

## MAP-Based Perceptual Modeling for Noisy Speech Recognition

YUNG-JI SHER<sup>1,3</sup>, YEOU-JIUNN CHEN<sup>4</sup>, YU-HSIEN CHIU<sup>5</sup>,  
KAO-CHI CHUNG<sup>1</sup> AND CHUNG-HSIEN WU<sup>2</sup>

<sup>1</sup>*Institute of Biomedical Engineering*

<sup>2</sup>*Department of Computer Science and Information Engineering  
National Cheng Kung University  
Tainan, 701 Taiwan*

<sup>3</sup>*Department of Physical Therapy  
Shu Zen College of Medicine and Management  
Kaohsiung, 821 Taiwan*

<sup>4</sup>*Department of Electrical Engineering  
Southern Taiwan University of Technology  
Tainan, 710 Taiwan*

<sup>5</sup>*Computer and Communications Research Laboratories  
Industrial Technology Research Institute  
Hsinchu, 310 Taiwan*

This study presents a maximum *a posteriori* (MAP) based perceptual modeling approach to deal with the issue of recognition degradation in noisy environment. In this approach, MAP-based noise detection is first applied to identify the noise segment in an utterance. Subtractive-type enhancement algorithm with masking properties of the human auditory system is then used to reduce the noise effect. Finally, MAP-based incremental noise model adaptation is developed to overcome the model inconsistencies between training and testing environments. For performance evaluation of the proposed approach, a Mandarin keyword recognition system was constructed. The experimental results show that the proposed approach achieves a better recognition rate compared to the audible noise suppression (ANS) and parallel model combination (PMC) methods.

**Keywords:** noisy speech recognition, speech enhancement, audible noise suppression, MAP-based perceptual modeling, noise detection, incremental model adaptation

### 1. INTRODUCTION

Robust speech processing plays an essential role in the development of human machine communication interface and has a wide range of applications in multi-modal/multi-media systems, mobile communication, car navigation systems, etc. In recent years, automatic speech recognition technologies have achieved promising performance. Unfortunately, while these systems are exposed to noisy environments, the performance degrades rapidly. The observed speech signals are subject to be distorted or noisy. For example, in car environment, these kinds of unwanted signal or interferences range over the acoustic background noise, radio-channel interference, vibration, moving cars, fans,

---

Received August 16, 2005; accepted January 17, 2006.  
Communicated by Jhing-Fa Wang, Pau-Choo Chung and Mark Billingham.

etc. Therefore, accurate modeling of noise patterns with both the temporal and spectral characteristics is the key task in this field.

Current research into noise reduction and distortion removal focuses on robust feature extraction, acoustic matching between training and recognition conditions, and speech enhancement [1-15]. Speech enhancement techniques in the front end intend to recover either the waveform or specific parameters of clean speech from noisy speech. This involves the transformation of noisy speech into as close an approximation of the training environment as possible. Spectral subtraction is a well-known family of enhancement algorithm. It attempts to estimate the short-time power spectrum of clean speech by explicitly subtracting noise estimation from the noisy speech assuming that the noise and speech are uncorrelated and additive. However, this method needs to be improved since it has a major drawback: the enhanced speech contains a “musical residual noise” with an unnatural structure. This perceptually annoying noise is composed of tones at random frequencies and has an increased variance [1-6]. Some methods have been recently developed with human perception [7-10].

Moreover, some model compensation methods have been developed to overcome the inconsistencies between training models and noisy input. The model parameters are trained and adapted in accordance with the noisy conditions. Parallel model combination (PMC) [11] is a representative approach. The noisy observation is modeled before recognition. However, for nonstationary noise modeled by hidden Markov models (HMMs), the optimal combination of speech and noise has to be done in the decoding stage. Another compensation method is Training Data Contamination [12]. The noise components experienced under test conditions are added to the training data in order to reduce the mismatch.

In this paper, a MAP-based perceptual speech model is proposed to deal with the issues of model inconsistencies. MAP-based noise detection is applied to identify the noise segment for accurate noise estimation. The auditory masking effect in the human perception phenomenon and the MAP-based adaptation algorithm are used to incrementally adapt the noise model for improving the recognition rate.

The rest of this paper is organized as follows. The next section gives theoretical details of the proposed MAP-based perceptual modeling framework. Section 3 summarizes experimental results. Finally, section 4 draws conclusions and offers suggestions for future works.

## 2. MATERIALS AND METHODS

The framework for MAP-based perceptual modeling approach is shown in Fig. 1. First, the noise is detected by MAP-based noise detection. For the noise segment, noise spectrum can be estimated and used to evaluate the auditory masking threshold. Then, an enhanced speech can be calculated using perceptual weighting based speech enhanced algorithm with auditory masking threshold and segment information. Incremental noise model adaptation is suitable for both recognition and verification of speech signals in noisy environment.

