ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

ON NOISE ROBUSTNESS OF DYNAMIC AND STATIC FEATURES FOR CONTINUOUS CANTONESE DIGIT RECOGNITION

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
{cyang,tanlee}@ee.cuhk.edu.hk frank.soong@atr.jp

ABSTRACT

It has been shown previously that augmented spectral features (static and dynamic cepstra) are effective for improving ASR performance in a clean environment. In this paper we investigate the noise robustness of static and dynamic cepstral features, in a speaker independent, continuous recognition task by using a noise-added, Cantonese digit database (CUDigit). We found that the dynamic cepstrum is more robust to additive, background noise than its static counterpart. The results are consistent across different types of noise and under various SNRs. Exponential weights which can exploit the unequal robustness of two features are optimally trained in a development set. A relative word error rate reduction of 41.9%, mainly on a significant reduction of insertions, is obtained on the test data under various noise and SNR conditions.

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 practical applications, ASR performance degrades rapidly when there is a substantial mismatch between models trained in a clean environment and noisy testing conditions [1]. The most direct way to reduce this mismatch is to train or to adapt the speech recognizer by using condition-specific, noisy data. However, given the fact that there are just too many kinds of noise and their levels can vary from one operating environment to the next, this approach is virtually infeasible. We therefore need a more practical approach to noisy speech recognition.

There are three main modules in a Hidden Markov Model (HMM) based ASR system: a front-end for feature extraction, acoustic and language models for modeling speech and a pattern matching decoder for recognition decision. To deal with the noise interference in ASR, approaches can be taken to address specifically the corresponding modules, including: 1) finding front-end acoustic features invariant or insensitive to noise interference or compensating the noisy features to equalize the noise effect; 2) adapting acoustic models to compensate for the noise distortion; 3) weighting features to exploit their unequal noise robustness in decoding.

In this study we will concentrate on characterizing front-end features by quantifying their relative robustness to noise and exploiting the unequal robustness by applying different weightings on the corresponding likelihood components in decoding. We will use *clean* HMMs for all experiments in this study.

Speech, a quasi-stationary, stochastic process can be analyzed by short-time spectral analysis with an appropriate frame size and rate. It is then represented as a sequence of quasi-stationary, static snap-shots. The state-of-art speech recognizer uses HMM to model speech as a stochastic, Markovian state sequence with corresponding output probability density functions (pdf's). In addition to the static feature, dynamic features are helpful to characterize the speech trajectory more precisely and it has been shown that an augmented representation (static and dynamic cepstral features) yields higher speech and speaker recognition performance than the static cepstra only [2-3] in a clean environment.

However, not too many quantitative studies have been done to examine the robustness of static and dynamic features in noise for ASR applications [4]. In this study we try to quantify the robustness of static and dynamic features under different types of noise and varying SNR's. Furthermore, based on the findings we design a simple but effective noise robust recognizer by weighting the likelihoods of dynamic and static features unevenly in decoding, motivated partially by the approaches in [5-6] where only clean signals were considered. A discriminative training procedure is proposed to train the exponential weights automatically using a small development dataset. The trained weights are evaluated in a continuous Cantonese digit database and a relative word (digit) error rate reduction of 41.9% over the conventional, un-weighted baseline recognizer is obtained.

2. DATABASE

2.1. Clean Database

Cantonese, a major Chinese dialect spoken by 60 million people in Southern China, Hong Kong and overseas, is both tonal and monosyllabic where almost every Chinese character (morpheme) is pronounced as a monosyllable. In this study, CUDigit [7], a continuous Cantonese digit database collected at the Chinese University of Hong Kong is used. It consists of 25 male and 25 female speakers. Each speaker recorded around 560 continuous strings of 1, 2, 3, 4, 7, 8, and 16 digits.

Training data:
All digit strings recorded by the first 20 male and 20 female speakers in CUDigit database is selected as the training set.

