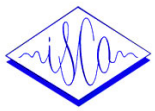

International Symposium on Chinese Spoken Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

A SYSTEM FOR MANDARIN SHORT PHRASE RECOGNITION ON PORTABLE DEVICES

XU Chao[1,2], LIU Yi[1], YANG Yongsheng[1], Pascale FUNG[1], CAO Zhigang[2]

HKUST RandD Corporation, University of Science and Technology, Hong Kong[1]
Department of Electronic Engineering, Tsinghua University, Beijing, 100084[2]

ABSTRACT

With the proliferation of portable devices, speech recognition, especially name, address and command recognition on these devices is a topic of growing relevance. A mandarin short phrase recognition system is introduced in consideration of the limited resources and calculation ability of portable devices. A fixed-point front-end is developed, discrete hidden Markov model is employed for acoustic modeling, and a SNR based likelihood weighting method is proposed to improve the noise robustness of the system. The memory size of the model set is 269kB, the decoding time is 0.89 times of the speech duration, and the method for robustness gives a relative 15.2% word error rate reduction in a complex practical environment with both channel distortion and non-stationary noise presence.

1. INTRODUCTION

As portable devices become more powerful and popular, speech recognition applications on such devices become more attractive and feasible. These applications may include voice dialing, command recognition, name recognition and address recognition, and can be summarized as short phrase recognition. A state-of-the-art automatic speech recognition (ASR) system, e.g., built with HTK [1], can work well for short phrase recognition in noise-free environments on personal computers. However, many challenges still exist when the ASR system is implemented on portable devices which have limited resources and calculation capability and are often used in a changing environment.

The hardware resources of portable devices mainly pose two challenges. First, most portable devices (e.g., Pocket PCs, PDAs and mobile devices) have no floating-point calculation supported. Therefore, directly implementing the PC-based ASR systems to portable devices leads to low computation efficiency [2]. Secondly, context-dependent acoustic models (e.g, triphone and biphone models) are commonly used in current PC-based ASR systems in order to achieve a good recognition performance. The size of triphone or biphone model with multiple Gaussian components often exceeds 1MB. However, the hardware resources in portable devices are very limited compared to PCs. For example, sunplus SPCE chips [3] provide with only 600kB or less ROM, while a discrete biphone model set for Mandarin Chinese requires 1080kB storage memory in our experiments. As a result, model size should be carefully controlled to meet the hardware limitation. Moreover, environmental noise is a key issue in implementing ASR system on portable devices. Compared to PC-based ASR systems, portable devices, e.g., PDAs and mobile phones are often used in different environments, and thus environmental robustness becomes a prominent problem. Although there are many noise robustness techniques, no single method is expected to be equally effective in all noise conditions [4]. For example, adaptation techniques like MAP [4, 5] require the coincidence between the environments of adaptation and testing. Unfortunately, the working environment of the portable devices such as mobile phones is often changing and unpredicted. We hence need a tailored method for this kind of condition.

In this paper, in order to implement ASR systems on portable devices efficiently, we first develop a fixed-point front-end for feature extraction. We employ discrete monophone hidden Markov models (HMMs) in the system due to the limited hardware resources of portable devices. Finally, we propose a SNR based likelihood weighting approach in the decoder to alleviate the effects of environment noise.

The paper is organized as follows. Section 2 describes the system organization. Section 3 describes the fixed-point front-end and acoustic model selection. Section 4 proposes the method of the SNR based likelihood weighting in decoding. Section 5 discusses the experimental results on Mandarin short phase recognition. Finally, we conclude in Section 6.

2. SYSTEM ORGANIZATION

An ASR system consists of three parts: feature extraction, acoustic model training and Viterbi recognizer [1]. The flow chart of the whole system is illustrated in Figure 1. In our mandarin short phrase recognition task on portable devices, a fixed-point front-end is employed, discrete HMM (DHMM) is selected to model the mandarin monophone, and a SNR based likelihood weighting method is introduced into the Viterbi decoding algorithm in the recognizer.

