Static and Dynamic Spectral Features: Their Noise Robustness and Optimal Weights for ASR

Author : *Chen Yang, Frank K. Soong and Tan Lee*

Professor: 陳嘉平

Reporter: 吳國豪

Outline

- Introduction
- Noise Robustness Analysis For Cepstral Features
- Weighting of Dynamic and Static Features
- Discriminative Training of Feature Weights
- Experiments

Introduction

- In this paper, we investigate the relative noise robustness of **dynamic** and **static** spectral features in speech recognition.

- The **optimal weights** are discriminatively trained with a small amount of development data.

Introduction

- The proposed approach is appealing to practical applications because

  1) noise estimation is not required.

  2) model adaptation is not required.

  3) only a minor modification of the decoding process is needed.

  4) only a few feature weights need to be trained.

Noise Robustness Analysis For Cepstral Features

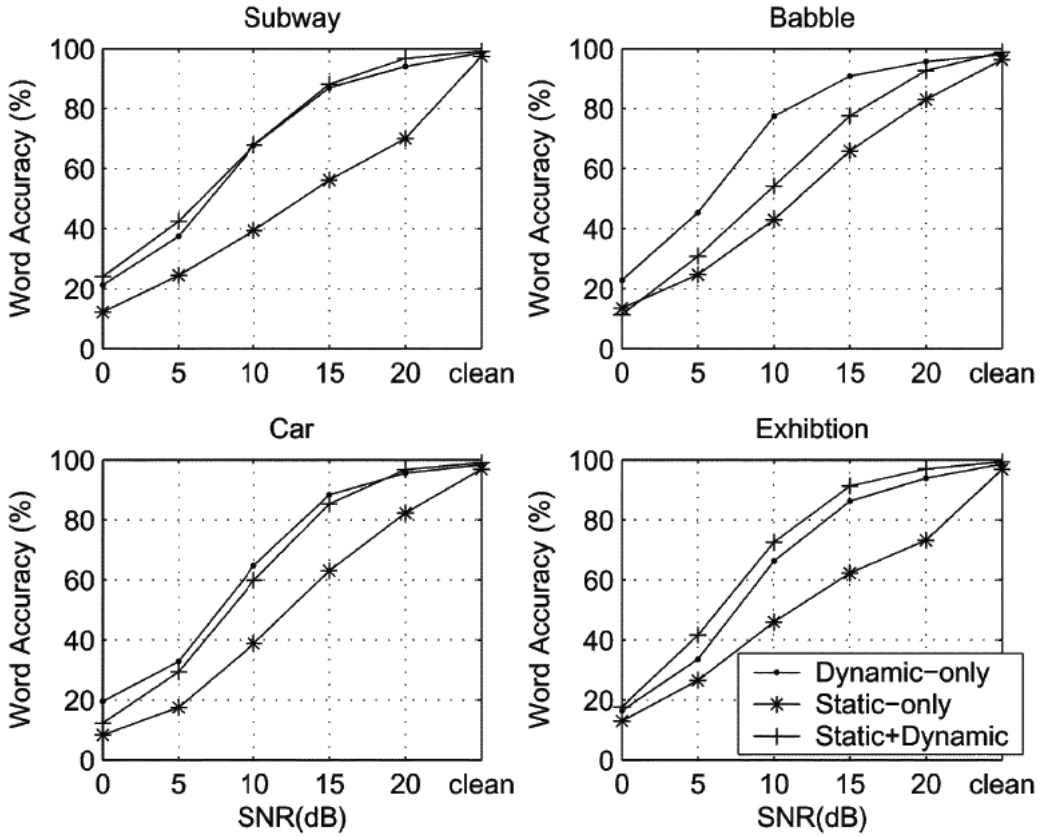

Fig. 1. Performance comparison of "static-only," "dynamic-only," and "static + dynamic" systems (Aurora 2).

Noise Robustness Analysis For Cepstral Features

- To simplify the analysis, we assume that each HMM state is represented by a single Gaussian probability density function (pdf). At a particular state *j*, the acoustic probability of the noisy observation y_t is computed as

$$b_j(y_t) = \frac{1}{\sqrt{(2\pi)^D \left|\sum_j\right|}} e^{-\frac{1}{2}(y_t-\mu_j)' \sum_j^{-1}(y_t-\mu_j)} \qquad (1)$$

  - x_t denote a clean speech feature vector at time *t* and y_t be the corresponding noise-corrupted version.
  - D is the dimension of the feature vector, μ_j and Σ_j denote, respectively, the mean vector and the covariance matrix of the Gaussian pdf of that state.

Noise Robustness Analysis For Cepstral Features

- The mismatch between y_t and the clean model is reflected by the exponent term in (1), which can be rewritten as

$$
\begin{aligned}
&(y_t-\mu_j)'\Sigma_j^{-1}(y_t-\mu_j)\\
&=(y_t-x_t+x_t+\mu_j)'\Sigma_j^{-1}(y_t-x_t+x_t-\mu_j)\\
&=(y_t-x_t)'\Sigma_j^{-1}(y_t-x_t)+2(y_t-x_t)'\Sigma_j^{-1}(x_t-\mu_j) \qquad (2)\\
&\quad+(x_t-\mu_j)'\Sigma_j^{-1}(x_t-\mu_j)
\end{aligned}
$$

- The expected value of the second term in (2) is zero. Only the first term, i.e., $(y_t-x_t)'\Sigma_x^{-1}(y_t-x_t)$, accounts for the likelihood difference between noisy and clean speech.

