for each band. During recognition, decisions from individual sub-band recognizers are recombined to arrive at a final decision at some phonetic/linguistic level. Under the hidden Markov modeling framework, this reduces to a recombination strategy of sub-band likelihoods. Another approach is taken by Bocchieri *et al.* [7] in which sub-band features are recombined in juxtaposition before model estimation. They called their method "frequency recombination" and the other method "likelihood recombination". This paper concerns only the likelihood recombination strategy in which sub-band likelihoods are linearly combined. The aim is to emphasize the more reliable sub-bands in the final decision, and the degree of emphasis is to be determined automatically and optimally.

Two approaches are commonly used to determine the weightings of sub-band likelihoods:

- Sub-band weights are set according to some reliability measure of the sub-bands. For instance, SNR or entropy-based measure of the sub-bands have been tried [2, 5].
- Optimization techniques are used to derive the "best" [1] sub-band weights. Discriminative training with minimum classification errors (MCE) criterion is used in [3]; and in the context of multiple-stream systems, stream weights have been found with maximum mutual information (MMI) training [4], and maximum likelihood (ML) estimation with additional constraint on weights [6].

Since the first approach does not guarantee to find the optimal sub-band weights, we focus only on the second approach. The referenced works all derive the weights during model training with clean speech. In practice, sub-band weights should be estimated online from little adaptation data. In this paper, we study various optimization methods of finding linear sub-band weights from noisy speech.

In the next Section, we first derive the MCE estimation formulas for sub-band weights, and briefly point out ML and MMI estimations in Section 3. This is followed by recognition experiments in Section 4 and conclusions in Section 5.

2. MINIMUM CLASSIFICATION ERROR ESTIMATION OF LINEAR SUB-BAND WEIGHTS

For simplicity, we will only show estimation formulas in which sub-band weights depend only on the sub-band and not on the model nor the state; the latter can easily be derived in a similar fashion. Model-independent sub-band weights are more interesting because in most practical scenarios, the sub-band weights have to be determined online using very few adaptation data, and model- or state-dependent sub-band weights may not be estimated reliably.

[1] "best" here applies only to linear re-combination of sub-band likelihoods. In fact, they can be re-combined non-linearly using MLP.

Suppose we have a K-sub-band isolated-word recognition system with a vocabulary of M words. The probability of an utterance X_i, $i = 1, 2, \ldots, N$, belonging to the class C_j with sub-band models $\lambda_{j,k}$, $k = 1, 2, \ldots, K$, is given by

$$P(X_i|C_j) = \prod_{k=1}^{K} P(X_i|\lambda_{j,k})^{\omega_k} \ , \tag{1}$$

$$\text{where} \quad 0 \le \omega_k \le 1 \text{ and} \sum_{k=1}^{K} \omega_k = 1 \tag{2}$$

assuming sub-band independence. We will denote $P(X_i|C_j)$ by P_{ij} and $P(X_i|\lambda_{j,k})$ by P_{ijk}, and their log-likelihoods by L_{ij} and L_{ijk} respectively. Thus we have

$$P_{ij} = \prod_{k=1}^{K} {P_{ijk}}^{\omega_k} \tag{3}$$

$$\text{and,} \quad L_{ij} = \sum_{k=1}^{K} \omega_k \, L_{ijk} \ . \tag{4}$$

The misclassification measure $d(X_i)$ is then given by the log-likelihood difference between the mean likelihood of all competing words and the likelihood of X_i. i.e.

$$d(X_i) = \log \left\{ \frac{1}{M-1} \sum_{m \neq j} P_{im} \right\} - L_{ij} \ . \tag{5}$$

The misclassification measure is smoothed using the sigmoid function, $l(d) = 1/\{1 + \exp(-\gamma d + \theta)\}$, to obtain the total loss function over all utterances, $R_{mce} = \sum_i l(d(X_i))$. To satisfy Eqn.(2) throughout the optimization process, parameter transformation is performed:

$$\overline{\omega_k} = log(\omega_k) \tag{6}$$

Taking derivative with respect to the k-th sub-band weight, $\overline{\omega_k}$, we have

$$\frac{\partial R_{mce}}{\partial \overline{\omega_k}} = \sum_i \frac{\partial l}{\partial d} \frac{\partial d(X_i)}{\partial \overline{\omega_k}}, \quad \text{where} \tag{7}$$

$$\frac{\partial l}{\partial d} = \gamma l(d(X_i))[1 - l(d(X_i))] \tag{8}$$

$$\text{and,} \quad \frac{\partial d(X_i)}{\partial \overline{\omega_k}} = \frac{\sum_{m \neq j}^{M} P_{im}(\Delta L_{imk} - \Delta L_{ijk})}{\sum_{m \neq j}^{M} P_{im}} \tag{9}$$

where $\Delta L_{imk} = exp(\overline{\omega_k}) L_{imk}$.

