STATIC AND DYNAMIC SPECTRAL FEATURES: THEIR NOISE ROBUSTNESS AND OPTIMAL WEIGHTS FOR ASR

Chen Yang[1] *Frank K. Soong*[1, 2] *Tan Lee*[1]

[1]Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong
[2]Spoken Language Translation Labs, ATR, Kyoto, Japan
{cyang, tanlee}@ee.cuhk.edu.hk frank.soong@atr.jp

ABSTRACT

In this paper we investigate the relative noise robustness between dynamic and static spectral features, by using two speaker independent, continuous digit databases in English (Aurora2) and Cantonese (CUDigit). It is found that dynamic cepstrum is more robust to additive noise than its static counterpart. The results are consistent across different types of noise and under various SNRs. Optimal exponential weights for exploiting unequal noise robustness of the two features are discriminatively trained in a development set. When tested under various noise conditions, the optimal weights yielded relative word error rate reductions of 36.6% and 41.9% for Aurora2 and CUDigit, respectively. The proposed weighting is attractive for many ASR applications in noise because (1) no noise estimation for feature compensation; (2) no adaptation of clean HMMs to a noisy environment; and (3) only a trivial change in the decoding process by weighting log likelihoods of static and dynamic components separately.

1. INTRODUCTION

Automatic speech recognition (ASR) has achieved a high performance in controlled, laboratory environments where background noise and channel variation are rather benign. However, in many real world applications, ASR performance degrades rapidly when there is a substantial mismatch between models trained in a clean environment and noisy test conditions. The most direct way to reduce this mismatch is to train or to adapt the speech recognizer by using condition-specific, noisy data. However, since there are just too many different kinds of noise and operating SNRs can vary from one environment to the next, this approach is virtually infeasible. We need other more practical approaches to noisy speech recognition.

By focusing on the three modules in a Hidden Markov Model (HMM) based ASR system: acoustic feature front-end, acoustic and language models and pattern matching decoder, we can deal with the problems of ASR in noise with different strategies. For example, 1) finding front-end acoustic features which are more invariant or insensitive to noise interference or compensating features to equalize the noise effect; 2) adapting acoustic models to make them more resistant to noise distortions; 3) weighting features in decoding to exploit their possible unequal robustness to noise. In this study we concentrate on characterizing front-end features by quantifying their relative robustness to noise and exploiting this unequal robustness by applying different weightings to the corresponding likelihood components in decoding. Only HMMs trained in *clean* environments are used exclusively in all experiments in this study.

Dynamic cepstral features can help static features to characterize the speech trajectory on its time varying rate. It has been shown that such a representation yields higher speech and speaker recognition performance than static cepstra only [1-2]. However, not too many quantitative studies have been done to examine the robustness of static and dynamic features for ASR in noise [3]. In this paper we try to quantify the robustness of static and dynamic features under different types of noise and variable SNRs. Furthermore, based on the findings we design a simple but effective noise robust recognizer by weighting exponentially the likelihoods of static and dynamic features unevenly in decoding, motivated partially by the approach in [4] where only clean signals were considered. A discriminative training procedure is proposed to train the weights automatically using a small development set. The approach was evaluated on two connected digit databases, one in English (Aurora2) [5] and the other in Cantonese (CUDigit) [6].

2. NOISE ROBUSTNESS ANALYSIS

2.1 Recognition with Only Static or Dynamic Features

To investigate the robustness of static and dynamic features to noise in recognition, we build two separate HMMs, based upon static-only and dynamic-only cepstral features, and test them in various types of noise digitally added to clean speech at different SNRs, using the Aurora2 database.

The performance of digit accuracy is shown in Fig.1 where three curves, labeled as “baseline”, “dynamic-only”, and “static-only”, are compared. In clean condition, the baseline system of the augmented static and dynamic features performs the best, better than either the static-only or the dynamic-only system, as expected. In additive noise, dynamic-only system starts to outperform the static-only one and the performance gap enlarges with decreasing SNRs till the noise level becomes too high. In babble and car noises, dynamic features even outperform the full features.