Noise Robustness Analysis For Cepstral Features

- Thus, for each utterance, we define the following **cepstral distance** to measure the mismatch caused by noise

$$CD = E[(y_t - x_t)' \sum_x^{-1} (y_t - x_t)]$$

  - $E[\cdot]$ denotes the time average over the utterance, and $\sum_x$ is the diagonal covariance matrix computed from the utterance.

- Assuming that dynamic and static features are independent of each other, the distances contributed by them are separately evaluated as

$$CD(s) = E[(y_t^s - x_t^s)' (\sum_x^s)^{-1} (y_t^s - x_t^s)]$$

$$CD(d) = E[(y_t^d - x_t^d)' (\sum_x^d)^{-1} (y_t^d - x_t^d)]$$

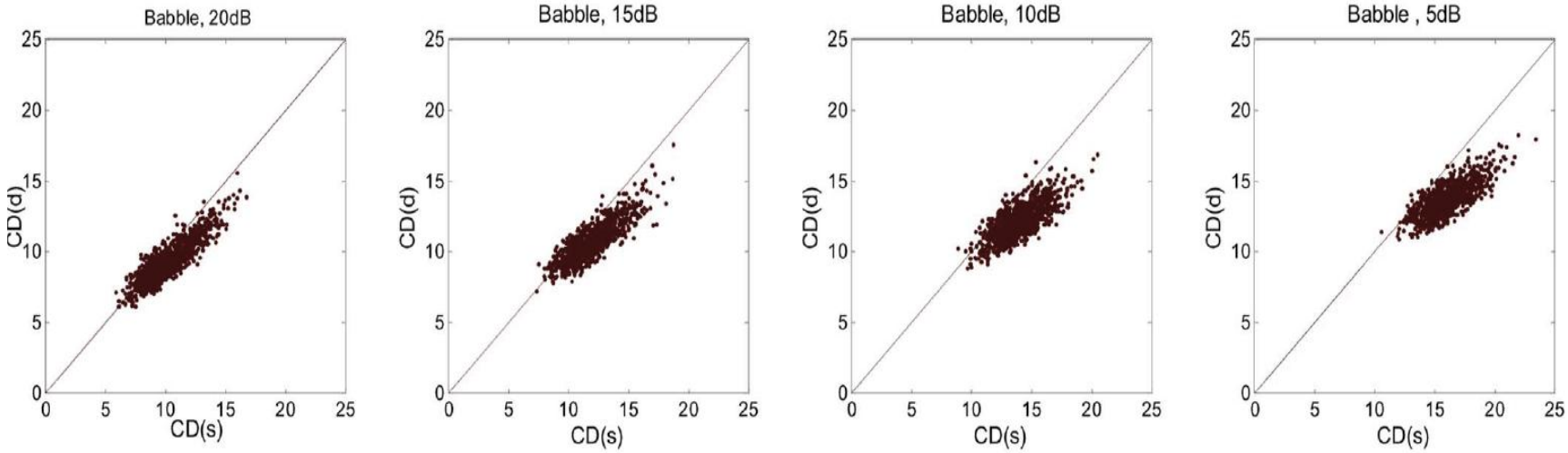

Fig. 2. Scatter diagrams of dynamic cepstral distance versus static cepstral distance (Aurora 2, babble noise).

TABLE I
RECOGNITION ACCURACY AND ERROR PATTERNS OBTAINED WITH DIFFERENT FEATURES (AURORA 2, BABBLE NOISE, 5 DB SNR)

System	Corr	Acc	Del	Sub	Ins
Static+dynamic	61.70%	30.80%	347	920	1022
Dynamic-only	45.53%	45.41%	1600	202	4
Static-only	38.09%	24.67%	973	1075	444

TABLE II
RELATIVE PERFORMANCE DIFFERENCE FROM "STATIC -ONLY" TO "DYNAMIC-ONLY" (AURORA 2, BABBLE NOISE, 5 DB SNR)

Deletion	Substitution	Insertion	Word error rate (WER)
64.44%	-81.21%	-99.10%	-27.53%

TABLE III
RECOGNITION ACCURACY AND ERROR PATTERNS OBTAINED WITH DIFFERENT FEATURES UNDER CLEAN CONDITION (AURORA 2)

System	Corr	Acc	Del	Sub	Ins
Static+dynamic	98.9%	98.9%	17	18	2
Dynamic-only	98.4%	98.1%	22	31	10
Static-only	96.7%	96.3%	40	70	12

Weighting of Dynamic and Static Features

- The output pdf at a particular HMM state is expressed as a Gaussian mixture, that is

$$b_j(o_t) = \sum_{k=1}^{K} c_{jk} N(o_t; \mu_{jk}, \Sigma_{jk})$$

  - *k* indexes the Gaussian mixture component, and c_{jk} is the corresponding mixture weight.
- Assuming that the dynamic and the static features are conditionally independent of each other (as implied by the diagonal covariance assumption), $b_j(o_t)$ can be split into two separate corresponding terms

$$b_j(o_t) = \sum_{k=1}^{K} c_{jk} e^{\left[\log N(o_t^d; \mu_{jk}^d, \Sigma_{jk}^d) + \log N(o_t^s; \mu_{jk}^s, \Sigma_{jk}^s)\right]}$$

Weighting of Dynamic and Static Features

- Different exponential weights can be applied to the acoustic likelihood components as

$$b_j(o_t) = \sum_{k=1}^{K} c_{jk} e^{\left[\alpha \log N(o_t^d; \mu_{jk}^d, \Sigma_{jk}^d) + \beta \log N(o_t^s; \mu_{jk}^s, \Sigma_{jk}^s)\right]}$$

  - α is the dynamic feature weight and β is the static feature weight, subject to the constraint of unity sum: $\alpha + \beta = 1$.