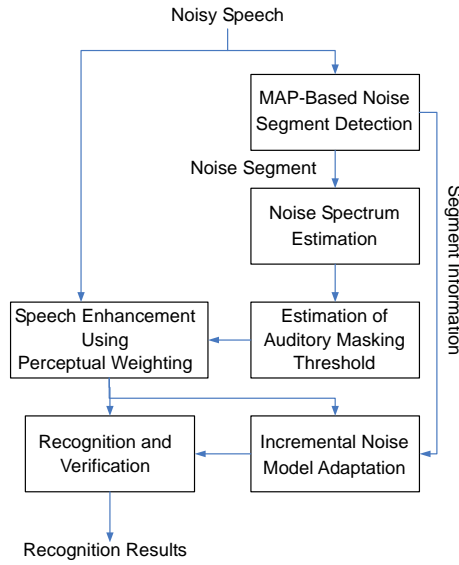


Fig. 1. The framework for MAP-based perceptual modeling approach.

## 2.1 Perceptual Speech Enhancement

### 2.1.1 MAP-based noise detection

For a speech utterance  $S$  with  $L_s$  feature vectors, it can be segmented into a sequence of speech segments denoted as

$$S = \{s_1, s_2, \dots, s_{B+1}\}, \quad (1)$$

where  $B$  is the number of boundaries. Moreover, the corresponding speech type sequence  $N^s$  and position information of boundaries  $R^s$  can be denoted as

$$N^s = \{n_1^s, n_2^s, n_3^s, \dots, n_B^s, n_{B+1}^s\}, \quad (2)$$

and

$$R^s = \{r_1^s, r_2^s, r_3^s, \dots, r_B^s\}, \quad (3)$$

where  $1 < r_1^s < r_2^s < r_3^s < \dots < r_B^s < L_s$  and  $n_i^s$  is the speech type for speech segment  $s_i$ .  $\forall n_i^s \in \hat{N} = \{\hat{N}_1, \hat{N}_2, \dots, \hat{N}_K\}$  in which  $\hat{N}_i$  is the  $i$ -th predefined speech type label and  $K$  is the number of speech types. The speech type is the classification of noise and speech. The noise detection can be estimated by MAP as shown below:

$$\begin{aligned} \max P(N^s, R^s, B | S) &= \max \frac{P(N^s, R^s, S | B)P(B)}{P(S)} \\ &\approx \max P(N^s, R^s, S | B) \\ &= \max P(S | N^s, R^s, B)P(N^s | R^s, B)P(R^s | B), \end{aligned} \quad (4)$$

where  $P(R^s | B)$  is the conditional probability of position information of boundaries,  $R^s$ , given the number of boundaries,  $B$ .  $P(N^s | R^s, B)$  is the conditional probability of speech type sequence,  $N^s$ , given  $R^s$  and  $B$ .  $P(S | N^s, R^s, B)$  is the conditional probability of speech utterance  $S$  given  $N^s$ ,  $R^s$ , and  $B$ .

Since there is no prior knowledge for boundary positions and no permutation of positions by the constraint,  $1 < r_1^s < r_2^s < r_3^s < \dots < r_B^s < L_s$ , a uniform distribution is considered to estimate the conditional probability, as follows:

$$P(R^s | B) = \frac{1}{\binom{L_s - 2}{B}}, \quad (5)$$

where  $\binom{L_s - 2}{B}$  means that  $L_s - 2$  possible positions are taken  $B$  positions for speech segment boundaries. For the conditional probability,  $P(N^s | R^s, B)$ , speech type sequence,  $N^s$ , is independent of the speech position information of boundaries,  $R^s$ . Thus, a multinomial distribution estimation of speech type sequence is adopted and shown as

$$\begin{aligned} P(N^s | R^s, B) &= P(N^s | B) = P(n_1^s, n_2^s, n_3^s, \dots, n_B^s, n_{B+1}^s | B) \\ &= \binom{L_s}{l_1^s, l_2^s, \dots, l_B^s, l_{B+1}^s} P(n_1^s)^{l_1^s} P(n_2^s)^{l_2^s} \dots P(n_{B+1}^s)^{l_{B+1}^s}, \end{aligned} \quad (6)$$

where  $l_i^s$  is the number of feature vectors for the  $i$ -th speech segment,  $s_i$ .  $P(n_i^s)$  is the probability that speech segment  $s_i$  is  $n_i^s$  and used to represent the *a priori* probability of speech type. It can be estimated from the training corpus by maximum likelihood estimation as

$$P(n_i) = \frac{C(s_j | s_j \in n_i)}{C(s_j)}, \quad (7)$$

where  $C(s_j)$  is the frequency of segments in the training corpus and  $C(s_j | s_j \in n_i)$  is the frequency of segments with speech type  $n_i$ . According to the boundary number,  $B$ , and position information of boundaries,  $R^s$ , the speech utterance  $S$  can be divided into a segment sequence  $S = \{s_1, s_2, \dots, s_{Q+1}\}$ . Thus,  $P(S | N^s, R^s, B)$  can be simplified as

$$\begin{aligned} P(S | N^s, R^s, B) &= P(s_1, s_2, \dots, s_B, s_{B+1} | N^s) \\ &= P(s_1, s_2, \dots, s_{B+1} | n_1^s, n_2^s, n_3^s, \dots, n_B^s, n_{B+1}^s). \end{aligned} \quad (8)$$