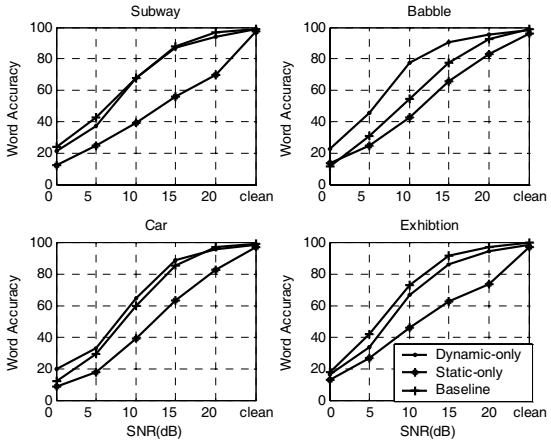

Fig. 1 Digit accuracy (%) of the baseline (full features), dynamic-only and static-only HMMs (Aurora2)

2.2 Static and Dynamic Cepstral Distances between Clean and Noisy Speech

For a given sequence of noisy speech observation, $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_T)'$, the output likelihood is:

$$b_j(\mathbf{y}_t) = \frac{1}{\sqrt{(2\pi)^V |\mathbf{\Sigma}_j|}} \exp\{-\tfrac{1}{2}(\mathbf{y}_t - \boldsymbol{\mu}_j)' \mathbf{\Sigma}_j^{-1} (\mathbf{y}_t - \boldsymbol{\mu}_j)\} \quad (1)$$

where we use a single Gaussian pdf to simplify our analysis. For an HMM of multi-mixture of Gaussians we can follow the same but somewhat more complicated analysis. The mismatch between clean and noisy conditions lies mainly on the exponent term which can be re-written as:

$$\begin{aligned} &(\mathbf{y}_t - \boldsymbol{\mu}_j)' \mathbf{\Sigma}_j^{-1} (\mathbf{y}_t - \boldsymbol{\mu}_j) \\ &= (\mathbf{y}_t - \mathbf{x}_t + \mathbf{x}_t - \boldsymbol{\mu}_j)' \mathbf{\Sigma}_j^{-1} (\mathbf{y}_t - \mathbf{x}_t + \mathbf{x}_t - \boldsymbol{\mu}_j) \\ &= (\mathbf{y}_t - \mathbf{x}_t)' \mathbf{\Sigma}_j^{-1} (\mathbf{y}_t - \mathbf{x}_t) + 2(\mathbf{y}_t - \mathbf{x}_t)' \mathbf{\Sigma}_j^{-1} (\mathbf{x}_t - \boldsymbol{\mu}_j) + (\mathbf{x}_t - \boldsymbol{\mu}_j)' \mathbf{\Sigma}_j^{-1} (\mathbf{x}_t - \boldsymbol{\mu}_j) \end{aligned} \quad (2)$$

where $\mathbf{y}_t$ and $\mathbf{x}_t$ are the corresponding noisy and clean speech observations; $\boldsymbol{\mu}_j$ and $\mathbf{\Sigma}_j$, corresponding mean and covariance of the clean HMM at state j.

Since the expected value of the second term is zero, the difference of likelihood between noisy and clean speech is just the first term, $(\mathbf{y}_t - \mathbf{x}_t)' \mathbf{\Sigma}_j^{-1} (\mathbf{y}_t - \mathbf{x}_t)$, which can be viewed as a weighted cepstral distance normalized by the variance of the clean model. We thus define a cepstral distance to measure this mismatch as:

$$CD = E[(\mathbf{y}_t - \mathbf{x}_t)' \mathbf{\Sigma}_{\mathbf{x}}^{-1} (\mathbf{y}_t - \mathbf{x}_t)] \quad (3)$$

where the diagonal covariance of the utterance, $\mathbf{\Sigma}_{\mathbf{x}}$, is used to approximate the diagonal covariance, $\mathbf{\Sigma}_j$, in the clean speech model; $E[\cdot]$ denotes the time average over the whole utterance.

In the following analysis, we select test set A in Aurora2 database as the analysis data set. The clean speech data are digitally contaminated by subway, babble, car, and exhibition noises at SNRs from 5dB to 20dB, incrementing at a step of 5dB. We computed the weighted distances between clean and noisy speech for both the static and the dynamic features, respectively:

$$CD(d) = E\left[(\mathbf{y}_t^d - \mathbf{x}_t^d)' (\mathbf{\Sigma}_{\mathbf{x}}^d)^{-1} (\mathbf{y}_t^d - \mathbf{x}_t^d)\right] \quad (4)$$

$$CD(s) = E\left[(\mathbf{y}_t^s - \mathbf{x}_t^s)' (\mathbf{\Sigma}_{\mathbf{x}}^s)^{-1} (\mathbf{y}_t^s - \mathbf{x}_t^s)\right] \quad (5)$$

where the superscripts *d* and *s* denote the dynamic and the static features.

