A NOVEL ROBUST SPEECH RECOGNITION ALGORITHM BASED ON MULTI-MODELS AND INTEGRATED DECISION METHOD *

Pan Shengxi Liu Jia Jiang Jintao Wang Zuoying Lu Dajin

Electronic Engineering Department of Tsinghua University

Beijing, P. R. of China (100084)

ABSTRACT

In this paper, a new robust speech recognition algorithm of multi-models and integrated decision(MMID) is proposed. A parallel MMID(PMMID) algorithm is developed. By using this new algorithm the advantages of different models can be integrated into one system. This algorithm uses different acoustic models at the same time based on DDBHMM (duration distribution based Hidden Markov Model)[2]. These different models include the channel-mismatch-correct(CMC) model, more-alternative-pronunciation model, tone and non-tone models of Chinese Mandarin speech, voice activity detection(VAD) model and state-skip model. The speech recognition accuracy of the multi-model system is better than that of single-model system in the adverse environments. The experimental results show that the error rate of the recognition system is 2.9% and reduced by 81% compared with the baseline system of the single-model.

1. INTRODUCTION

The research of robustness is a main problem in automatic speech recognition (ASR)[1]. The varieties of environment that influence recognizer are mainly that: (1) noise, (2) channel mismatch, and (3) articulation effect. For example, the channel mismatch in telephone speech recognition, the different accents and speech rate, et al.

The methods of robust speech recognition are currently divided into three categories [1]. Although these different methods are effective for solving the problems of robust speech recognition, the integrated characteristics of recognition algorithm in different environments and the multi-model problems of recognition system are less considered. In this paper a new robust speech recognition algorithm with multi-models and integrated decision (MMID) is proposed, and different models are integrated into one system based on the duration distribution based HMM (DDBHMM)[2]. The channel mismatch and some articulation effects are mainly concerned in this MMID algorithm. The new algorithm shows good performance of recognition in the adverse environments.

The paper is arranged as follows: Section 2 proposes the formula of MMID based on maximum a posterior (MAP) and Section 3 gives a speech recognition system based on Parallel MMID(PMMID). The experimental results and conclusions are shown in Section 4 and 5.

II THE ROBUST MMID ALGORITHM BASED ON MAP

Let $\boldsymbol{W} = W_1, W_2, \cdots, W_M$ represent the Chinese characters (words). An utterance $\boldsymbol{A}$ consists of M pronunciations of $A_1, A_2, \cdots, A_M$, where A_m consists of consonant C_m, vowel V_m and tone T_m ($m = 1,2,\cdots, M$). And $\boldsymbol{A}$ includes T frames of acoustic features $\boldsymbol{E} = e_1, e_2, \cdots, e_T$. The task of speech recognition is to establish the relationship of $\boldsymbol{A}$ and the recognition result $\hat{\boldsymbol{W}}$. That is choosing the recognition result by MAP of intersection set of $\boldsymbol{A}$, $\boldsymbol{W}$ [5].

$$\begin{aligned}\hat{\boldsymbol{W}} &= \arg\left\{\max_{(\boldsymbol{W},\boldsymbol{A})} P(\boldsymbol{WA/E})\right\} \\ &= \arg\left\{\max_{(\boldsymbol{W},\boldsymbol{A})} P(\boldsymbol{WAE})\right\} \\ &= \arg\left\{\max_{(\boldsymbol{W},\boldsymbol{A})} P(\boldsymbol{W})P(\boldsymbol{A/W})P(\boldsymbol{E/A})\right\}\end{aligned} \quad (1)$$

We call $P(\boldsymbol{W})$, $P(\boldsymbol{A/W})$ and $P(\boldsymbol{E/A})$ as *language* model, *confusion* model, and acoustic model respectively[5], which consider the last two models in this paper.