The approximation of Eq. (8) is based on the assumption that the probabilities of each speech segments  $s_i$  labeled as speech type  $n_i^s$  are independent. Therefore, Eq. (8) can be represented as

$$P(s_1, s_2, \dots, s_B, s_{B+1} | n_1^s, n_2^s, n_3^s, \dots, n_B^s, n_{B+1}^s) = \prod_{i=1}^{B+1} P(s_i | n_i^s). \quad (9)$$

Consider that a speech segment  $s_i$  can be represented as a vector sequence,  $s_i =$

$o_1^{s_i} o_2^{s_i} \dots o_{m_i^s}^{s_i} \cdot P(S_i | n_i^s)$  can be written as

$$P(s_i | n_i^s) = P(o_1^{s_i} o_2^{s_i} \dots o_{m_i^s}^{s_i} | n_i^s) = \prod_{j=1}^{m_i^s} P(o_j^{s_i} | n_i^s). \quad (10)$$

Based on the acoustic characteristics of sounds, a speech type can be modeled by different acoustic properties.  $P(o_j^{s_i} | n_i^s)$  can be written as

$$P(o_j^{s_i} | n_i^s) = \left( \prod_{k=1}^{K_{n_i^s}} P(o_j^{s_i} | M_k^{n_i^s}) \right)^{\frac{1}{K_{n_i^s}}}, \quad (11)$$

where  $K_{n_i^s}$  is the number of models in speech type  $n_i^s$ .  $M_k$  is the acoustic model clustered in speech type  $n_i^s$  and can be modeled by Hidden Markov Models, Gaussian Mixture Models, etc.

### 2.1.2 Estimation of auditory masking threshold

Masking effects are usually described by a masking threshold function [10, 16]. In this paper, the mel-scale, which corresponds to the auditory sensation of tone height, is adopted instead of critical bands. After the mapping process of spreading function, which takes the masking effect over several adjacent critical bands into accounts, an alternative spectral flatness measure (SFM) function to further determine which speech frame is close to either noise-like or tone-like is proposed and defined as

$$SFM_{dB} = 10 \log_{10} \left\{ \frac{\exp \left( \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} \log |E_t(\omega)|^2 d\omega \right)}{\frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} |E_t(\omega)|^2 d\omega} \right\}, \quad (12)$$

where  $E_t(\omega)$  is defined as

$$E_t(\omega) = Y_t(\omega) - \hat{Y}(\omega). \quad (13)$$

$\hat{Y}(\omega)$  is the spectrum resulted from applying a low-pass filter over the noisy speech  $Y_t(\omega)$ . The basic idea of maximizing the spectral flatness is used to eliminate the prediction error in glottal inverse filtering process. But this is difficult to achieve automatically. We adopt the median filter as an alternative. The SFM was defined as the ratio of the geometric mean to the arithmetic mean of the power spectrum. The SFM was used to generate the ‘‘tonality factor’’  $\gamma$  that aids in selecting the appropriate masking threshold for each frame:

$$\gamma = \min \left( \frac{SFM_{dB}}{SFM_{dB \max}}, 1 \right), \quad (14)$$

where  $SFM_{db\ max}$  is  $-60$ dB. It is used to estimate that the signal is entirely tone like and an SFM with  $0$ dB is used to indicate a signal that is completely noise-like.

The resulted threshold is compared with the normalized,  $T_N(i)$ , and absolute auditory threshold,  $T_A(i)$ , [5]:

$$T_F(i) = \max\{T_N(i), T_A(i)\}, \quad (15)$$

where  $T_A(i)$  can be represented as

$$T_A(i) = 3.64 \left( \frac{i}{1000} \right)^{-0.8} - 6.5e^{-0.6 \left( \frac{i}{1000} - 3.3 \right)^2} + 10^{-3} \left( \frac{i}{1000} \right)^4. \quad (16)$$

The normalized auditory threshold is shown as follows:

$$T_N(i) = \frac{10^{\log_{10}(C_i) - \frac{\gamma(14.5+i) + 5.5(1-\gamma)}{10}}}{P_i}, \quad (17)$$

where  $P_i$  is the number of points in critical band  $i$ .  $C_i$  is the spread spectrum of the  $i$ -th critical band:

$$C_i = B_i \cdot SF_i, \quad (18)$$

where  $B_i$  is the power energy of the  $i$ -th critical band.  $SF_i$  is the spreading function and used to represent the masking between signals in different critical bands. A model that approximates the basilar membrane spreading function is defined as [16]:

$$10\log_{10} SF_i = 15.91 + 7.5 \cdot (i + 0.474) - 17.5 \cdot \sqrt{1 + (i + 0.474)^2}. \quad (19)$$

Fig. 2 shows the power spectrum, AMT and the absolute hearing threshold for Mandarin speech/ling/(zero).