Fig. 2 depicts the scatter diagrams of dynamic distance (between clean and noisy dynamic cepstra) vs. its static counterpart. Four diagrams are shown for babble noise at SNRs of 5, 10, 15, and 20dBs. Scatter diagrams of other noises follow similar patterns but not shown here. In each scatter plot, the cepstral distance (between clean and noisy speech at a specified SNR) of static feature and that of dynamic feature form a 2-dimensional point for each utterance. The diagonal line in each plot represents a trace where the two distances, dynamic and static, are equal.

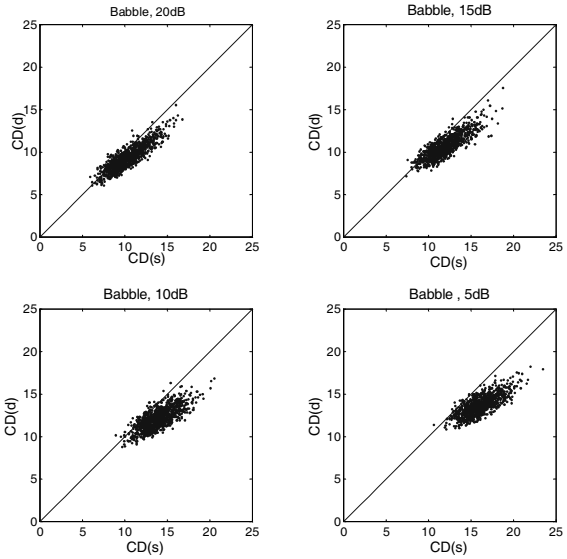

Fig. 2 Scatter diagrams of dynamic cepstral distance vs. static cepstral distance (variance normalized) in babble noise

Two observations can be made on the figure:

(1) With decreasing SNRs, the points move away from the origin when both cepstral features are mismatched with their clean counterparts. Or both distances are larger for increasingly mismatched conditions at lower SNRs.

(2) For all types of noise, majority points fall below the diagonal line. In other words, the dynamic cepstral distance between noisy and clean features is smaller than its static counterpart, after variance normalization. It indicates that the dynamic features are more resilient to noise than the static features, hence better recognition performance in noise.

3. EXPONENTIAL WEIGHTINGS IN DECODING

3.1. Exponential Weightings

Based on the findings in the previous section, we propose to weight the log likelihoods of static and dynamic features differently in decoding to exploit their uneven noise robustness. Assuming the dynamic and static features are mutually independent (as implied by the diagonal covariance), the output

likelihood of an observation can be split into two separate corresponding terms, *d* and *s*, as:

$$b_j(\mathbf{o}_t) = \sum_{k=1}^{K} c_{jk} \exp\{\log[N(\mathbf{o}_t^d; \boldsymbol{\mu}_{jk}^d, \boldsymbol{\sigma}_{jk}^d)] + \log[N(\mathbf{o}_t^s; \boldsymbol{\mu}_{jk}^s, \boldsymbol{\sigma}_{jk}^s)]\} \quad (6)$$

where *k* is the *k*-th mixture index; c_{jk}, the mixture weight.

The acoustic likelihood components can be computed with different exponential weightings as:

$$b_j(\mathbf{o}_t) = \sum_{k=1}^{K} c_{jk} \exp\{\alpha \log[N(\mathbf{o}_t^d; \boldsymbol{\mu}_{jk}^d, \boldsymbol{\sigma}_{jk}^d)] + \beta \log[N(\mathbf{o}_t^s; \boldsymbol{\mu}_{jk}^s, \boldsymbol{\sigma}_{jk}^s)]\} \quad (7)$$

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

3.2 Recognition with Bracketed Weightings

We tested our idea of exponential weights by bracketing the two weights at a step of 0.1 with the constraint of unity sum. The recognition performance obtained with bracketed weights is shown in Fig. 3, where recognition accuracy improves continuously with increasing dynamic feature weight and reaches the best performance at the largest weight, $\alpha = 0.9$.

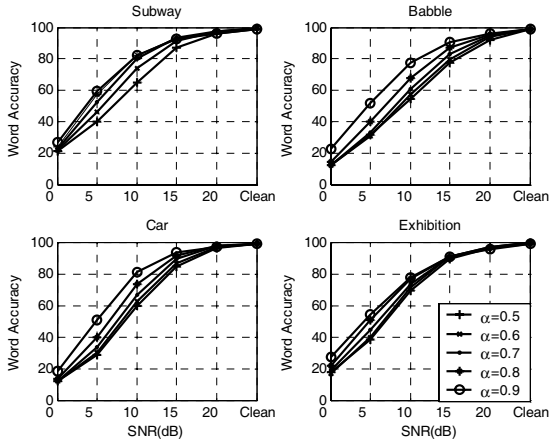

Fig. 3 Recognition performance of word accuracy (%) using bracketed weightings (Aurora2)