- The best performance is attained with $\alpha = 0.9$ and $\beta = 0.1$.

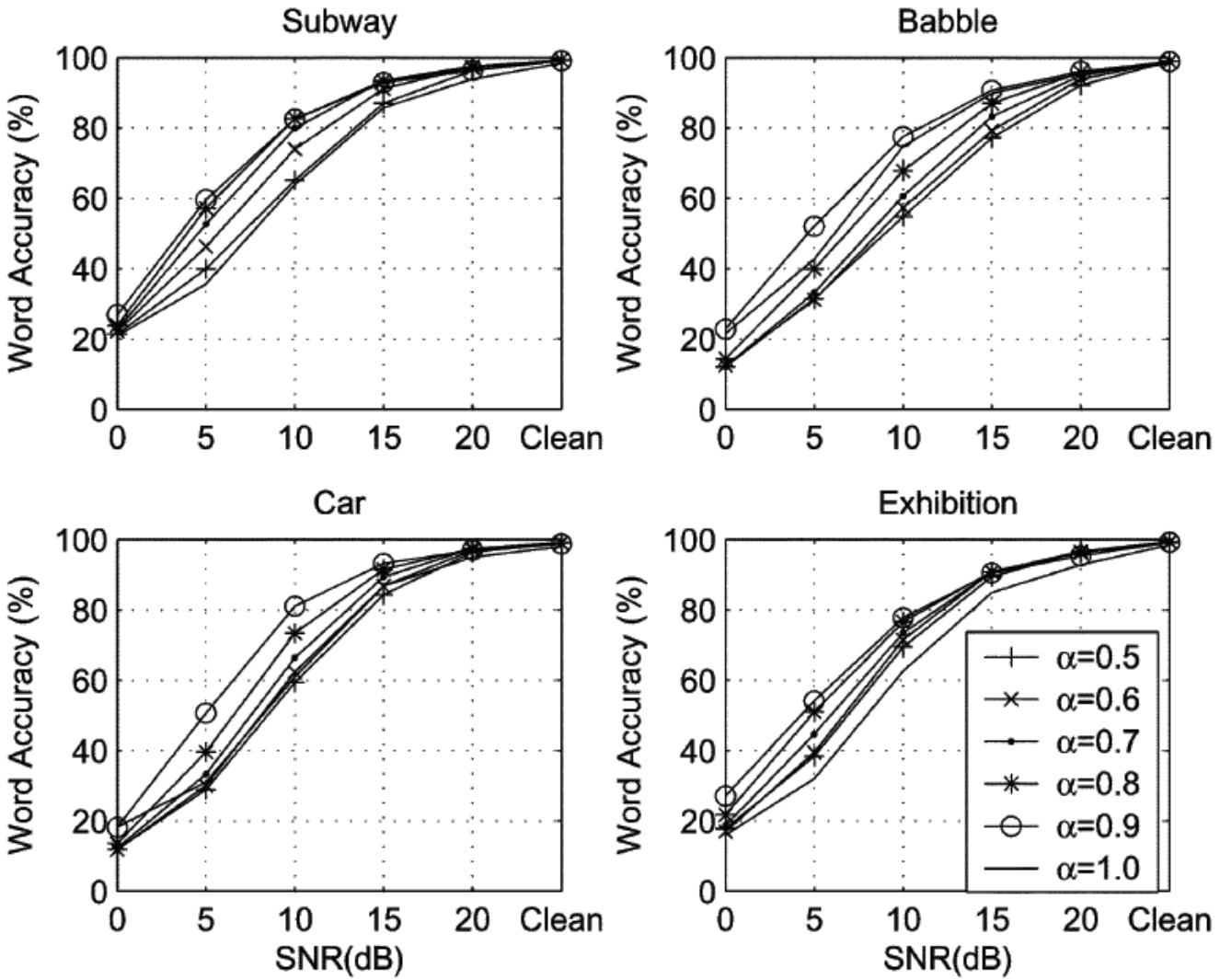

Fig. 4. Recognition performance using bracketed weights (test Set A of Aurora 2).

Discriminative Training of Feature Weights

- The goal of discriminative training for speech recognition is to **minimize the recognition error rate**.

- We use the log likelihood difference between **the recognized** and **the correct word sequences** as the optimization criterion. Given an utterance $o_u = \{o_{u1}, o_{u2}, \ldots, o_{uT}\}$

$$lld(o_u) = g^r(o_u) - g^c(o_u)$$

  - $g^r(o_u)$ is the log likelihood of the recognized word sequence for o_u
  - $g^c(o_u)$ is the log likelihood computed by forced alignment with the correct answer.