If the successive acoustic features are un-correlative and probabilities of different tones and accents are equal, then the integrated Formula (2) of single-model can be led out[5] from Formula (1):

* This project is supported by National Natural Science Funds (Item: 69772020), National 863 High Technology Projects (Contract No. 863-306-03-02-1), and 211 Engineering Project (Item: Man-machine Interactive Support Environment).

$$\hat{\boldsymbol{W}} = \arg\left\{\max_{(\boldsymbol{W},\boldsymbol{A})} P(\boldsymbol{W})P(\boldsymbol{A}/\boldsymbol{W})P(\boldsymbol{E}/\boldsymbol{A})\right\}$$
$$= \arg\left\{\min_{(\boldsymbol{W},\bar{N}_C,\bar{N}_V)}\left\{l(\boldsymbol{W}) + \sum_{m=1}^{M}[l(C_m/W_m) + l(V_m/W_m) + l(T_m/W_m)] + M\cdot l(d)\right.\right.$$
$$\left.\cdot(\bar{N}_C+\bar{N}_V) + \sum_{m=1}^{M}\left\{\left\{\sum_{i=1}^{\bar{N}_C}\sum_{t=d_{m_C}^{i-1}+1}^{d_{m_C}^{i}} l'(e_{St})/C_m^i\right)\right\}\right.$$
$$\left.\left.\left.+\left\{\sum_{i=1}^{\bar{N}_V}\sum_{t=d_{m_V}^{i-1}+1}^{d_{m_V}^{i}} l'(e_{St})/V_m^i\right)\right\}+\right\}\right\}\right\} \quad (2)$$

where $l(a)\overset{\Delta}{=}-\lg P(a)$. $P(d)$ is a uniform duration distribution over appropriate range of duration. Formula (2) includes five models.

(1) State-skip model: When speech rate is fast, the co-articulation is very strong and some pronunciations are lost. So the state-skip model and non-state-skip model are used in our system at the same time. here $1\le \bar{N}_V \le N_V$ • $1\le \bar{N}_C \le N_C$ • N_C and N_V represent the standard state number of Initial and Final.

(2) and (3): Tones and more-alternative-pronunciations model: The pronunciations of the Chinese Mandarin with the different accents are considered. The basic idea of this new function is that the system ties some easy confusing elemental pronunciations into a new unit and allows them to have the different pronunciations. For example, some people pronounce “zh” is equal to “z”, and “sh” is equal to “s”, and the like. Because Chinese Mandarin is a speech with the tone, the tone and non-tone are regarded as two models in our system. The formula can be expressed as:

$$l'(e_t/\Lambda) = \min_{(A,W)}\{l(e_t/\Lambda)\} \quad (3)$$

where, $l(e_t/\Lambda)$ is the likelihood, which does not consider the tones and more-alternative-pronunciations, and $l'(e_t/\Lambda)$ is the minimum likelihood of possible tones and/or more-alternative-pronunciations with the template Λ.

(4) CMC model: This is a model based on DDBHMM with an algorithm of correcting channel mismatch. The algorithm has published in [4] which. can be expressed as:

$$e_S = e_Y - e_H \quad (4)$$

where e_H is the estimation of channel parameter[4].

(5) VAD model•VAD is voice activity detection. It is mainly used for voiced and unvoiced detection and operates on time domain with some adapting methods. In our system the VAD algorithm of Recommendation G.723.1 of ITU[3] is applied for detecting silence. By some small modification the VAD model works very well in our ASR system. The search space of recognition decoding is reduced by VAD model because the frames of silence are deleted.

For the integrated formula of single model above, the different combinations of five models (for example, CMC model, tone and non-tone model, et al.) may be selected by program, and the combination number is 32($=2^5$), we call one of them as a single-model algorithm. The realization of each single-model of different types is a one-pass Viterbi searching processing using Formula (2). Generally speaking, The different types of Formula (2) could lead out different results due to different understanding for the physical process of speech (for example, state jumping or no state jumping), and different considering of the time and space when realization of the recognition process. These special algorithms were called single-model algorithm $f_i(i=1,2,\cdots,32)$. If the single-model algorithm set of $\boldsymbol{F}=\{f_1, f_2,\cdots, f_n,\cdots f_{32}\}$ is known, more information can be utilized to obtain the result by using the criterion of MAP, which can be expressed as:

$$\hat{\boldsymbol{W}} = \arg\left\{\max_{(\boldsymbol{W},\boldsymbol{A},\boldsymbol{F})} P(\boldsymbol{WAF}/\boldsymbol{E})\right\}$$
$$= \arg\left\{\max_{(\boldsymbol{W},\boldsymbol{A},\boldsymbol{F})}\left[P(\boldsymbol{F}/\boldsymbol{E})P(\boldsymbol{WAE}/\boldsymbol{F})/P(\boldsymbol{E})\right]\right\} \quad (5)$$
$$= \arg\left\{\max_{(\boldsymbol{W},\boldsymbol{A},f_i)}\left[P(f_i/\boldsymbol{E})P(\boldsymbol{WAE}/f_i)\right]\right\}$$