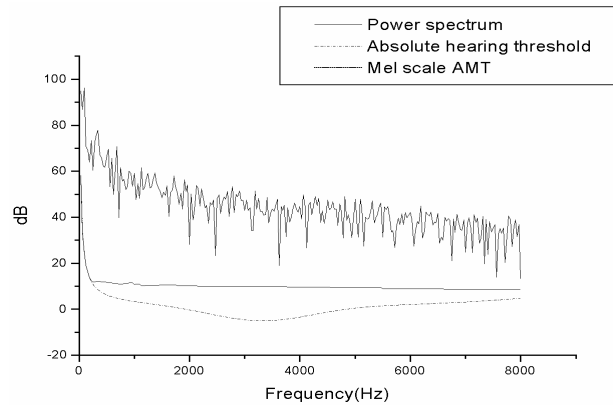


Fig. 2. Comparison of the power spectrum, AMT and the absolute hearing threshold.

### 2.1.3 Perceptual weighting

The traditional approach for estimating the clean speech is spectral subtraction. Subtractive-type algorithm can be carried out using a second approach: filtering of noise speech with a time-varying linear filter dependent on the characteristics of the noisy signal spectrum and on the estimated noise spectrum. The noise suppression process becomes a multiplication of the short-time spectral magnitude of the noisy speech  $|Y(\omega)|$  by a gain function  $G(\omega)$  and an enhanced speech short-time magnitude  $|\hat{S}(\omega)|$  can be obtained by

$$|\hat{S}(\omega)| = G(\omega) |Y(\omega)|. \quad (20)$$

The filter for power spectral subtraction is given by

$$G(\omega) = \begin{cases} \left(1 - \frac{|N(\omega)|^2}{|Y(\omega)|^2}\right)^{\nu(\omega)} & \text{if } |Y(\omega)|^2 - \alpha |N(\omega)|^2 > k^{1/\nu(\omega)} |Y(\omega)|^2, \\ k & \text{otherwise} \end{cases}, \quad (21)$$

where  $|N(\omega)|$  is the noise spectral magnitude estimated from the results of MAP-based noise segment detection.  $k$  is a minimum constant for avoiding the negative power spectrum. The transition shape of spectral filtering is controlled by  $\nu(\omega)$ . Obviously, large  $\nu(\omega)$  implies a sharp transition [1, 5]. Motivated by this concept, the auditory masking effect is adopted as the non-linear spectral modification. Since frequency components below the masking threshold are inaudible, the audible noise power spectrum can be expressed as:

$$A_n(\omega) = A_y(\omega) - A_s(\omega). \quad (22)$$

Examination of the expression for the audible noise spectrum implies that  $A_n(\omega)$  is given by

$$A_n(\omega) = \begin{cases} Y(\omega) - \hat{Y}(\omega) & \text{if } Y(\omega) > T(\omega) \text{ and } \hat{Y}(\omega) > T(\omega) \\ Y(\omega) - T(\omega) & \text{if } Y(\omega) > T(\omega) \text{ and } \hat{Y}(\omega) < T(\omega) \end{cases}. \quad (23)$$

## 2.2 Recognition and Verification

### 2.2.1 Incremental noise model adaptation

Using HMMs for estimation of an input signal  $x(t)$  observed in an additive noise  $y(t)$ , the posterior pdf from Bayes' rule can be defined as follows [17]:

$$f_{X|Y}(x(t) | y(t)) = \frac{f_{Y|X}(y(t) | x(t))}{f_Y(y(t))} f_X(x(t)) = \frac{1}{f_Y(y(t))} f_N(y(t) - x(t)) f_X(x(t)). \quad (24)$$