3. FEATURE EXTRACTION AND ACOUSTIC MODELING FOR PORTABLE APPLICATIONS

A fixed-point front-end is developed in our system. Mel-frequency cepstral coefficients (MFCCs) are widely employed by most current ASR systems. Figure 2 illustrates the extraction process of the MFCCs and the logarithm of the frame energy. All these process blocks employ floating-point calculations in

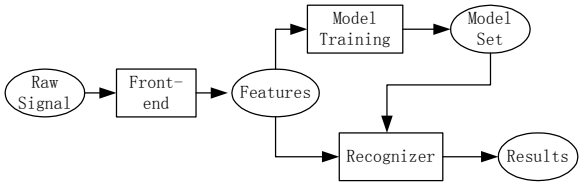

Figure 1: flow chat of the whole system

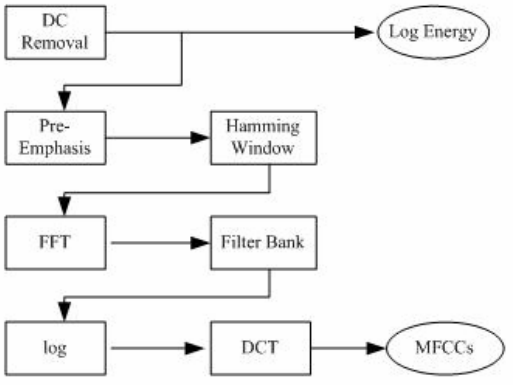

Figure 2: flow chart of the front-end.

PC-based systems like HTK. As most portable devices have no floating-point supported, we employ fixed-point calculations for all these processes. First, we use an approximation algorithm for the logarithm operation. The natural logarithm is first transformed into 2-based logarithm, and then, in 2-based logarithm, the input variable is treated bit by bit:

$$\begin{cases} \log_2 2n = 1 + \log_2 n \\ \log_2 (2n+1) = 1 + \log_2 n + \log_2 (1 + \dfrac{1}{2n}) \end{cases}, \quad (1)$$

where the last term is approximately calculated by:

$$\log_2 (1 + \frac{1}{2n}) \cong \frac{1.17}{2n}. \quad (2)$$

Thus, the logarithm operation can be realized by a series of arithmetical operations in limited steps.

In FFT and DCT, the transform coefficients can be calculated in advance after the vector size has been determined, and thus sine and cosine operations can be avoided.

Through the above simplification for logarithm, FFT and DCT, all the rest calculations are arithmetical operations of floating-point addition, subtraction, multiplication and division, and these operations can be transformed into fixed-point calculations by rescaling or quantization.

According to the output probability, HMM can be divided into continuous HMM (CHMM) and DHMM [1]. The output probability for DHMM is given by $p_j(o_i)$, where j is the state index and o_i is the vector quantization (VQ) index. The output probability for CHMM is:

$$p_j(o_t) = \prod_s \left[\sum_m c_{jsm} N(o_{st}; \mu_{jsm}, \Sigma_{jsm}) \right]^{\gamma_s}, \quad (3)$$

where subscript j is the state index, s the stream index, t the frame index, and m the Gaussian mixture index; γ is the stream weight, and c the Gaussian mixture weight; $N(o; \mu, \Sigma)$ denotes a single Gaussian distribution with the mean vector μ and the covariance matrix Σ. In decoding, each frame of observation will be compared with each possible state, and in every comparison, output probability should be calculated. As a consequence, the time of each output probability calculation is a key cost in the whole decoding period. Another consideration in model selection is the model's phoneme level. Our ASR system is to recognize the dynamically constructed Mandarin short phrase, so phone model rather than word or whole phrase model is preferred. When monophone, biphone and triphone models are compared, the tirphone and biphone model sets both take over 1MB memory. We consequently employ monophone model rather than biphone or triphone.