Discriminative Training of Feature Weights

- The empirical cost averaged over a set of *U* training utterances is given by $$LLD = \frac{1}{U}\sum_{u=1}^{U} lld(o_u)$$
- This empirical cost is minimized by iteratively adjusting the feature weights and , using the **gradient descent** method, that is $$\alpha(n+1) = \alpha(n) - \varepsilon\left(\frac{\partial LLD}{\partial \alpha}\right) \qquad \beta(n+1) = \beta(n) - \varepsilon\left(\frac{\partial LLD}{\partial \beta}\right)$$
  - ε is the step size of parameter adjustment and *n* indexes the iteration of adjustment.

$$\frac{\partial lld(o_u)}{\partial \alpha} = \frac{\partial g^r(o_u)}{\partial \alpha} - \frac{\partial g^c(o_u)}{\partial \alpha}$$

Discriminative Training of Feature Weights

- With the superscript symbols removed for clarity, we have

$$\frac{\partial g(o_u)}{\partial \alpha} = \sum_{t=1}^{T} \frac{\partial \{\log b_{q_t^*}(o_{ut})\}}{\partial \alpha}$$

  - q_t^* denotes the HMM state associated with o_{ut} and $b_{q_t^*}(o_{ut})$ is the corresponding acoustic probability.

$$\frac{\partial \{\log b_{q_t^*}(o_{ut})\}}{\partial \alpha} = \frac{1}{b_{q_t^*}(o_{ut})} \cdot \frac{\partial \{b_{q_t^*}(o_{ut})\}}{\partial \alpha} = \frac{1}{b_{q_t^*}(o_{ut})} \times$$

$$\sum_{k=1}^{K} \left\{ c_{jk} \exp\left[\alpha \log N(o_{ut}^d; u_{q_t^* k}^d, \sigma_{q_t^* k}^d) + \beta \log N(o_{ut}^s; u_{q_t^* k}^s, \sigma_{q_t^* k}^s) \right] \cdot \log N(o_{ut}^d; u_{q_t^* k}^d, \sigma_{q_t^* k}^d) \right\}$$

Experiments

- Aurora 2 database:
  - There are three different sets of test data in Aurora 2.
  - Test Set A is used for matched-condition tests while Test Set B and C are for mismatched-condition tests.

- The CUDIGIT Database:
  - Cantonese is a major Chinese dialect with a speaker population of over 60 million.
  - CUDIGIT contains connected digit strings spoken by 25 male and 25 female speakers. Each speaker has about 560 utterances of variable lengths from 1 to 16 digits.

Recognition Results With Condition-Specific Weights

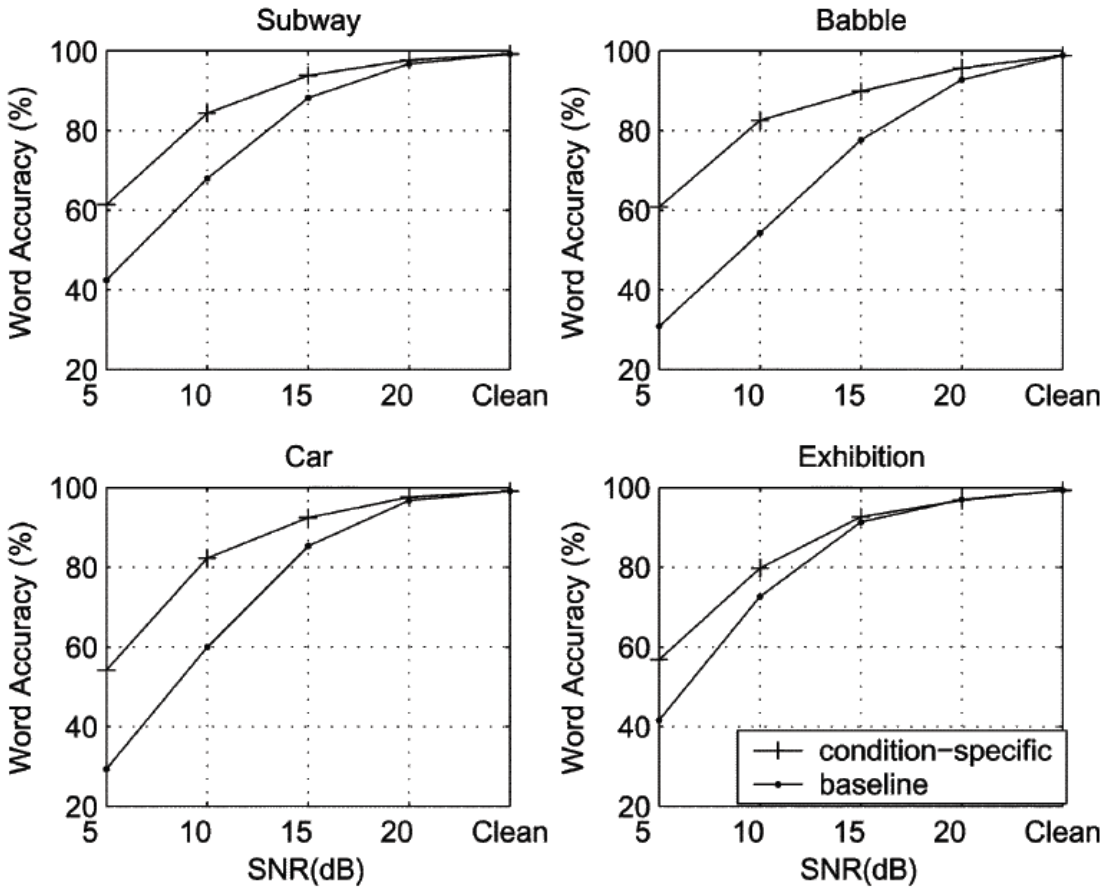

Fig. 6. Recognition performance with condition-specific feature weights (Test Set A of Aurora 2).

