EuroSpeech'99, Vol. 3, pp.1255-1258, Budapest, Hungary, Sept. 1999

A FAST AND EFFECTIVE STATE DECODING ALGORITHM

Mingxing Xu, Fang Zheng, Wenhu Wu
Speech Laboratory, Department of Computer Science and Technology, Tsinghua University
Beijing, 100084, P.R. China
xumx@sp.cs.tsinghua.edu.cn

ABSTRACT

In this paper a fast and effective algorithm named equal feature variance sum (EFVS) frame-synchronous searching is presented for state decoding. EFVS controls the state transition by using only the feature variance of the speech, instead of by using the state dwell distribution. The basic hypothesis of this new algorithm is the equality of feature variance sum in each state of the speech. Given the boundaries of the speech recognition unit (SRU), EFVS can generate the state sequence without dynamic searching. In a continuous speech word recognition system, this novel algorithm reduces the error rate by 36.8% and speed up the system by 65.6% compared with the traditional state decoding methods.

Keywords: State Decoding, Frame-Synchronous Searching, Equal Feature Variance Sum

1. INTRODUCTION

The state decoding algorithms based on HMM[1], not only the traditional frame-synchronous searching (FSS) algorithms[2], but also some modified Viterbi algorithms, are mainly based on dynamic programming theory. This theory does not perfectly fit the speech recognition. Firstly, these algorithms try to get the best matching between the utterance to be recognized and the acoustic model without considering whether they belong to the same speech recognition unit (SRU). In such matching one speech feature sequence may correspond to a few different state sequences considered as the best results. Therefore, some recognition errors are produced. Secondly, these algorithms are dependent on the exponential distribution of state dwell which may not be suitable speech recognition.

Moreover, the left-to-right HMM has some shortcomings, such as the last state of the SRU is a trap for the state transition, the state transition matrix is unable to describe the speech variance well.

We propose a novel algorithm for state decoding to resolve the above problems. The basic hypothesis of this new algorithm is the equality of feature variance sum in each state of speech. Under the hypothesis, any transition between two states is only controlled by the average feature variance sum of the speech feature sequence without using state dwell distribution. Because there is only one unique sequence of feature in the state decoding, the unique state decoding result will be obtained for later use in recognition matching. Given the boundaries of the SRUs, a simplified algorithm based on EFVS is presented, which can determine the state of each frame by calculating the average feature variance sum of state, instead of by dynamic searching. Therefore the recognition is made faster.

In addition, EFVS uses only the feature variance, thus not only it is fast and effective, but also it can deal with the speech with variable speed or wizened utterances.

This paper is organized as follows. In Section 2 the general methods for state decoding are discussed. and the basic theory of EFVS state decoding algorithm is described in details. In Section 3 the searching rules of EFVS are presented. A simplified algorithm for EFVS method is given in Section 4. Experimental results and conclusoions are provided in Section 5 .

2. STATE DECODING METHOD TYPE

In order to use the information of state duration, some state decoding algorithms different from classical Viterbi algorithm are used in many modified HMMs. We can classify these algorithms into several classes according to the way how the state duration information is introduced and the strategies of searching are used.

(1) State dwell information described by a distribution, state transition controlled by probability.

Methods of this type use the parameters of distribution to describe state through modeling the state dwell distribution[1]. In the state decoding processing, the probability of the current state dwell and the occurance probability the of feature vector in the state affect the transition and dwell of the current state. The state sequence with the max probability score is considered as the best one. Some of these methods describe the state dwell distribution by the statistical information of speech signal[3][4][5]. And some methods do not assume the homogeneous hypothesis of HMM[6].

(2) State dwell information described by a range of duration, state transition controlled by the relationship between the state duration and the range.

These methods get the range of duration by statistics and assume the state dwell obeys the uniform distribution. In controlling of the state transition, the range of the current state dwell which controls whether the current state should stay is determined dynamically by the history of the duration information of the states passed[7].

The novel state decoding algorithm presented in this paper uses the variance of speech features to reflect the information of state dwell and control the state transition. It is very different from above methods.

3. EQUAL FEATURE VARIANCE SUM (EFVS) STATE DECODING ALGORITHM

3.1. Basic Theory

Speech signal is assumed to be a stable signal during a short-term period. Each state of the acoustic model must be equal to or have the same ability to describe the speech varying. So a hypotheses is provided.

Basic Hypotheses: The best state sequence is the one that makes every state of the acoustic model have the equal feature variance sum.

According to the hypotheses, a novel state decoding algorithm is presented named Equal Feature Variance Sum state decoding algorithm, in short EFVS. In order to describe the algorithm, we deduce some properties.

Suppose the state number of the SRU is *N*, and the state space is $\{S_0, S_1, \cdots, S_{N-1}\}$. Let the feature vector sequence be $\{o_0, o_1, \cdots, o_{T-1}\}$, and the length of the SRU in frame be *T*. The feature variance (FV) at the *t*-th frame is defined as

$$\Delta_t = \left\| o_{t+1} - o_t \right\|, \quad 0 \le t < T-1 \tag{1}$$

At the *t*-th frame, the partial accumulated feature variance sum is

$$FVS(t) = \sum_{k=0}^{t} \Delta_k \tag{2}$$

and the number of states passed is $STATE(t)$. The feature variance sum of state *s* is

$$SFVS(s) = \sum_{t=t_1}^{t_2} \Delta_t = FVS(t_2) - FVS(t_1 - 1) \tag{3}$$

where t_1 and t_2 are the boundaries of state *s*.