4. LIKELIHOOD WEIGHTING FOR THE DECODER

As the portable devices are used in various places from office to street, from home to field, environmental robustness for the system becomes a key problem. There exist many noise robustness techniques, but most of them, such as MLLR, PMC and VTC [4, 5], are designed aiming at CHMM and do not fit DHMM. Moreover, these techniques employ some noise assumptions (e.g., stationary noise) or environment constrain (e.g., the test environment matches the adaptation environment). However, the working environment of portable devices such as mobile phones is often changing and unpredicted. The idea of the missing data techniques [6] is not limited in specific noise or specific environment. It use only reliable time-frequency parts of noise-corrupted signal and ignore the unreliable parts in recognition. As time-frequency information is not involved in VQ indexes in DHMM, we use frame level information instead, and propose the SNR based likelihood weighting method that can be applied for both CHMM and DHMM.

A HMM decoder often employs Viterbi algorithm [1], where delta scores of the current feature vector to each possible states are calculated, and a state trace with a maximal final score is marked as the result. The delta score contains two parts, i.e., the transition probability and the output probability. Assuming that the state in the previous frame and the current frame are s_{t-1} and s_t, respectively, the delta score can be expressed as:

$$\Delta = \log P(s_t \mid s_{t-1}) + \log P(x_t \mid s_t), \quad (4)$$

where $P(s_t \mid s_{t-1})$ is the transition probability from s_{t-1} to s_t (it is a general form and the two states may come from different HMMs), and $\log P(x_t \mid s_t)$ is the likelihood of the current observation x_t given the model state s_t.

In missing data techniques, each frequency channel of x_t is assumed independent and marked reliable or unreliable, thus the likelihood is the sum of each channel's likelihood. To use only reliable parts, the likelihood term in (4) is revised and the delta score is given by:

$$\Delta' = \log P(s_t \mid s_{t-1}) + \sum_i \gamma_n(x_t^{(i)}, s_t), \quad (5)$$

where $\gamma_n(x_t^{(i)}, s_t)$ is given by:

$$\gamma_n(x_t^{(i)}, s_t) = \begin{cases} \log P(x_t^{(i)} \mid s_t), & x_t^{(i)} \text{ is reliable} \\ \log \int P(x_t^{(i)} \mid s_t) dx_t, & x_t^{(i)} \text{ is unreliable} \end{cases} . \quad (6)$$

As no time-frequency parts are involved in VQ indexes, we give a reliability measure to each frame according to its SNR, and then introduce the reliability measure as a weight into the likelihood term instead. In this way, the delta score is:

$$\Delta' = \log P(s_t \mid s_{t-1}) + \gamma(x_t, s_t, R_t), \quad (7)$$

where R_t is the frame's SNR in logarithm domain, and $\gamma(x_t, s_t, R_t)$ is given by:

$$\gamma(x_t, s_t, R_t) = w(R_t) \log P(x_t \mid s_t). \quad (8)$$

As better SNR reveals less corruption, an increasing function should be employed for the weight function in (8). Furthermore, the weight function should be easy in calculation in consideration of the decoding speed. Hence, we employ a linear function:

$$w(R_t) = a + b \cdot R_t, \quad (9)$$

where a and b are two constants and are set experientially as $a = 0.5$ and $b = 0.01$, respectively.

The delta score calculation in decoder with the SNR based likelihood weighting applied is illustrated in Figure 3 in comparison with the original delta score calculation.

To estimate the frame SNRs, we employed an iterative process similar to the extended spectral subtraction technique [7], where a Weiner filter incorporated noise estimate is first employed and then the noise estimate is subtracted from the corrupted signal to get the clean signal. Let r_t denote the SNR in linear domain, i.e. $R_t = 10 \log r_t$. In each iteration, the noise energy is estimated first:

$$N_t = (1-\alpha) N_{t-1} + \alpha \frac{X_t}{1 + r_{t-1}} \quad (10)$$

where α is a small positive constant, N denotes the energy of the noise and X of the noisy signal. Then the frame SNR is estimated by:

$$r_t = \max(\frac{X_t}{N_t} - \beta, r_{\min}) \quad (11)$$

where β is a constant, and $r_{\min}$ is a floor value for SNRs.