Recognition Results With Condition-Specific Weights

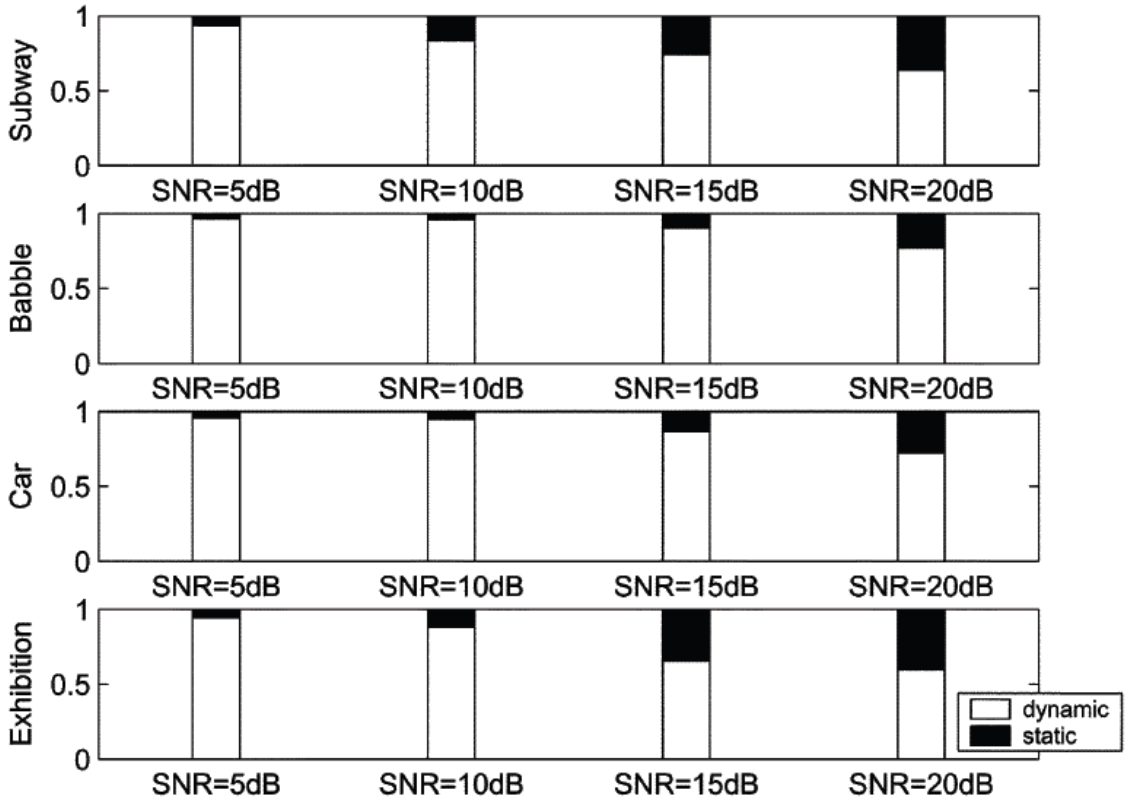

Fig. 7. Optimal condition-specific weights for different noise conditions (Test Set A of Aurora 2).

Universal Weights on Matched Conditions

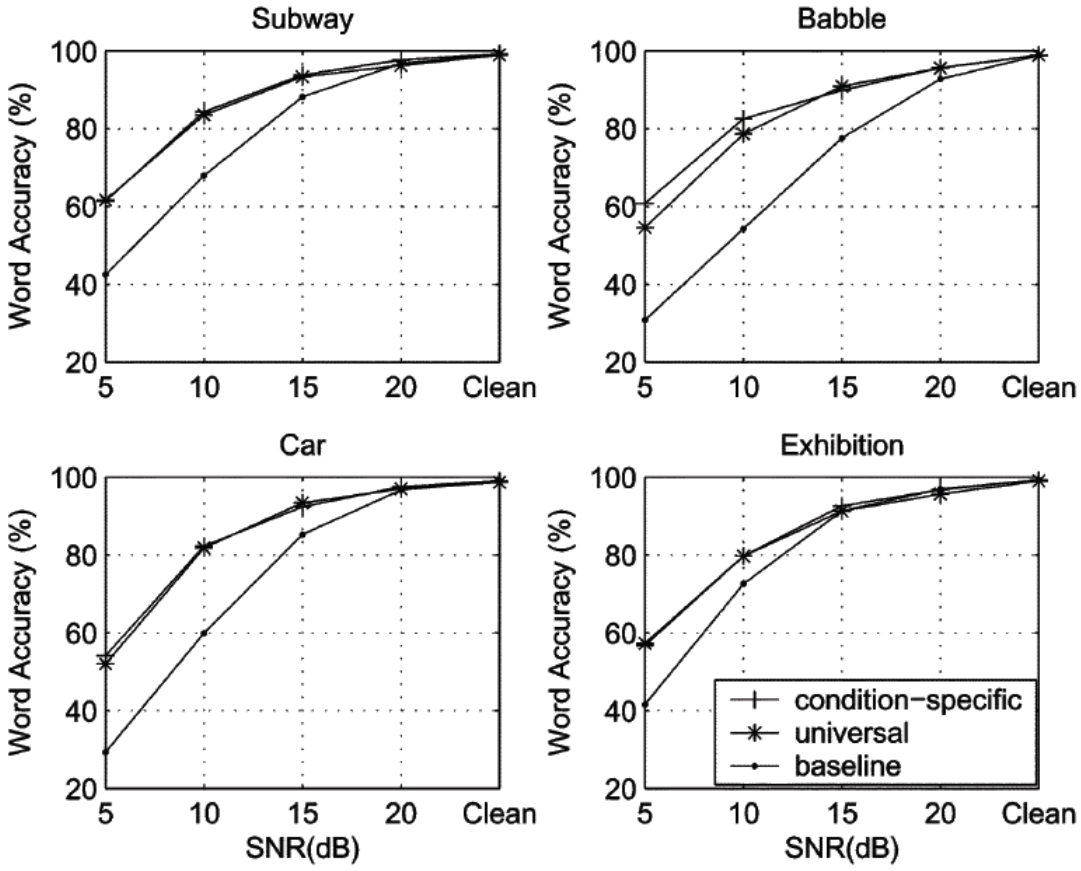

Fig. 8. Recognition performance with universal weights for matched conditions (Test Set A of Aurora 2).