Development and test data:
The digit data recorded by the rest 10 (5 male and 5 female) speakers is used as development and test sets. Ten digit strings from each of the 10 speakers are selected as the development set and they will be used to train the exponential weights of static and dynamic cepstral features. The rest of the data is then divided into four subsets. One hundred digit strings are selected from each of the 10 speakers to form a test subset and each subset consists of 1,000 digit strings.

2.2. Noise addition

The noise samples are selected from NOISE-ROM-0[8], including: white, babble, car and factory noises. They are first down-sampled to 16 kHz and then digitally added to the clean development set and 4 test subsets, one subset for each type of noise, at specified signal-to-noise ratios (SNRs), from 0 to 20dB at a step of 5dB.

3. CANTONESE DIGIT RECOGNIZER

3.1. Recognition System

HMM models
A whole-word HMM was trained for each of the ten Cantonese digits (from "0" to "9" as "ling4","jat1", "ji6", "saam1", "sei3", "ng5", "luk6", "cat1", "baat3" and "gau2" [1]). Each model consists of 8 left-to-right states without skipping. Each state output pdf is a mixture of 3 Gaussians with diagonal covariances. In addition, there are a three-state "silence" model and a single-state "short pause" model whose only state is tied with the "silence" model [9]. HMMs were trained using 22,390 clean utterances in the training set.

Acoustic features
Cepstral features were computed in a frame of 25 msec, shifted every 10 msec in a mel-frequency scaled filter bank of 32 filters. Each feature vector contains the first 13 MFCCs (including the log energy). The dynamic features, i.e. ΔMFCC, are derived from the static features in a window of 7 successive frames.

3.2. Baseline Recognition Results

The baseline result used for comparison is the recognition word accuracy (%) obtained by using the full features (i.e., static cepstral features augmented by their 1st-order dynamic counterpart) as shown in Table 1.

SNR	0dB	5dB	10dB	15dB	20dB	Clean
White	9.58	14.89	20.09	40.27	69.89	97.99
Babble	-22.06	-6.82	14.27	46.26	76.68	97.17
Car	16.13	27.67	56.25	82.07	92.03	97.53
Factory	3.8	26.85	60.53	85.28	94.05	97.49

Table 1 Baseline performance of CUDigit database

[1] The pronunciations are written using the transcription scheme by the Linguistic Society of Hong Kong

The negative digit accuracies in the table are due to many non-substitution (i.e., insertion or deletion) errors at low SNR's. In fact, the insertion errors are the most prominent errors in many testing conditions for this Cantonese digit database.

4. MODELING AND RECOGNITION USING SEPARATE FEATURES

To investigate the robustness of static and dynamic features to noise and their individual contributions to recognition, we build two separate models, based upon static and dynamic cepstral features, and test them in various kinds of noise and at different SNRs. The resultant performance is shown in Fig.1 where three different performance curves in digit accuracy, labeled as "baseline", "dynamic only", and "static only", are compared. From the figure, in clean condition, the baseline system of the augmented static and dynamic features performs better than either dynamic- or static-only features as expected. However, in additive noise, dynamic features start to outperform either static or augmented full features. The performance differences enlarge with decreasing SNRs till the noise level becomes too high. The dynamic only HMM shows spectacular robustness to the highly low-passed, fairly stationary car noise. For other noises, the dynamic only systems still perform significantly better than either the baseline or the static only systems.

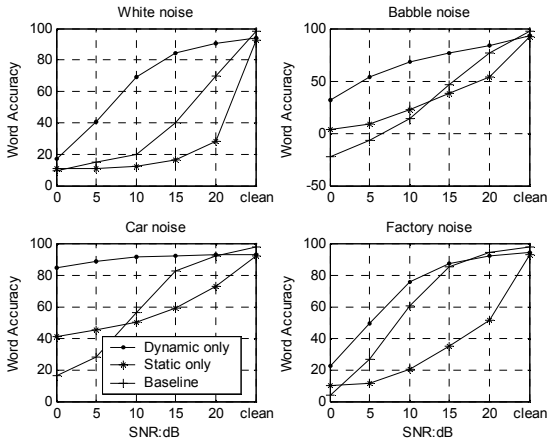

Fig.1 Digit accuracy (%) of the baseline (augmented cepstral feature), dynamic only and static only HMMs