4. DISCRIMINATIVE WEIGHT TRAINING

Encouraged by the results in the previous subsection, we decide to optimize the weight values. The log likelihood difference (*lld*) between the recognized and the correct states is chosen as the objective function for optimization. For the *u*-th speech utterance of T observations, $\mathbf{O}_u = (\mathbf{o}_{u_1}, \mathbf{o}_{u_2}, ..., \mathbf{o}_{u_T})$, the *lld* [7-8] is:

$$lld(\mathbf{O}_u) = g^r(\mathbf{O}_u) - g^l(\mathbf{O}_u) \quad (8)$$

where $g^r(\mathbf{O}_u)$ is the log likelihood of the recognition result and $g^l(\mathbf{O}_u)$, that of the correct (forced) alignment.

The cost averaged over the whole training set of *U* utterances is:

$$LLD = \frac{1}{U} \sum_{u=1}^{U} lld(\mathbf{O}_u) \quad (9)$$

This cost is minimized by adjusting iteratively the dynamic weight, α, and the static weight, β, via the steepest descent as:

$$\alpha(n+1) = \alpha(n) - \varepsilon \frac{\partial LLD}{\partial \alpha} \quad (10)$$

$$\beta(n+1) = \beta(n) - \varepsilon \frac{\partial LLD}{\partial \beta} \quad (11)$$

where

$$\frac{\partial lld(\mathbf{O}_u)}{\partial \alpha} = \frac{\partial g^r(\mathbf{O}_u)}{\partial \alpha} - \frac{\partial g^l(\mathbf{O}_u)}{\partial \alpha} \quad (12)$$

$$\frac{\partial g(\mathbf{O}_u)}{\partial \alpha} = \sum_{t=1}^{T} \frac{\partial \{\log b_j(\mathbf{o}_{u_t})\}}{\partial \alpha} \quad (13)$$

and T, total number of frames of the utterance $\mathbf{O}_u$; *n*, the iteration index; and ε, an appropriate step size.

5. DATABASES AND EXPERIMENTAL RESULTS

5.1 Databases and Experimental Setups

Databases

Two speaker independent, continuous digit databases are used in this study, one in English (Aurora2) [5] and the other one in Cantonese (CUDigit) [6].

Acoustic features

13 dimensional static MFCC (including log energy) vectors are computed in a frame of 25 msec, shifted every 10 msec. The dynamic features, i.e. ΔMFCC, are derived from the static features, in a window of 7 successive frames.

Clean speech models

Whole-word, digit HMMs are 16 (English) or 8 (Cantonese) left-to-right states without skipping. Each state's output pdf is a mixture of 3 Gaussians with diagonal covariance. There are a three-state "silence" and a single-state "short pause" models [5].

5.2 Condition Specific Weights on Aurora2 Database

In Fig. 4 the recognition performance obtained by using the optimally trained weights is depicted together with that of the unweighted, baseline system. The exponentially weighted system yields a substantial improvement over the baseline system. Overall, a 36.6% relative WER reduction is obtained.

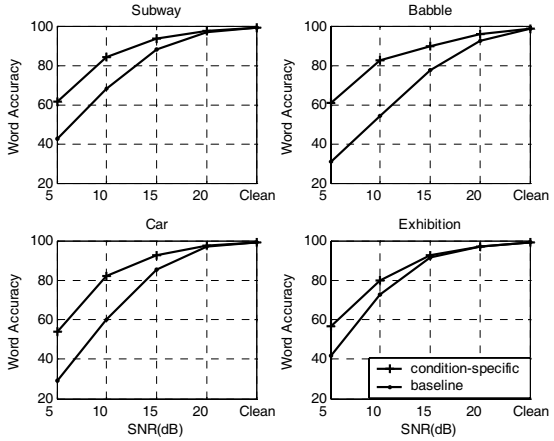

Fig. 4 Recognition performance using optimal weights (Aurora2)

5.2 Universal Weights on Aurora2 Database

In the previous experiments, condition-specific weights were tested. However, condition-specific development data for weight training may not always be available. Also, we like to investigate the weight sensitivities to mismatched testing conditions. Here, universal weights, i.e., two weights only, were trained by using multi-style training data of different noises at various SNRs and tested under various noise conditions. Fig. 5 depicts the recognition performance of word accuracy (%) by using condition-specific and universal weights. We find that the two performance curves, which are virtually on top of each other in three out of four noise conditions, are always better than the baseline results. For the babble noise, the condition-specific weights show some advantages over the universal weights.