Universal Weights on Mismatched Conditions

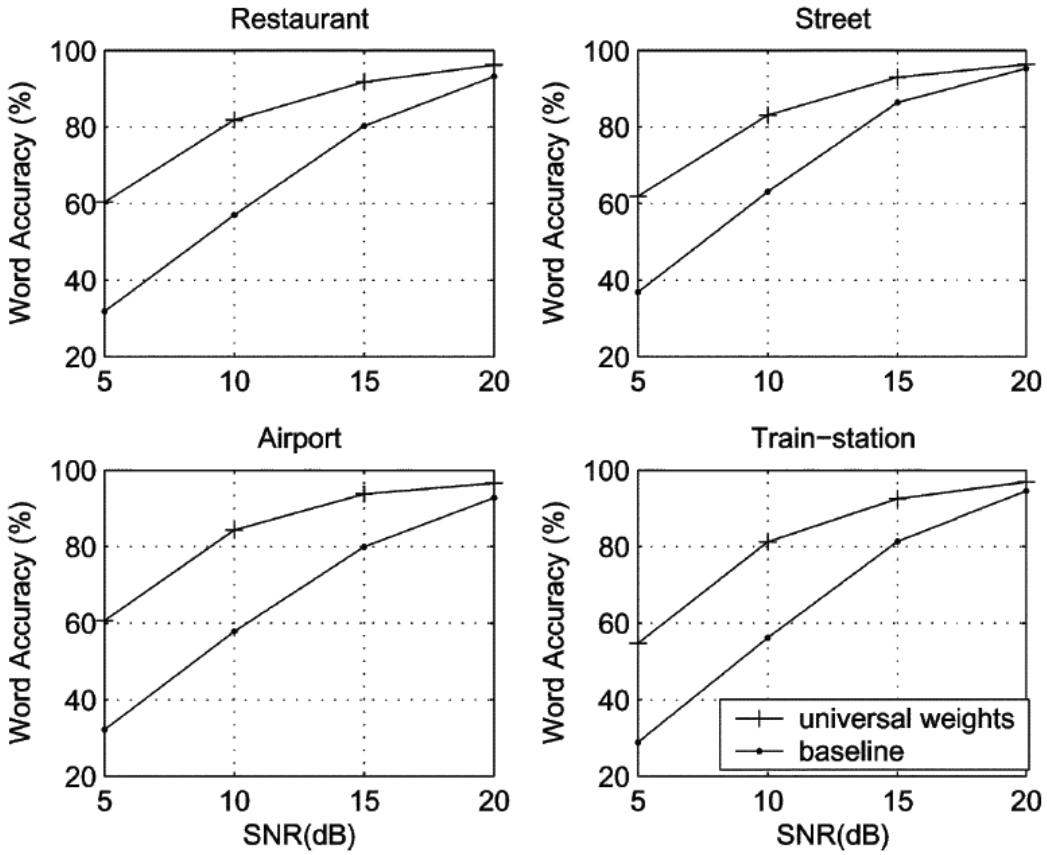

Fig. 9. Recognition performance with universal weights for mismatched noise types (Test Set B of Aurora 2).

Universal Weights on Mismatched Conditions

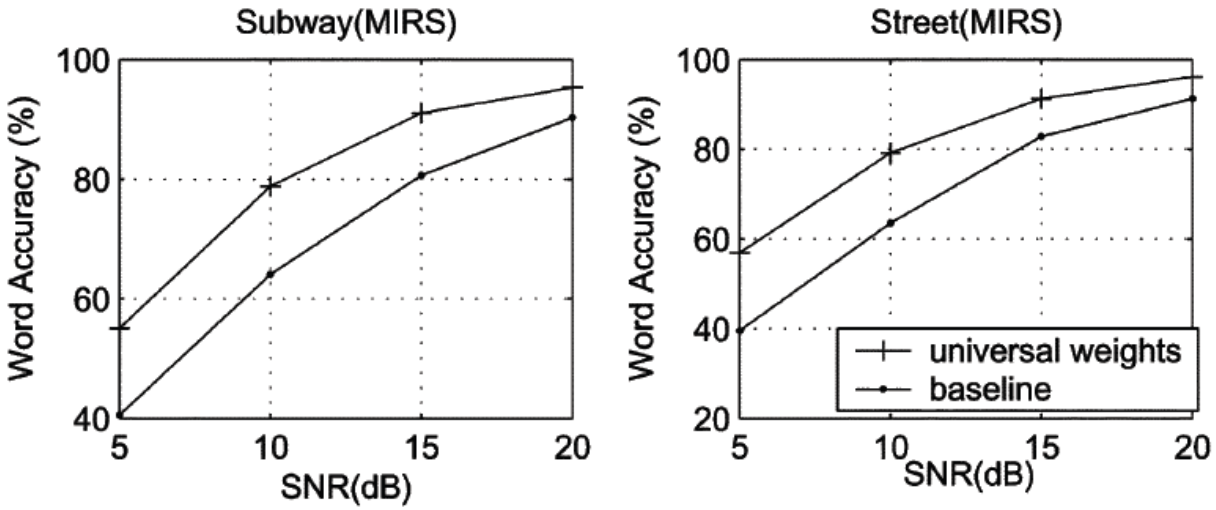

Fig. 10. Recognition performance with universal weights when there exists channel distortion (Test Set C of Aurora 2).