Starting from an initial guess of ω_k where $0 < \omega_k < 1$ and transforming it to $\overline{\omega_k}$ using Eqn.(6), we may use a gradient-descent algorithm to get a better estimate of the sub-band weight for the $(t+1)$-th iteration from its estimate from the t-th iteration:

$$\overline{\omega_k}^{(t+1)} = \overline{\omega_k}^{(t)} - \epsilon_t \left(\frac{\partial R_{mce}}{\partial \overline{\omega_k}} \right) \tag{10}$$

where ϵ_t is the learning rate at the t-th iteration.

Finally, $\overline{\omega_k}$ is transformed back to ω_k after the gradient-descent procedure completes.

3. MAXIMUM MUTUAL INFORMATION AND MAXIMUM LIKELIHOOD ESTIMATION OF LINEAR SUB-BAND WEIGHTS

One may also estimate the sub-band weights by other means. In the following, we will briefly discuss two common methods.

3.1. MMI Estimation

The mutual information between an utterance X_i and all the models $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_M\}$ is defined as

$$I(X_i; \Lambda) \quad \approx \quad L_{ij} - \log \left(\frac{1}{M} \sum_{m=1}^{M} P_{im} \right) \tag{11}$$

assuming equal priors for all the M words. Thus, the MMI criterion is

$$R_{mmi} = \sum_i I(X_i; \Lambda) \quad \text{and} \tag{12}$$

$$\frac{\partial R_{mmi}}{\partial \overline{\omega_k}} = \sum_i \frac{\sum_{m=1}^{M} P_{im}(\Delta L_{ijk} - \Delta L_{imk})}{\sum_{m=1}^{M} P_{im}} \tag{13}$$

and the optimal weights may be found by a gradient-descent algorithm as in MCE training.

Notice that the estimation formulas, Eqn.(13) and Eqn.(7) for MMI and MCE training are very similar. The main difference is the additional term $\frac{\partial l}{\partial d}$ due to smoothing of the log-likelihood difference by the sigmoid function.

3.2. ML Estimation

In principle, the sub-band weights cannot be determined analytically by the maximum-likelihood method, since it will simply give all weights to the sub-band with the highest probability if we assume $\sum_{k=1}^{K} \omega_k = 1$. However, if one is willing to impose certain constraints on the weights, then ML estimation of sub-band weights is plausible. For example, in [6], J. Hernando suggested constraints in the form

$$\sum_{k=1}^{K} {\omega_k}^n = 1, \quad n > 1 \ . \tag{14}$$

The solution to this constrained ML criterion is

$$\omega_k = \frac{\left(\sum_i L_{ijk}\right)^{\frac{1}{n-1}}}{\left(\sum_{m=1}^{K} \left(\sum_i L_{ijm}\right)^{\frac{n}{n-1}}\right)^{\frac{1}{n}}} \ . \tag{15}$$

The main drawback of the method is that the constraint on sub-band weights is not well justified, and is imposed in an ad hoc manner.

4. RECOGNITION EXPERIMENTS AND RESULTS

We evaluated the various methods of finding sub-band weights on isolated TI digits [8]. Data from 55 male speakers were used for training and data from the remaining 56 male speakers were used for testing. Noisy data for testing were

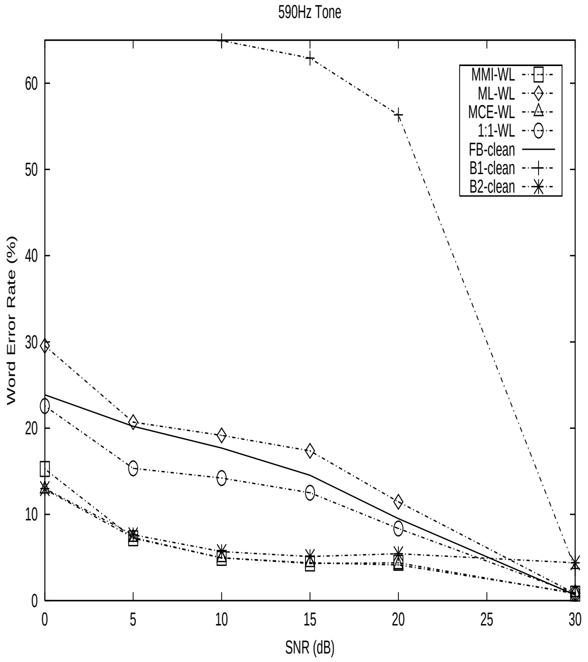

Figure 1: Recognition results with 590Hz tone: FB=full-band, B1=Band-1, B2=Band-2, MCE-WL=word-level likelihood combination with MCE-derived weights, MMI-WL=word-level likelihood combination with MMI-derived weights, ML-WL=word-level likelihood combination with ML-derived weights, 1:1-WL=word-level likelihood combination with equal weights

created from the corresponding clean speech by manually adding the noise at a prescribed signal-to-noise-ratio (SNR). Noisy speech were also created for training sub-band weights in a similar manner from a randomly picked one-fifth of the training data — of about 48 seconds. (In practice, one may pre-compute sub-band weights at various SNRs for each type of noises as in this paper, or they may be estimated online from adaptation data.)