For a given observation,  $f_Y(y(t))$  is a constant and the maximum *a posteriori* (MAP) estimate is obtained as follows:

$$\hat{x}^{MAP}(t) = \arg \max_{x(t)} f_N(y(t) - x(t)) f_X(x(t)). \quad (25)$$

The MAP estimation of Eq. (25) requires the pdf models of the signal and the noise process. Stationary and continuous-valued processes are often modeled by a Gaussian or a mixture Gaussian pdf that is equivalent to a signal-state HMM. For a non-stationary process an  $n$ -state HMM can model the time-varying pdf of the process as a Markov chain of  $n$  stationary Gaussian subprocesses. Now assume that there is an  $n$ -state HMM,  $\pi$ , for signal and one state HMM,  $\eta$ , for noise. Given the observed state sequence  $Y$ , the noise model can be estimated as:

$$s_{\text{signal}}^{\text{MAP}} = \arg \max_{s_{\text{signal}} \ s_{\text{noise}}} (\max_{s_{\text{signal}} \ s_{\text{noise}}} f_Y(Y, s_{\text{signal}}, s_{\text{noise}} | \pi, \eta)), \quad (26)$$

and

$$s_{\text{noise}}^{\text{MAP}} = \arg \max_{s_{\text{noise}} \ s_{\text{signal}}} (\max_{s_{\text{noise}} \ s_{\text{signal}}} f_Y(Y, s_{\text{signal}}, s_{\text{noise}} | \pi, \eta)). \quad (27)$$

In the MAP-based noise detection, the speech and noise segments are detected and used to adapt the initial noise model as follows:

$$\hat{\mu}_n = \frac{\sum_{i=1}^q y_i}{\tau + q}, \quad (28)$$

where  $\tau$  is a balancing factor between prior mean and the estimated mean and  $q$  is the number of the estimated noise frames. Therefore, the MAP signal estimate and the adaptive noise threshold  $\theta$  are given by:

$$\hat{x}(m) = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_n^2} (y(m) - \mu_n) + \frac{\sigma_n^2}{\sigma_x^2 + \sigma_n^2} \mu_x, \quad (29)$$

and

$$\theta = \frac{\mu_x + \mu_n}{2}. \quad (30)$$

For a noise-free signal  $\mu_n = \sigma_n^2 = 0$ , we have  $\hat{x}(m) = y(m)$ .

### 2.2.2 Keyword recognition

In the recognition process, the Viterbi algorithm is employed to find the most likely keyword  $W_k$ , where

$$W_k = \arg \max_j L(O | W_j), \quad (31)$$

and  $L(O | W_j)$  is the likelihood of the observation sequence  $O$  given word  $W_j$ . In the context of subsyllable recognition,  $W_k$  is a concatenation of subsyllable units that can be written as

$$W_k = ss_1^{(k)} ss_2^{(k)} \dots ss_N^{(k)}, \quad (32)$$

where  $N$  is the number of subsyllables, and the subsyllable string  $ss_1^{(k)} ss_2^{(k)} \dots ss_N^{(k)}$  is the subsyllable lexical representation of keyword  $W_k$ .

### 2.2.3 Keyword verification

Keyword verification can be treated as the problem of statistical hypothesis testing. The null hypothesis,  $H_0$ , represented by the input speech containing a given keyword  $K_i$  is tested against the alternative hypothesis,  $H_1$ , that  $K_i$  does not exist. According to the Neyman-Pearson Lemma [18], the optimal test is the likelihood ratio test such that the null hypothesis,  $H_0$ , is accepted if

$$R(O; K_i) = \frac{L(O | H_0)}{L(O | H_1)} > \gamma_i, \quad (33)$$

where  $\gamma_i$  is the critical threshold of the test. As a result, two types of errors can occur: false rejection (Type I) and false acceptance or false alarms (Type II) errors.

For a subsyllable  $ss_n^{(k)}$  of keyword  $k$ , the normalized confidence measure is defined as

$$LR(O_{t_{n-1}}^{t_n}; ss_n^{(k)}) = \frac{1}{T_n^{(k)}} \log \left[ L(O_{t_{n-1}}^{t_n} | ss_n^{(k)}) \right] - \frac{1}{T_n^{(k)}} \log \left[ L(O_{t_{n-1}}^{t_n} | \overline{ss}_n^{(k)}) \right], \quad (34)$$

where  $\overline{ss}_n^{(k)}$  is the anti-syllable model of  $ss_n^{(k)}$  and  $T_n^{(k)}$  is the number of frames allocated for subsyllable  $ss_n^{(k)}$ . For an  $N$ -subsyllable string  $ss_1^{(k)} ss_2^{(k)} \dots ss_N^{(k)}$  corresponding to the most likely keyword  $W_k$ , the whole word phonetic verification function is defined as follows:

$$D(O; W_k) = \log \left[ \frac{1}{N} \sum_{n=1}^N \exp \left[ -\eta \cdot LR(O_{t_{n-1}}^{t_n}; ss_n^{(k)}) \right] \right]^{\frac{1}{\eta}}, \quad (35)$$

where  $\eta$  is a positive constant, and  $N$  is the number of subsyllable of the signal.