This is the basic formula of MMID. The second term in Formula (5) could be one kind of single-model algorithm in Formula (2). And now we consider the problem of estimating $P(f_i/\boldsymbol{E})$. Because $P(f_i/\boldsymbol{E})$ is the probability of selecting algorithm f_i under the feature $\boldsymbol{E}$, the simplest estimation method is to set probability equally according to the criterion of maximum entropy. But $P(f_i/\boldsymbol{E})$ can be estimated from training database as follow: Each of *N* utterances in training database can be recognized using the algorithm $f_i(i=1,2,\cdots K_F)$, where $K_F\le 32$. If algorithm f_i is used for k_i times according to a strategy, then $P(f_i/\boldsymbol{E})$ can be estimated by frequency:

$$P(f_i/\boldsymbol{E}) \approx \frac{k_i}{N},\ i=1,2,\cdots,K_F \quad (6)$$

Selecting different algorithms has different methods. The algorithm with maximum likelihood and best word recognition

rate of the utterance is chosen.

III THE PARALLEL REALIZATION OF SPEECH RECOGNITION WITH MMID

In this Section, a parallel MMID(PMMID) algorithm of frame synchronism is developed, which uses beam searching algorithm under the guidance of Formula (5) with a parallel Viterbi searching processing, and tree structure language model of limited vocabulary commands[5]. The command number is about 16,000 and the recognition result is the key words. We give a simplification diagram of PMMID where the detailed algorithm can be found in [5].

Step 1. For a given utterance, Initial the structure of searching path. For the same possible choice with different single-models, initial it $K_F(\leq 32)$ times with different single-model parameters. $t=1$.

Step 2. For each active path, do the Viterbi searching within a Unit (for example, word) according to the parameters of each single-model.

Step 3. Check each active path to determine if it can jump from the current Unit to next Unit. If jumping, insert a new path to active path; else goto next Step.

Step 4. If the searching meets the end of utterance, that is $t=T$, then give the first choice result and searching processing ends; else $t=t+1$, goto Step 1

The basic idea of PMMID is that it is a frame synchronism processing when all single-models search recognition results. And due to this reason, the searching time can be saved and the recognition rate do not degrades.

IV EXPERIMENT RESULT

The following experiments has been done to verify the algorithm of MMID proposed in this paper. The cepstrum mean subtraction(CMS) algorithm is used in this system for increasing robustness of the system. Two Chinese Mandarin speech databases are used as training data in our experiment. One is utterances of 699 phrases (60 males and 40 females) designed by our laboratory; the other is utterances of 520 sentences (38 males and 38 females) provided by national 863 projects. All training data are passed through real telephone channels. The testing data are the utterances (2 males and 2 females) of 207 command sentences for common telephone, which was recorded in our Lab. Each of four students read 207 command sentences through real telephone channels. All of testing data pass the different telephone channel from that of training data.

In order to research the performance of MMID, recognition results of 32 different combination of single-models are collected according to Formula (2). We do serial simulated experiments of MMID by selecting $K_F(K_F=1,2,\cdots 32)$ from 32 single-models according to Formula (5), and find the optimal results and average results of MMID. At the same time, the lowest limit of error rate of MMID is given, which can be estimated when all K_F single-models can not recognized the command correctly. From Fig. 1, the following conclusion can be obtained. (1) The error rate decreases quickly when the number of single-model increases from 1 to 5, and the error rate keep almost unchanged when the number of single-model exceed 5. It shows the advantage of MMID. (2) The optimal error rate has some distance from lowest limit error rate, even if it reaches the lowest limit of error rate, which still has scope to improve the MMID algorithm.

The recognition error rates are compared in detail according to the different conditions, such as different single model, different local accents and speech rates, et al. Table 1 shows the experimental results. Baseline is the standard algorithm. Five single-model algorithms are the special simplification of Formula (2). Due to using CMS, the CMC model has only some improvement of error rate compared with baseline system[4]. The more-alternative-pronunciation model is effect on some people with accents; Tone model have good effect for all testing data; State skip model has good effect only for the data of faster pronunciation of file No. 2; VAD model has some improvement compared with baseline. These algorithms do not change greatly the average error rate due to their specific consideration for specific data type. The error rate of five single-model integration with Formula (2) has even a little degradation compared with baseline due to larger searching data domain of this algorithm.