Speech data were low-passed at 4400Hz and MFCCs were extracted from a window of 15ms at a frame rate of 100Hz. The full-band acoustic vector consists of 12 MFCCs and the normalized energy while a sub-band acoustic vector consists of 6 MFCCs and the normalized energy. Cepstral mean subtraction was performed as well.

All (full-band or sub-band) HMMs are left-right, whole-word models with 6 states and 4 mixture Gaussians per state. They are all trained with *clean* speech. Two sub-bands consisting of equal number of critical bands were used:

- Band-1: 100 – 1080 Hz
- Band-2: 1000 – 4400 Hz .

4.1. Experiment I: Tone at 590Hz

We first checked when only Band-1 was corrupted by a band-limited noise — a tone of 590Hz. Multi-band recognizers combined sub-band word likelihoods linearly before final decisions are made. The result is shown in Figure 1.

Table 1: Changes in sub-band weight of Band-1 as SNR decreases using 590Hz tone noise. ML^2 means n = 2 according to Eqn.(14)

Weights	0dB	5dB	10dB	15dB	20dB	30dB
MCE	0.159	0.143	0.138	0.167	0.200	0.487
MMI	0.210	0.212	0.137	0.123	0.119	0.500
ML^2	0.610	0.626	0.630	0.626	0.610	0.441

Since only Band-1 is corrupted, Band-2 maintains good performance under all SNRs while the full-band recognizer's accuracy degrades with lower SNR. Among the multi-band recognizers, performance is poor when their sub-band weights are derived by ML training or set equal, while MCE/MMI-derived weights give the highest accuracy. From the sub-band weights found by the various methods in Table 1, we see that MCE training is more effective than the other methods and gradually moves more weights from Band-1 — the corrupted band — to Band-2 as the SNR decreases.

The results serve as a theoretical upper bound on how good multi-band recognition can be under band-limited noises.

4.2. Experiment II: NOISEX-92

In reality, noises often spread over a wide spectrum. We performed a set of experiments with three noise types from the NOISEX-92 database: white, m109, and babble noise. White noise and babble noise are chosen because they are often encountered in real life, and m109 noise represents a more characteristic noise.

We have the following observations from the results shown in Figure 2–4.

- Performance of the full-band recognizer drops drastically once the SNR falls below 15dB.
- When the individual sub-band recognizers have similar performance, all recognizers, full-band or multi-band with weights trained by different methods, have similar accuracies.
- However, when the performance of the two sub-band recognizers diverges, the acclaimed advantage of multi-band recognizers sets in if the sub-band weights are trained with MMI or MCE: they effectively weigh the two bands to emphasize more on the more reliable band; and their performance is at least as good as that of the best sub-band recognizer.
- Sub-band weights derived from MMI or MCE training result in similar performance; MCE-trained weights give just slightly better performance.

5. CONCLUSION

Our experiments show that sub-band weights derived from MCE or MMI training can effectively emphasize the more reliable bands. Subsequent multi-band speech recognition by linear combination of sub-band likelihoods at word level reduces word-error-rate (WER) by 17.9% on average over the full-band recognizer trained with clean speech. In practice, one may train sub-band recognizers with clean speech, and either use simulated noisy speech to pre-compute the sub-band weights using MCE training method, or derive the weights from adaptation data online. We will expand our work to continuous speech recognition under noisy environments.

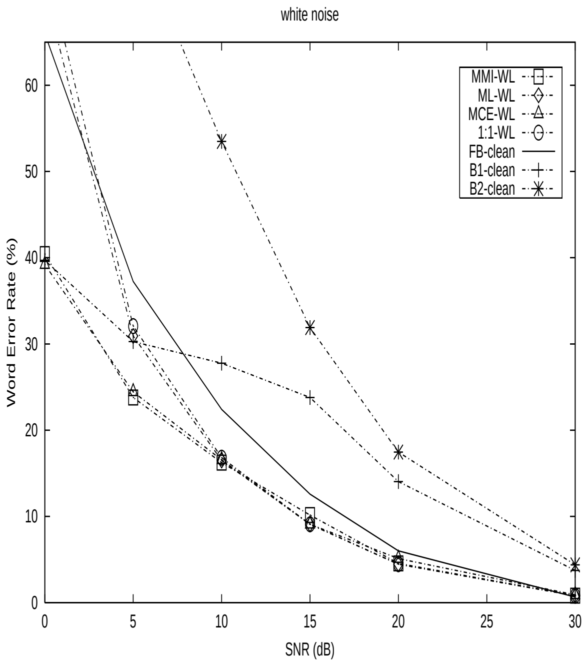

Figure 2: Recognition results with white noise.