## 3. EXPERIMENTAL RESULTS

### 3.1 Acoustic Model Training

Among the database TCC 300 [19] collected in Taiwan area, 2676 sentence were selected and used as the training database. 817 words were defined as the keyword set. The testing speech database was pronounced by 14 speakers (12 males, 2 female) in clean, street, and car environments (70km/hr and 90km/hr in high way). In order to evaluate the performance for noise segment detection, 539 collected speech utterances in clean, street, and car environment were manually tagged with boundary information. The baseline system was built without auditory masking processing. A 26-dimension feature

vector with 12 Mel-Frequency Cepstrum Coefficients (MFCCs), 12 delta MFCCs, one delta log energy, and one delta delta log energy were adopted as features.

In Mandarin speech, each syllable can be phonetically decomposed into an INITIAL followed by a FINAL. The INITIAL of a syllable is optional and comprises a single consonant if it exists. The FINAL comprises a vowel or diphthong nucleus preceded by an optional medial, followed by an optional nasal. Based on the linguistic information of Mandarin, there are 21 INITIALS and 38 FINALS. The *K*-means clustering algorithm is used to perform group clustering based on minimizing the overall inter-subsyllable group distance. Thus, the INITIALS and FINALS can be clustered into 3 groups and 9 groups, respectively. The constituency of each group is given in Table 1. Based on those acoustic characteristic, 100 right context-dependent INITIAL and 37 context-independent FINAL HMMs were used to construct the keyword recognizer. Each INITIAL HMM consists of 3 states and each FINAL HMM consists of 5 states, each with 10 Gaussian mixture densities. In general, for each subsyllable model, a corresponding anti-subsyllable model was also constructed specifically for the word verification task [13].

**Table 1. The constituency of subsyllable groups for (a) FINALS and (b) INITIALS.**

(a)		(b)	
Group 1	iai	Group 1	p, g, k, h
Group 2	uei	Group 2	f, ji, chi, shi, j, ch, sh, tz, ts, s
Group 3	eng, iou, iung	Group 3	b, m, d, t, n, l, r
Group 4	ai, an, uai, uan, uen, uang		
Group 5	i, in, ing		
Group 6	o, ou, u, ua, uo, ueng		
Group 7	au, e, er		
Group 8	a, ang, ia, iau, iang		
Group 9	null, ei, en, ie, iou, iu, iue, iuan, iun		

### 3.2 Experiments on the Noise Segment Detection

In this experiment, the proposed MAP-based noise segment detection was adopted for the identification of noise speech. Delta Bayesian information criterion (BIC) approaches was adopted for comparison [20]. Considering the variations in manually tagged boundaries, a tolerated boundary that the detected boundary falls into a predefined interval within an offset of 10 frames against manually tagged noise segment boundaries was defined. For the performance evaluation of noise segmentation, the harmonic mean *F* criterion was adopted [21]. This evaluation index also takes the precision and the recall rates into accounts [22] and is shown as follows:

$$F = 2 / \left( \frac{1}{r_p} + \frac{1}{r_r} \right), \quad (36)$$

in which  $r_p$  and  $r_r$  represent the precision and the recall rates, respectively. *F* ranges from 0 to 1 and the greater *F* value represents the better accuracy of boundary detection. The detailed formulas are given as follows:

$$r_p = \frac{|R_a|}{|R|} \times 100\%, \quad (37)$$

$$r_r = \frac{|R_a|}{|A|} \times 100\%, \quad (38)$$

in which  $|A|$  is the number of total boundaries manually tagged in our collection;  $|R|$  is the number of total detected noise segment boundaries;  $|R_a|$  is the number of valid noise segment boundaries to be detected. The approach of delta BIC with penalty weight 0.7 and window size 0.5s can achieve a satisfactory result which the precision rate, recall rate and  $F$  value were 0.68, 0.41, and 0.51, respectively. Using MAP-based noise segment detection, the precision rate, recall rate, and  $F$  value were 0.75, 0.44, and 0.55, respectively. The proposed approach outperformed than the delta BIC approach. However, some problems still arise in noise segment in different noisy environments. The most occurred problem contain shot noise, hubbubs, and vibration in either street or car environment.

### 3.3 Performance Evaluation and Comparison

In this experiment, the optimal value of the parameter,  $v(\omega)$ , used in auditory noise suppression was examined. Table 2 show the experiment results in choosing optimal transient factor with respect to the average recognition rate in clean, street, and car environments. Given  $v(\omega) = 1$ , the system achieved the best recognition performance and was adopted in the following experiments.