PMMID algorithm is a parallel-integrated decision algorithm from K_F single-models. The realization is a frame-synchronism Viterbi algorithm[5]. Fig. 1 and Table 1 all show that the advantage of different single models could be complemented each other. PMMID gets the best results. Using the parallel structure, the error rate is reduced 81% compared with baseline. The experiment verifies that MMID is an effective algorithm.

V CONCLUSION

In this paper, a new robust speech recognition algorithm of MMID based on MAP is proposed, and the channel mismatch, more alternative pronunciation, tone, speed of pronunciation and VAD are considered in this algorithm. The PMMID algorithm is developed. The experimental result show MMID can process the channel mismatch and some articulation effects well and show that the error rate of the recognition system is 2.9% and reduced by 81% compared with the baseline system of the single-model.

REFERENCE

[1] Mazin G. Rahim & Biing-Hwang Juang,“ Signal Bias Removal by Maximum Likelihood Estimation for Robust Telephone Speech Recognition” , IEEE Trans. on Speech & Audio Processing, pp19-30, Vol. 4. No.1, January 1996

[2] Wang zuoying, Gao Hongge, “An Inhomogeneous HMM Speech Recognition Algorithm”, Chinese journal of electronics, 1998.1, Vol.7, No.1, pp.73-77.

[3] ITU-T Annex A to Recommendation G.723.1 commendation

G.723.1.

[4] Pan Shengxi, Liu Jia et al., "The Channel Correct Algorithm signal processing conference on speech, image and communication, Zheng Zhou City, China, Oct. 1inese national signal processing conference on speech, image and communication, Zheng Zhou City, China, Oct. 1997. pp. 146-149.

Based on DDBHMM," The 8th Chinese national

[5] Pan Shengxi, " The Research of Robust Speech Recognition Based on Telephone Speech", the doctoral dissertation of Department of Electronic Engineering, Tsinghua University, Beijing, China, July 1998.

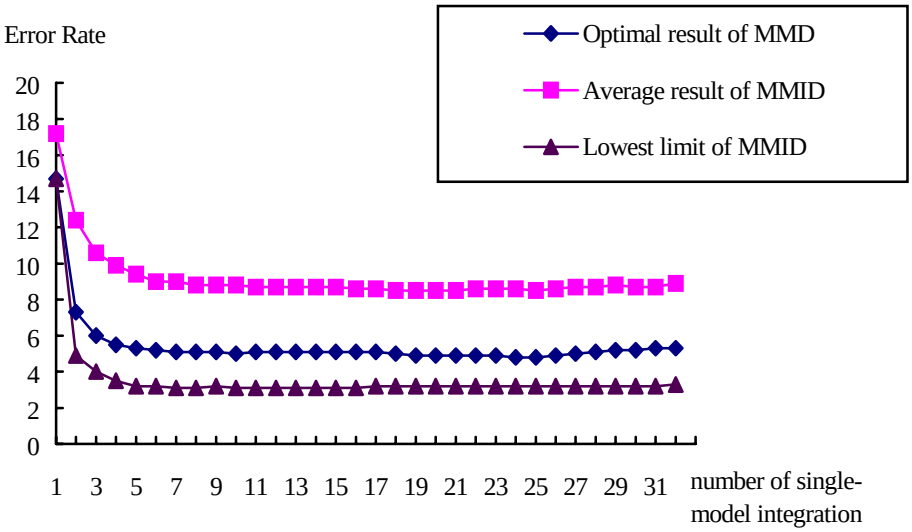

Fig. 1 The serial simulation experiments of MMID with Formula (5)

Table 1. Description of testing data and their recognition results (Error rate: %)

File No.	1	2	3	4	AVERAGE
Sex	Female	male	Female	male	
Normalized speed of speech	1.0	1.27	1.06	1.16	1.12
Baseline with Formula (2)	20.3	19.8	12.6	8.2	15.2
CMC model with Formula (2)	21.3	18.0	10.1	7.7	14.3
VAD model with Formula (2)	15.0	20.3	12.6	7.2	13.8
more-alternative-pronunciation with Formula (2)	25.1	16.9	12.0	6.8	15.2
Tone model with Formula (2)	16.4	17.9	10.1	6.3	12.7
State-skip model with Formula (2)	31.9	12.1	15.9	8.2	17.0
Five single-model integration with Formula (2)	30.0	14.0	15.9	4.8	16.2
PMMID algorithm with Formula (5) $K_F = 5$	3.4	4.8	2.4	1.0	2.9