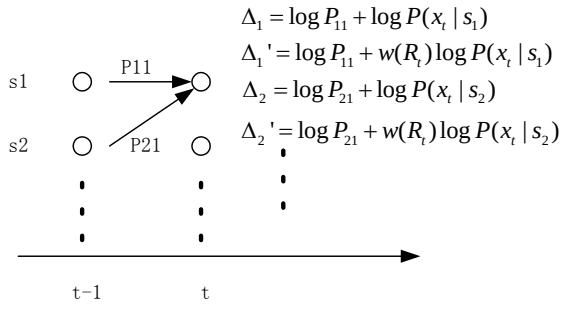

Figure 3: delta score calculation in decoding (Δ_i is the conventional method and Δ_i' the proposed method)

5. EXPERIMENTS AND RESULTS

To show the efficiency of DHMM, we first compare the recognition speed, acoustic model size and the recognition accuracy of using DHMM and CHMM on portable devices. Secondly, the SNR based likelihood weighting scheme is tested to show the environment robustness performance.

5.1. DHMM vs. CHMM

We compare monophone DHMM and CHMM here. As discussed in section 2, we don't employ biphone or triphone as they result in too large model size.

We employ MFCC_E_D_A [1] for model training and recognition. The feature extraction process is illustrated in figure 2, where pre-emphasis coefficient is 0.97; a 200 point Hamming window shifted in 80 point is employed for frame segmentation (8kHz sampled speech); 256 point FFT is applied and the amplitudes of FFT coefficients are analyzed by a 23-channel Mel-frequency filter bank [1]; and after DCT, the cepstral features are liftered with a cepstral lifter factor [1] of 22. Finally, the logarithm of frame energy is incorporated with cepstral features and their delta and accelerate features are calculated both with regression window [1] size of 2.

In CHMM schemes, 64 mandarin monophones and the silence model are all described by 3-state (3 emitting state and two non-emitting states for begin and end) multi-mixture Gaussian HMMs. Left-right model is employed for 64 mandarin monophones and transition probabilities between the first emitting state and the last emitting state of silence model are additionally estimated. Models with Gaussian mixture of 1, 2, 4 and 8 are tested and denoted as 4 different schemes in Table 1, i.e., C-1, C-2, C-4 and C-8 respectively.

In DHMM scheme, MFCC_E_D_A features are divided into four streams. The first three streams are cepstral features, delta cepstral features and accelerate cepstral features. The last stream contains the frame energy and its delta feature and accelerate feature. The VQ codebook takes sizes of 256, 256, 128 and 32 for the four streams respectively. The DHMM scheme employs an identical model topology to that employed in CHMM schemes.

We employed 37326 sentences of Mandarin Chinese speech through telephone channels to training the 64 mandarin monophone models and silence model. In testing, we employed 900 telephone channel utterances from a lexicon of 233 mandarin short phrases.

Table 1: comparison of different model type (C-i for CHMM with i Gaussian mixtures)

	C-1	C-2	C-4	C-8	DHMM
Model Size(kB)	71	137	264	519	269
Word Acc(%)	77.6	80.7	84.7	87.9	83.0
Relative decoding time	1.91	2.65	4.35	7.70	0.89

A comparison of using DHMM and CHMM is shown in Table 1. The first row gives each scheme's model size in kilobyte, the second row gives the recognition results of word accuracy of each scheme, and the last row gives the relative time consumption (decoding time divided by the real time of the utterance) in decoding.

We can see that only the DHMM scheme obtains the recognition results inside the real time. In other schemes, users have to wait at least a period comparable to recording time to get the recognition results. In addition, DHMM scheme achieves a comparable accuracy with C-4 but takes less than a quarter of decoding time.

5.2. Experiments on environmental robustness

Two kinds of noise environment are employed for robustness test. First, we use artificially added Volve noise [8] to represent a steady environment with stationary additive noise. In the second case, we use speech recorded in a noisy office environment and through a Compaq iPAQ Pocket PC to test the system robustness in an environment with both channel distortion and non-stationary noise presence.