**Table 2. Experimental results (%) for choosing transient factor  $v(\omega)$ .**

$v(\omega)$	0.5	1	2
Top 1	56.43	59.37	58.77
Top 2	65.61	67.58	64.27

The performances for parallel model combination (PMC) and auditory noise suppression (ANS) proposed by Gales [11] and Virag [10] were compared using Receiver Operator Characteristic (ROC) plots. Fig. 3 shows the comparison of verification performance in different false alarm rates and false rejection rates. It is clear that our proposed approach outperformed the others. From this figure, the performance of PMC and that of ANS are similar. Furthermore, for the proposed approach in this paper, the false alarm rate achieved the lowest value, 32.73%, at 7.9% false rejection. Fig. 4 shows the keyword recognition rates for the baseline, PMC, ANS and the proposed approach in different environments. The average SNRs for clean, street, car with 70km/hr, and car with 90 km/hr environments are 20.1dB, 9.6dB, 10.3dB and 6.4dB, respectively. The experimental results show our proposed method achieves a better recognition rate.

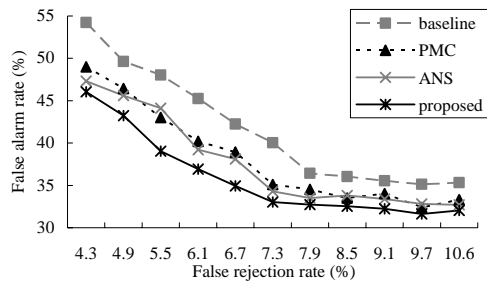


Fig. 3. Experimental results of keyword verification.

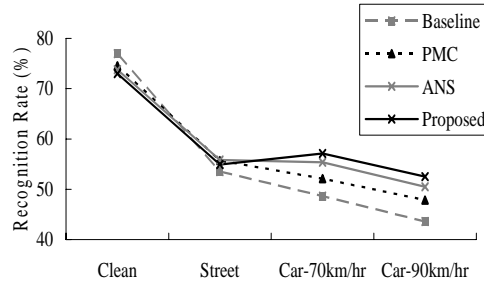


Fig. 4. Keyword recognition rates for different environments.

#### 4. CONCLUSION

This study proposed a MAP-based perceptual modeling approach to deal with the model inconsistency issues. The weighting method based on the masking effect was used in speech enhancement. Instead of using the complex inverse filtering criterion, an alternative median filter based spectral flatness measure was conducted. Combining with a proposed speech segmentation approach for noise segment detection, a MAP-based model adaptation scheme was adopted to incrementally adapt the noise model. Experimental results show the proposed perceptual modeling and incremental noise adaptation approaches are very promising in speech enhancement and achieves a better recognition rate than the traditional approaches.

#### REFERENCES

1. S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, Signal Processing*, Vol. ASSP-27, 1979, pp. 113-120.
2. M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Process*, 1979, pp. 208-211.
3. R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft decision noise suppression filter," *IEEE Transactions on Acoustics, Speech, Signal Processing*, Vol. ASSP-28, 1980, pp. 137-145.
4. Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, Signal Processing*, Vol. ASSP-32, 1984, pp. 1109-1121.
5. P. Lockwood and J. Boudy, "Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and projection, for robust recognition in cars," *Speech Communication*, Vol. 11, 1992, pp. 215-228.
6. J. H. L. Hansen, "Morphological constrained feature enhancement with adaptive cepstral compensation (MCE-ACC) for speech recognition in noise and Lombard effect," *IEEE Transactions on Speech Audio Processing*, Vol. 2, 1994, pp. 598-614.
7. D. Tsoukalas, M. Paraskevas, and J. Mourjopoulos, "Speech enhancement using

- psycho-acoustic criteria,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Process*, 1993, pp. 359-361.
8. T. Usagawa, M. Iwata, and M. Ebata, “Speech parameter extraction in noisy environment using a masking model,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Process*, Vol. 2, 1994, pp. 81-84.
  9. S. Nandkumar and J. H. L. Hansen, “Dual-channel iterative speech enhancement with constraints on an auditory-based spectrum,” *IEEE Transactions on Speech and Audio Processing*, Vol. 3, 1995, pp. 22-34.
  10. N. Virag, “Signal channel speech enhancement based on masking properties of the human auditory system,” *IEEE Transactions on Speech and Audio Processing*, Vol. 7, 1999, pp. 126-137.
  11. M. J. F. Gales and S. J. Young, “Cepstral parameter compensation for HMM recognition in noise,” *Speech Communication*, Vol. 12, 1993, pp. 231-239.
  12. B. Dautrich, L. Rabiner, and T. Martin, “On the effects of varying filter-bank parameters on isolated word recognition,” *IEEE Transactions on Acoustics, Speech, Signal Processing*, Vol. ASSP-31, 1992, pp. 793-806.
  13. C. H. Wu and Y. J. Chen, “Recovery of false rejection using statistical partial pattern trees for sentence verification,” *Speech Communication*, Vol. 43, 2004, pp. 71-88.
  14. C. H. Wu and Y. J. Chen, “Multi-keyword spotting of telephone speech using fuzzy search algorithm and keyword-driven two-level CBSM,” *Speech Communication*, Vol. 33, 2001, pp. 197-212.
  15. C. H. Wu, Y. J. Chen, and G. L. Yan, “Integration of phonetic and prosodic information for robust utterance verification,” *IEE Proceedings – Vision, Image and Signal Processing*, Vol. 147, 2000, pp. 55-61.
  16. M. R. Schroeder, B. S. Atal, and J. L. Hall, “Optimizing digital speech coders by exploiting masking properties of the human ear,” *Journal of the Acoustical Society of America*, Vol. 66, 1979, pp. 1647-1651.
  17. S. V. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*, John Wiley & Sons, Ltd., England, 2000.
  18. K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, New York, 1972.
  19. The Association for Computational Linguistics and Chinese Language Processing, <http://www.aclclp.org.tw>.
  20. C. H. Wu, Y. H. Chiu, C. J. Shia, and C. Y. Lin, “Automatic segmentation and identification of mixed-language speech using delta-BIC and LSA-based GMMs,” *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 14, 2006, pp. 266-276.
  21. R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison-Wesley, New York, 1999.
  22. M. Cettolo and A. Federico, “Model selection criteria for acoustic segmentation,” in *Proceedings of the ISCA Tutorial and Research Workshop on Automatic Speech Recognition*, 2000, pp. 221-227.