The CUDIGIT Database

BASELINE PERFORMANCE OF THE CUDIGIT TASK

Noise/SNR	0dB	5dB	10dB	15dB	20dB	clean
White	9.6%	14.9%	20.1%	40.3%	69.9%	97.9%
Babble	-22.1%	-6.8%	14.3%	46.3%	76.7%	97.2%
Car	16.1%	27.7%	56.3%	82.1%	92.0%	97.5%
Factory	3.8%	26.9%	60.5%	85.3%	94.1%	97.5%

The CUDIGIT Database

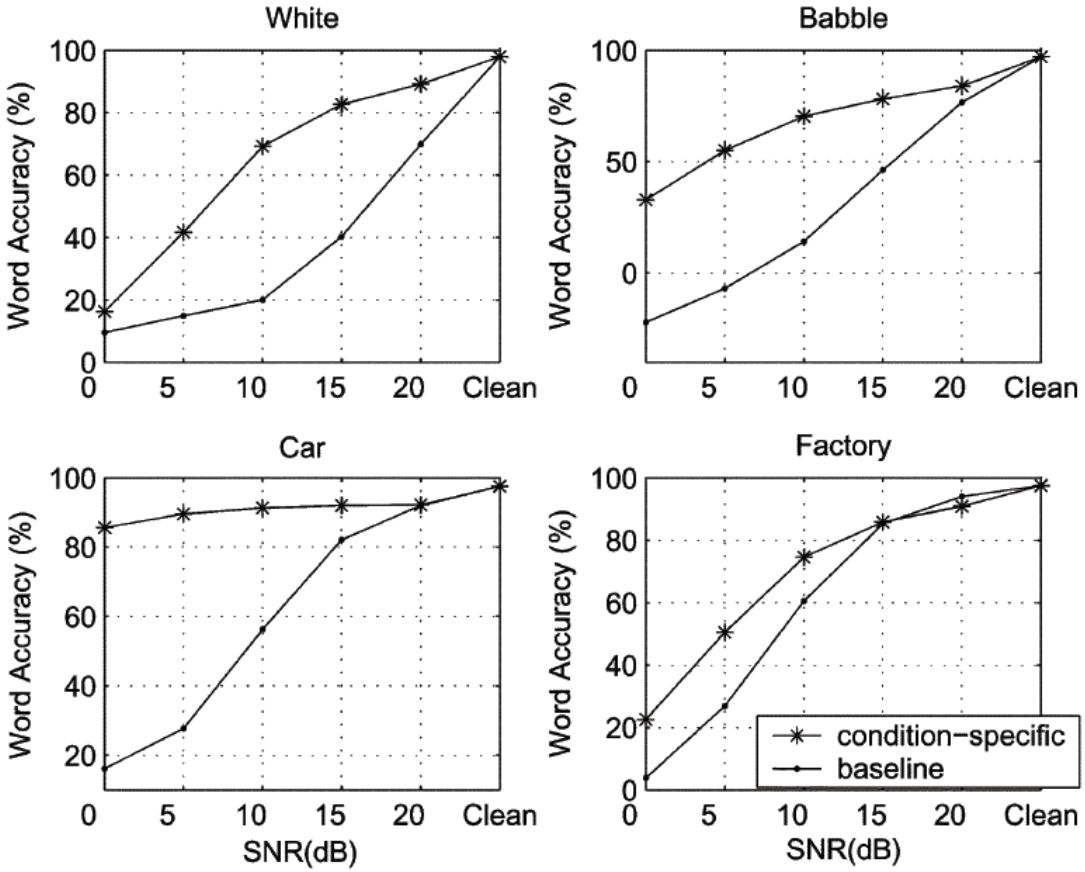

Fig. 11. Recognition performance with condition-specific weights (CUDIGIT).