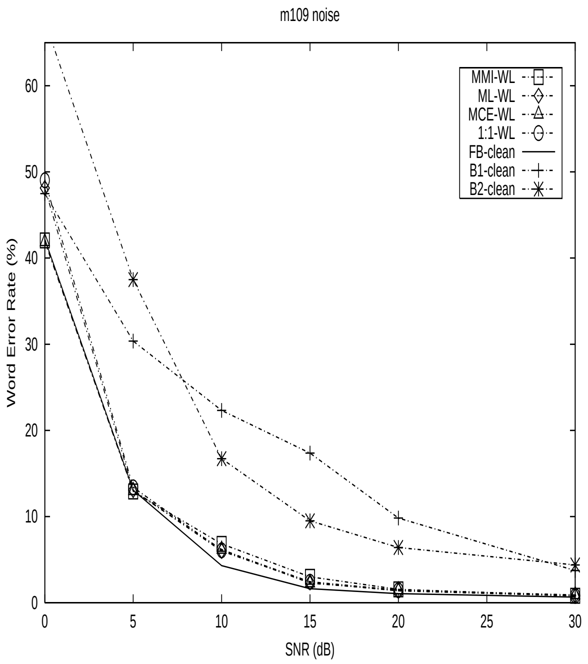

Figure 3: Recognition results with m109 noise.

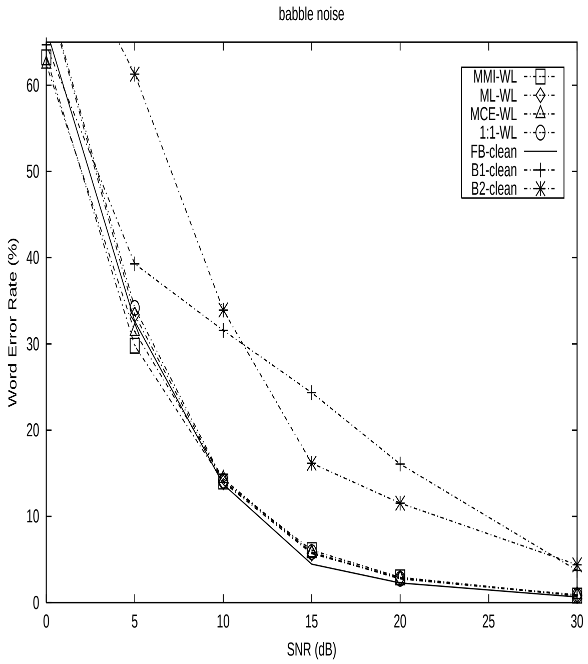

Figure 4: Recognition results with babble noise.

6. ACKNOWLEDGEMENTS

This work is supported by the Hong Kong RGC under the grant number CA97/98.EG02, and by the grant HKTIIT 98/99.EG01 from the Cable & Wireless HKT.

7. REFERENCES

[1] J. B. Allen. How Do Humans Process and Recognize Speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):567–577, 1994.

[2] H. Bourlard and S. Dupont. A New ASR Approach Based on Independent Processing and Recombination of Partial Frequency Bands. In *ICSLP*, October 1996.

[3] C. Cerisara, J.-F. Mari J.-P. Haton, and D. Fohr. A Recombination Model for Multi-band Speech Recognition. In *ICASSP*, volume II, pages 717–720, May 1999.

[4] Y.L. Chow. Maximum Mutual Information Estimation of HMM Parameters for Continuous Speech Recognition Using the N-Best Algorithm. In *ICASSP*, April 1990.

[5] H. Hermansky, S. Tibrewala, and M. Pavel. Towards ASR on Partially Corrupted Speech. In *ICSLP*, October 1996.

[6] J. Hernando. Maximum Likelihood Weighting of Dynamic Speech Features for CDHMM Speech Recognition. In *ICASSP*, pages 1267–1270, 1997.

[7] S. Okawa, E. Bocchieri, and A. Potamianos. Multiband Speech Recognition in Noisy Environments. In *ICASSP*, volume II, pages 641–644, May 1998.

[8] R.G. Leonard. A Database for Speaker-Independent Digit Recognition. In *ICASSP*, 1984.