**Yung-Ji Sher (余永吉)** received the B.S. degree in Department of Physical Therapy from National Yang Ming Medical College, Taipei in 1992, and the M.S. and the Ph.D. degree in Institute of Biomedical Engineering from National Cheng Kung University in 1997 and 2006. He is presently an assistant professor and Head of Department of Physical Therapy in Shu Zen College of Medicine and Management. His research interests include modern literal Taiwanese, natural language processing, computational linguistic, speech signal processing, speech communication, assistive technology, biomedical engineering, physical therapy, digital signal processing, and rehabilitation engineering.



**Yeou-Jiunn Chen (陳有圳)** was born in Taichung, Taiwan, 1972. He received the B.S. degree in Mathematics from Tatung Institute of Technology in 1995 and Ph.D. degrees in Institute of Information Engineering from National Cheng Kung University in 2000. From 2001 to 2005, he was a research at Advanced Technology Center, Computer and Communications Laboratories, Industrial Technology Research Institute. He is presently an assistant professor in the Department of Electrical Engineering at Southern Taiwan University of Technology. His current research interests include speech enhancement, speech conversion, speech recognition, and biomedical signal processing.

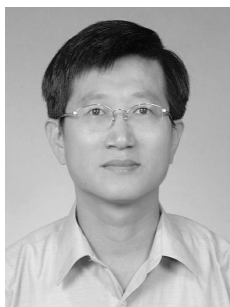


**Yu-Hsien Chiu (邱毓賢)** received the B.S. degree in Electrical Engineering from I-Shou University, Kaohsiung, Taiwan, R.O.C., in 1997, the M.S. degree in Biomedical Engineering from National Cheng Kung University, Tainan, Taiwan, in 1999, and the Ph.D. degree in Department of Computer Science and Information Engineering from National Cheng Kung University in 2002. He is presently an engineer in the Computer and Communications Laboratories, Industrial Technology Research Institute. His research interests include speech and biomedical signal processing, embedded system design, spoken language processing and sign language processing for the hearing-impaired.



**Kao-Chi Chung (鍾高基)** received the B.S. degree in Mathematics from National Cheng Kung University, Tainan in 1973, the M.S. degree in Mathematics from SUNY at Buffalo in 1979, the M.S. degree in Computer Science/Applied Mathematics and the Ph.D. in Biomedical Engineering from University of Virginia in 1981 and 1987. Since August 1994, he has been with

National Cheng Kung University as an associate professor in the Institute of Biomedical Engineering. His research interests include assistive technology, biomedical engineering, rehabilitation engineering, orthopedic biomechanics, soft tissue engineering, digital signal processing, speech signal processing and communication, medical instrumentation.



**Chung-Hsien Wu (吳宗憲)** received the B.S. degree in Electronics Engineering from National Chiao Tung University, Hsinchu, Taiwan, in 1981, and the M.S. and Ph.D. degrees in Electrical Engineering from National Cheng Kung University, Tainan, Taiwan, R.O.C., in 1987 and 1991, respectively. Since August 1991, he has been with the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan. He became a professor in August 1997. From 1999 to 2002, he served as the Chairman of the Department. He also worked at Massachusetts Institute of Technology Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, in summer 2003 as a visiting scientist. He is currently the Editor-in-Chief for International Journal of Computational Linguistics and Chinese Language Processing. His research interests include speech recognition, text-to-speech, multimedia information retrieval, spoken language processing and sign language processing for hearing-impaired. Dr. Wu is a senior member of IEEE. He is also a member of International speech communication association (ISCA) and ROCLING.