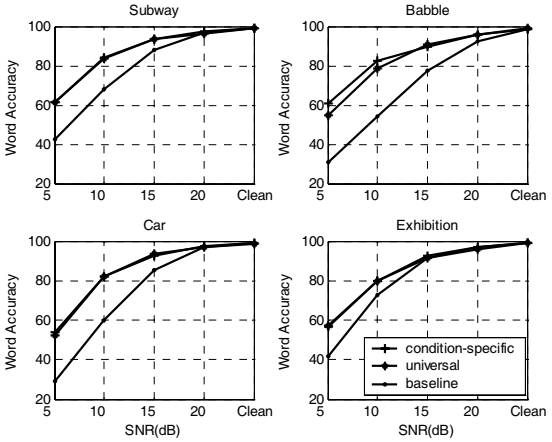

Fig. 5 Condition-specific vs. universal weight recognition performance (Aurora2)

5.3 Evaluation on CUDigit Database

The optimal weights were also tested on CUDigit, a connected Cantonese digit database [9-10]. Fig. 6 depicts the recognition digit accuracy of both weighted and baseline systems, tested on clean and digitally added noises. The relative WER improvement is 41.9%, averaged over all noise conditions.

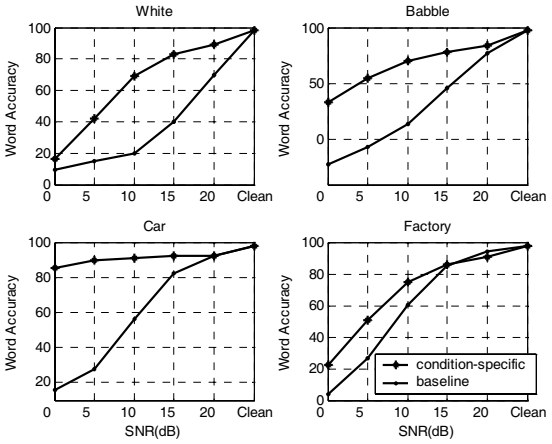

Fig. 6 Weighted and baseline performance (CUDigit)

6. CONCLUSIONS

In this paper we investigated the relative robustness of dynamic and static cepstral features for ASR in noise. The dynamic features were found to be more resilient to additive noise interference than their static counterpart. Optimal exponential weights for exploiting the unequal robustness of the two cepstral features were trained and tested on two continuous digit databases, Aurora2(English) and CUDigit(Cantonese). Relative error rate reductions of 36.6% and 41.9%, have been obtained over the baseline results, respectively. In our approach, the same clean HMM is used in the decoding process; hence no extra computation is needed. The fact that there is no need to estimate the noise or to adapt clean HMMs makes this approach rather attractive for many ASR applications in noise.

ACKNOWLEDGMENT

This research is supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region (Project No. CUHK4206/01E). C. Yang is supported by Research Studentship by a central allocation grant from Research Grants Council.

REFERENCES

[1] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoust. Speech Signal Proc.,* vol.34, pp.52-59, Feb. 1986.

[2] F. K. Soong and A. E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," *IEEE Trans. Acoust. Speech Signal Proc.,* vol. 36, pp.871–879, June 1988.

[3] B. A. Hanson and T. H. Applebaum, "Robust speaker-independent word recognition using static, dynamic and acceleration features: experiments with Lombard and noisy speech," *Proc. ICASSP-1990*, pp.857-860.

[4] J. Hernando, "Maximum likelihood weighting of dynamic speech features for CDHMM speech recognition," *Proc. ICASSP-1997*, pp.1267-1270

[5] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *ISCA ITRW ASR2000*, pp.181-188, Sept. 2000, Paris, France.

[6] T. Lee, W. K. Lo, P. C. Ching and H. Meng, "Spoken language resources for Cantonese speech processing," *Speech Communication*, pp.327-342, vol.36, 2002.

[7] J.-K. Chen and F. K. Soong, "An N-best candidates-based discriminative training for speech recognition applications," *IEEE Trans. Speech and Audio Proc.,* vol.2, pp.206–216, Jan.1994.

[8] B.-H. Juang, W. Chou and C.-H. Lee, "Minimum classification error rate method for speech recognition," *IEEE Trans. Speech and Audio Proc.,* vol.5, pp.257–265, May 1997.

[9] C. Yang, "On the robustness of static and dynamic spectral information for speech recognition in noise," Ph. D Dissertation, The Chinese University of Hong Kong, in preparation.

[10] C. Yang, F. K. Soong and T. Lee, "Noise robustness of dynamic and static features for continuous Cantonese digit recognition" submitted to *ISCSLP 2004, Hong Kong*.