Inclusion of The Acceleration Features

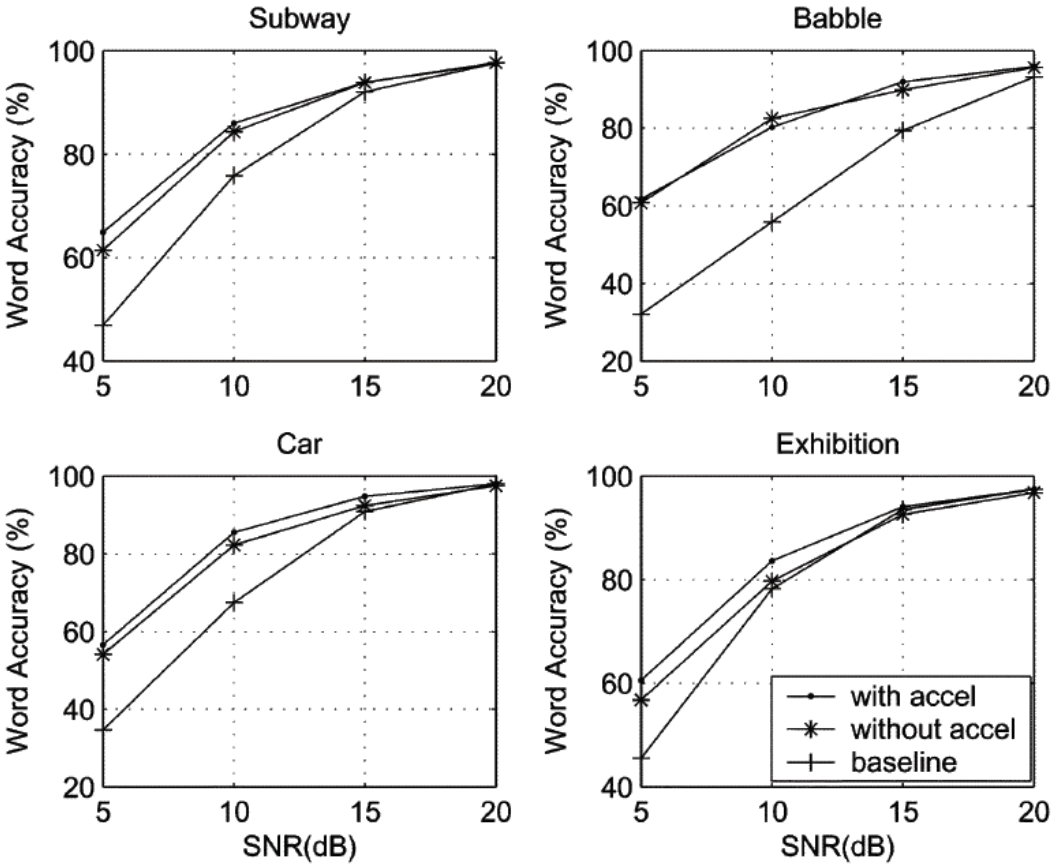

Fig. 14. Recognition performance with the inclusion of condition-specific weighted acceleration features (Test Set A of Aurora 2).

Inclusion of The Acceleration Features

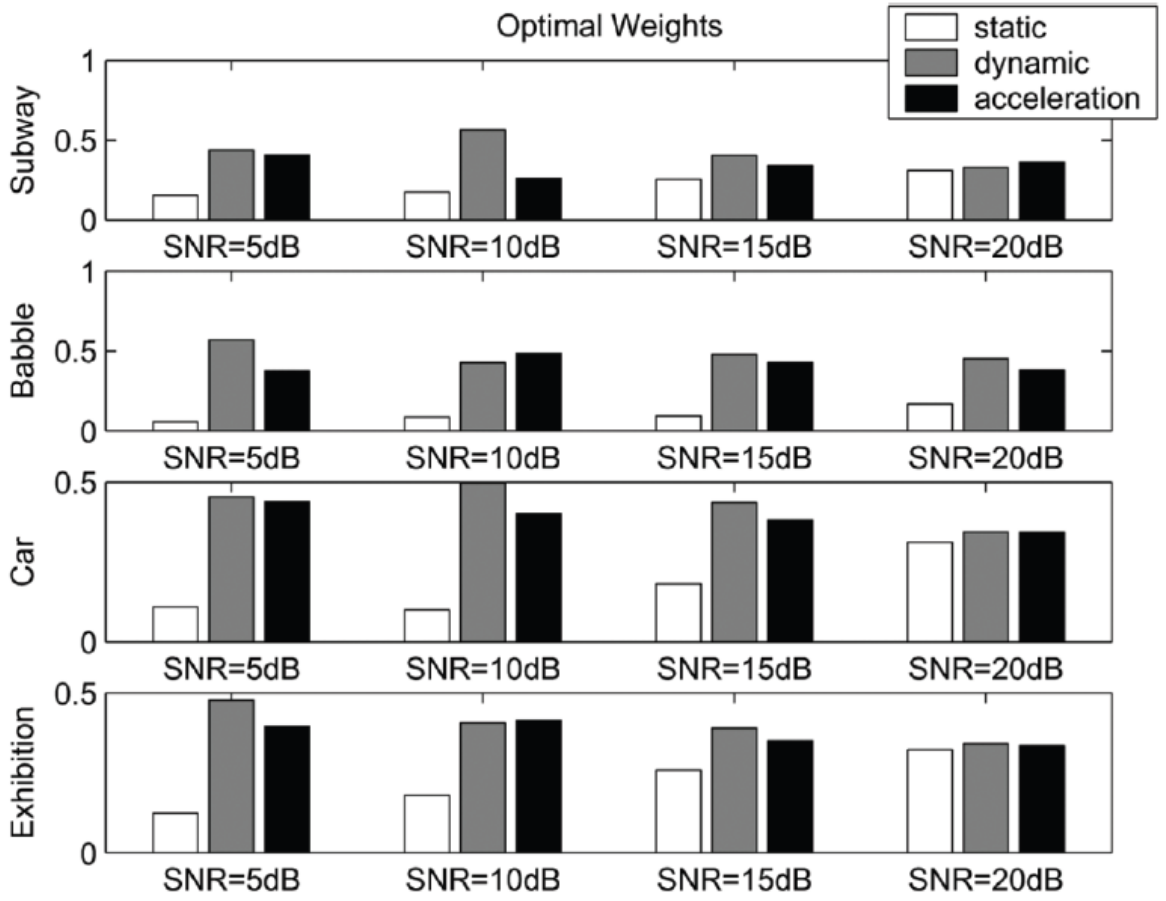

Fig. 15. Comparison of optimal condition-specific weights for static, dynamic, and acceleration features (Test Set A of Aurora 2).