5. UNEQUAL WEIGHTING OF DYNAMIC AND STATIC FEATURES IN DECODING

Given the *u*-th speech utterance of T observations, $\mathbf{O}_u = (\mathbf{o}_{u_1}, \mathbf{o}_{u_2}, ..., \mathbf{o}_{u_T})$, the acoustic log likelihood function is:

$$g(\mathbf{O}_u; \Lambda) = \sum_{t=1}^{T} [\log a_{q_{t-1}q_t} + \log b_{q_t}(\mathbf{o}_{u_t})] + \log \pi_{q_0} \quad (1)$$

where $\mathbf{\Lambda}$ denotes HMMs, and q_t , the decoded state at frame *t*. The output likelihood is

$$b_{q_t}(\mathbf{o}_{u_t}) = \sum_{k=1}^{K} c_{q_t k} N(\mathbf{o}_{u_t}; \boldsymbol{\mu}_{q_t k}, \boldsymbol{\sigma}_{q_t k}) \quad (2)$$

Assuming dynamic and static features are independent from each other, the output likelihood can then be split into two terms as:

$$b_{q_t}(\mathbf{o}_{u_t}) = \sum_{k=1}^{K} c_{q_t k} \exp\{\log[N(\mathbf{o}_{u_t}^d; \boldsymbol{\mu}_{q_t k}^d, \boldsymbol{\sigma}_{q_t k}^d)] + \log[N(\mathbf{o}_{u_t}^s; \boldsymbol{\mu}_{q_t k}^s, \boldsymbol{\sigma}_{q_t k}^s)]\} \tag{3}$$

where k is the mixture component index; $c_{q_t k}$, the corresponding mixture weight; and superscript "d" denotes the dynamic feature whereas "s", the static feature.

As observed in the previous section, the effectiveness of static feature and dynamic feature in noisy speech recognition are quite different. Different weights should then be used to exploit their relative effectiveness. We propose to weight the output likelihood exponentially as:

$$b_{q_t}(\mathbf{o}_{u_t}) = \sum_{k=1}^{K} c_{q_t k} \exp\{\alpha \log[N(\mathbf{o}_{u_t}^d; \boldsymbol{\mu}_{q_t k}^d, \boldsymbol{\sigma}_{q_t k}^d)] + \beta \log[N(\mathbf{o}_{u_t}^s; \boldsymbol{\mu}_{q_t k}^s, \boldsymbol{\sigma}_{q_t k}^s)]\} \tag{4}$$

where α is the dynamic weight; β, the static weight.

The weights are discriminatively trained subjected to a unity sum constraint, $\alpha + \beta = 1$.

6. DISCRIMINATIVE TRAINING OF WEIGHTS

To train the optimal weights of static and dynamic features automatically, an objective cost function is defined first. The likelihood difference between the likelihoods of the recognized and correct state (i.e., through forced alignment between the acoustic observations and the correct given digit transcriptions) is used here. For a given speech utterance, $\mathbf{O}_u$, the likelihood difference [10-11] is:

$$lld(\mathbf{O}_u) = g^r(\mathbf{O}_u) - g^l(\mathbf{O}_u) \tag{5}$$

where $g^r(\mathbf{O}_u)$ is the log likelihood of the recognition result and $g^l(\mathbf{O}_u)$, that of the correct alignment. The empirical cost is defined as the average of log likelihood difference over U utterances in the training set:

$$LLD = \frac{1}{U}\sum_{u=1}^{U} lld(\mathbf{O}_u) \tag{6}$$

According to the definition, the more negative the empirical cost function, the better the performance. This empirical cost LLD can be minimized by adjusting iteratively the dynamic weight, α, and the static weight, β via the steepest descent as:

$$\alpha(n+1) = \alpha(n) - \varepsilon \frac{\partial LLD}{\partial \alpha} \tag{7}$$

$$\beta(n+1) = \beta(n) - \varepsilon \frac{\partial LLD}{\partial \beta} \tag{8}$$

where

$$\frac{\partial lld(\mathbf{O}_u)}{\partial \alpha} = \frac{\partial g^r(\mathbf{O}_u)}{\partial \alpha} - \frac{\partial g^l(\mathbf{O}_u)}{\partial \alpha} \tag{9}$$

$$\frac{\partial g(\mathbf{O}_u)}{\partial \alpha} = \sum_{t=1}^{T} \frac{\partial\{\log b_{q_t}(\mathbf{o}_{u_t})\}}{\partial \alpha} \tag{10}$$

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

We use the development data to train the weights. Fig. 2 gives the recognition word accuracy obtained from thus trained weights along with the baseline performance. The relative performance improvement is 41.9%, averaged over all noise conditions.

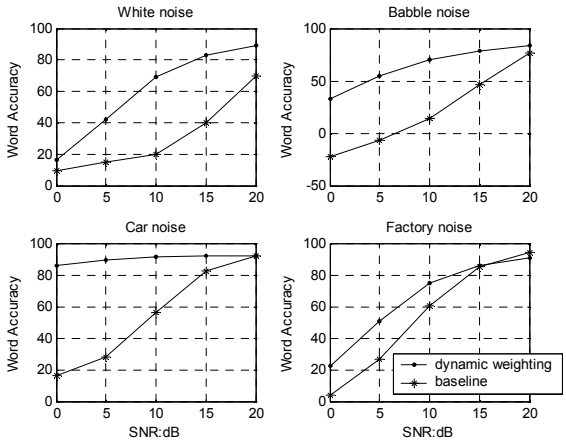

Fig.2 Digit accuracy (%) of optimally weighted and equally weighted (baseline) HMMs

7. COMPARISON AND DISCUSSIONS

We also have evaluated the dynamic feature weighting scheme in Aurora2 task and obtained similar recognition improvements by weighting the static and dynamic cepstrum unequally [12-13]. However, an interesting contrast exists between the optimal weights derived from CUDigit and Aurora2 databases. In table 2 the optimal dynamic feature weights derived using the two databases are tabulated, where the corresponding static weights are the complement values to one.

Optimal dynamic weights in CUDigit

SNR	5dB	10dB	15dB	20dB
White	0.99	0.99	0.99	1.00
Babble	1.00	1.00	1.00	1.00
Car	1.00	1.00	1.00	1.00
Factory	0.99	0.99	1.00	1.00

Optimal dynamic weights in Aurora2

SNR	5dB	10dB	15dB	20dB
Subway	0.94	0.83	0.74	0.64
Babble	0.96	0.95	0.90	0.77
Car	0.96	0.94	0.86	0.72
Exhibition	0.94	0.88	0.66	0.60

Table 2 Optimal weights for the dynamic features in CUDigit and Aurora2 at different SNRs

We observed that the ratio of the optimal dynamic feature weight to static feature weight is much larger for CUDigit than Aurora2. In Aurora2, larger weights are obtained for the dynamic features, but never as large as the weights in the CUDigit, which are overwhelmingly biased toward the dynamic features.