The acoustic model set is the DHMM model set used in acoustic model comparison experiments above.

In the stationary additive noise testing, the test data are the above telephone channel speech artificially added with Volve noise. The telephone speech contains noise originally, and we add additional Volve noise at global SNR levels of 10dB, 5dB, 0dB and -5dB (the understandability to human is affected little in all these cases when we listen to the noisy data).

The recognition results are listed in Table 2. The first row gives the SNR levels, the second row gives the word error rates of a baseline system, and the third row gives the word error rates of our system with SNR based likelihood weighting applied (denoted as SNRW). The last column is the average word error rates of the former columns, where the SNRW scheme shows on average a relative decrease of 5% in word error rate. And in each of these cases, SNRW scheme shows overall improvements in noise robustness.

Table 2: Word error rates for additive Volve noise test.

SNR	10dB	5dB	0dB	-5dB	Avg
baseline	19.6	20.1	22.2	28.2	22.5
SNRW	18.8	19.0	21.9	26.0	21.4

Table 3: Results for a complex environment with both channel distortion and non-stationary noise presence.

	Baseline	SNRW
Word Error Rate (%)	23.1	19.6
Relative decrease (%)	0	15.2

The experimental results for a complex practical environment with both channel distortion and non-stationary noise presence are given in Table 3. As the training data is recorded through the telephone channel and the test data is recorded with the Compaq iPAQ Pocket PC, there exists remarkable channel mismatch. The test speech is spoken in a noisy office. There exists stable noise like that of computers and air-conditions, and also non-stationary noise like background babbles and discussing. In the recognition results, we can see that the SNR based likelihood weighting scheme shows a good performance of relative decrease of 15.2% in word error rate compared with the baseline system.

6. CONCLUSIONS

In this paper, we present a system for Mandarin short phrase recognition on portable devices. Due to the limited CPU capability of the portable devices, we develop the fixed-point front-end for feature extraction. Due to the limitation of memory size, we use DHMM for acoustic modeling. Compared to using float-point calculation and CHMM, we achieve comparable recognition accuracy and greatly reduce the computation time. To alleviate the effects of environment noise, we further propose the SNR based likelihood weighting approach in the decoder. With this method, we obtain a relative decrease of 15.2% in word error rate in a complex real-world environment with both channel distortion and non-stationary noise presence.

7. ACKNOWLEGEMENTS

This work is supported by Grant# S/P584/03A of the Innovation & Technology Commission of the Hong Kong government.

8. REFERENCES

[1] S. Young, J. Jansen, J. Odell, D. Ollason, and P. Woodland. The HTK Book. Entropic, 1999.

[2] S. Deligne, etc., "A Robust High Accuracy Speech Recognition System for Mobile Applications", IEEE Trans. on speech and audio processing, pp.551-561, vol.10, Nov, 2002.

[3] http://www.sunplus.com.tw/products/speechandmusic/spce.asp

[4] G. M. Davis, Noise Reduction in Speech Applications, CRC Press, FL, 2002.

[5] X. Huang, A. Acero and H-W. Hon, Spoken Language Processing: A Guide to Theory, Algorithm and System Development, Prentice Hall PTR, New Jersey, 2001.

[6] M. Cooke, A. Morris, and P. Green., "Missing data techniques for robust speech recognition", In Proc. of ICASSP'97, Munich, April 1997.

[7] P. Sovka, P. Pollak, and J. Kybic. "Extended Spectral Subtraction", In Signal Processing VIII Teories and Applications, volume 2, pp.963-966, EUSIPCO-96, Sep. 1996.

[8] A.P. Varga, H.J.M. Steeneken, M. Tomlinson, and D. Jones, "The noisex-92 study on the effect of additive noise on automatic speech recognition", Tech. Rep., Speech Research Unit, Defence Research Agency, Malvern, U.K., 1992.