To investigate why such a difference exists between the weights in the Cantonese and English digit databases, we first

check the corresponding recognition error patterns. CUDigit results in more insertion errors than those in Aurora2. This disproportionate distribution is also true in noise–free, clean conditions. The high insertion rate, very possibly, is due to the fact that all Cantonese digits are monosyllabic; the short duration and simple phonetic content also make them prone to insertions, especially in noise. Similar observations have been reported before: "One of the major sources of errors was due to frequent insertions of digit '5', pronounced as a mono-syllabic nasal[ng5], which may be confused with and treated as part of the nasal coda in the digits '0'[ling4] or '3'[saam1]" [7]. For different noises, error patterns show more varieties: in white noise, digits and silence tend to be misrecognized as '3' [saam1] and '4' [sei3]; in babble noise, the major errors are insertions of '5' [ng5] and '0' [ling4]; in car noise, there are many insertions of '5' [ng5]. Different types of noises and SNR's result in different recognition errors, but overall there are many more insertions in CUDigit than in Aurora2. Fig. 3 shows the pie charts of error distributions of baseline results for CUDigit and Aurora2 at 10 dB SNR with babble noise.

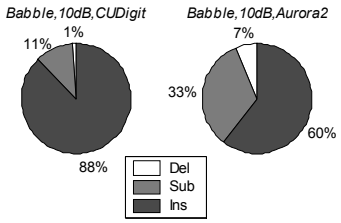

Fig. 3 Baseline error distributions of CUDigit and Aurora2

The sensitivities of the error patterns to SNR's can be further illustrated in Fig. 4 by the bar graphs of deletion, substitution, insertion, and total errors (note that they are in different scales) in babble noise. The errors are shown in different shades for bracketed feature weights. Other three noises, although not plotted here, show similar trends but somewhat different error proportions.

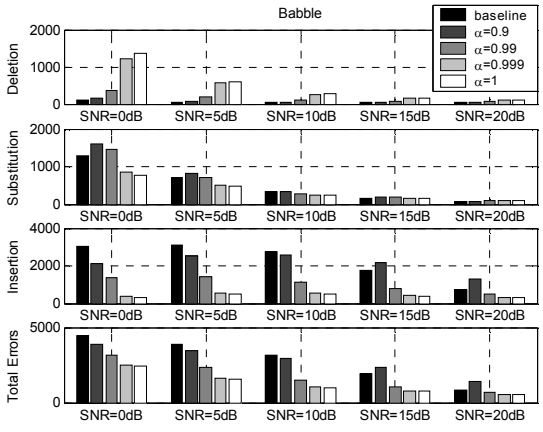

Fig. 4 Recognition error count by using different feature weightings (in babble noise)

By using a larger weighting for the dynamic features, insertions, which constitute majority of recognition errors at low SNRs, are significantly reduced at an expense of more deletions. Substitution errors are also reduced. Overall the total errors are significantly reduced; it confirms that weighting the dynamic likelihood more than its static counterpart is effective for recognizing noisy speech.

8. CONCLUSION

In this paper we investigate the noise robustness of dynamic and static cepstral features in speech recognition. The dynamic feature has been found to be more robust than its static counterpart in noise. Optimal exponential weights are trained in a development set. When tested on CUDigit, a continuous Cantonese digit database, an average of 41.9% relative error reduction is obtained, comparing with the baseline results. Except training the optimal weights with a small amount of development data, the decoder stays the same and no extra computation is needed. Easy training of the compact weights and no need to adapt the clean models (hence no extra decoding effort) to different noise conditions, make the proposed approach an ideal candidate for many potential ASR applications in noise.

ACKNOWLEDGEMENT

This research is substantially 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] Y.-F. Gong, "Speech recognition in noisy environments: a survey," *Speech Communication*, pp.261-291, vol.16, 1995.

[2] 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.

[3] 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.

[4] 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.

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

[6] I. Rogina and A. Waibel, "Learning state-dependent stream weights for multi-codebook HMM speech recognition systems," *Proc. ICASSP-1994*, pp.217-220.

[7] 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.

[8] H.J.M. Steeneken and F.W.M. Geurtsen, "Description of the RSG.10 Noise data-base", Report IZF 1988-3, TNO Institute for perception, Soesterberg, The Netherlands,1988.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] C. Yang, F. K. Soong and T. Lee, "Static and dynamic spectral features: their noise robustness and optimal weights for ASR," submitted to *ICASSP 2005*.