Corollary 1: Let

$$\Delta_{state} = \frac{1}{N} \sum_{t=0}^{T-1} \Delta_t = FVS(T-1)/N \tag{4}$$

then

$$SFVS\ (s) \ge \Delta_{state}\ , \quad 0 \le s \le N-1 \tag{5}$$

Corollary 2: The last frame of the SRU must be dealt with specially.

According to the definition of FV, we can not calculate the FV for the last frame because the next frame belongs to another SRU. Let

$$\Delta_{T-1} = 0 \tag{6}$$

then we get

$$FVS(T-2) = FVS(T-1) \tag{7}$$

This will not be happened at other frames.

Corollary 3: If the length of unit is unknown, the first state must be dealt with specially.

Firstly, the first state can not be skipped for its being responding to the beginning of pronounce. Secondly, we can not use Corollary 1 to get Δ_{state} when the length of the SRU is unknown. We must control the transition of the first state by using the first statistical information obtained from mass data.

Corollary 4: $\sum_{s=0}^{n} SFVS\ (s) \ge n * \Delta_{state}$ (8)

Corollary 5: The state skipping exists.

If $FVS(t) \ge (n+1) * \Delta_{state}$ (9)

and $FVS(t-1) < n * \Delta_{state}$ (10)

From Eq. (8), we get

$$STATE(t) \ge n+1 \tag{11}$$

and $STATE\ (t-1) < n$. (12)

From Eq. (11) and (12), it is clear that the *n*-th state is skipped.

Corollary 6: The last state of the SRU is not a trap for the state transition.

Since *N* and Δ_{state} are limited, there exists *t* such that $FVS\ (t) \ge N * \Delta_{state}$ and the frame belongs to the last state. Therefore the SRU will end at the *(t+1)*-th frame normally.

Corollary 7: If a state transition occurs at the *t*-th frame,

$$FVS(t-2) < STATE\,(t-1) * \Delta_{state} \le FVS(t-1) \tag{13}$$

Since a state transition occurs at the *t*-th frame, *STATE(t-1)* states have been passed at the *(t-1)*-th frame. From Eq. (8), we have

$$FVS\ (t-1) \ge STATE\ \ (t-1) * \Delta_{state} \tag{14}$$

Following is a proof for the left part of the inequation (13) with reduction to absurdity.

If $FVS(t-2) \ge STATE\,(t-1) * \Delta_{state}$ (15)

From Eq. (8) we get

$$STATE(t-2) \geq STATE(t-1) \quad (16)$$

Since the state sequence is ascend, it is clear that

$$STATE(t-2) \leq STATE(t-1) \quad (17)$$

So

$$STATE(t-2) = STATE(t-1) \quad (18)$$

From this equation and (15),

$$FVS\,(t-2) \geq STATE\,(t-2) * \Delta_{state} \quad (19)$$

This shows that the state of the *(t-1)*-th frame is different from the previous frame according to Corollary 4, that is

$$STATE(t-2) \neq STATE(t-1) \quad (20)$$

This equation is conflict with (18), so

$$FVS\,(t-2) < STATE\,(t-1) * \Delta_{state} \quad (21)$$

3.2 Searching rules

Let $\{o_0, o_1, \cdots, o_{L-1}\}$ denote the feature vector sequence of the current speech segment, where *L* is the length of the speech segment which may contain one or more SRUs.

When the length of the SRU feature vector sequence is unknown, we can not use Corollary 1 to calculate Δ_{state}. In the realizing of EFVS state decoding algorithm, according to Corollary 3, the first state transition is controlled by the state duration range which can be obtained from mass data by means of the statistical method.

Suppose the first state ends at the *t*-th frame, namely

STATE(t)=1 and *STATE(t+1)=2*.

According to Corollary 7, we can easily obtain

$$FVS\,(t-1) < \Delta_{state} \leq FVS\,(t),$$

here we define $FVS(-1) = 0$ when the duration of the first state is one frame. Thus we get a dynamic variable range of Δ_{state} at the *t*-th frame which is denoted by $RANGE(t)$.

Let

$$RANGE\,(t) = (R_{\min}(t),\ R_{\max}(t)].$$

It is clear that

$$R_{\min}\,(t) < \Delta_{state} \leq R_{\max}\,(t).$$

According to Corollary 7, we have

$$R_{\min}\,(t) = FVS(t-1), \quad R_{\max}\,(t) = FVS(t).$$

We consider all cases at the *(t+1)*-th frame.

Case 1:

$$FVS\,(t+1) \geq STATE\,(t) * R_{\max}\,(t).$$

From this equation and Corollary 4, we obtain

$$RANGE(t+1) = RANGE(t)$$

and

$$STATE(t+1) > STATE(t).$$

Therefore the *(t+1)*-th frame belongs to a different state from the *t*-th frame. This case is corresponding to (*I)* in *Figure 1*.

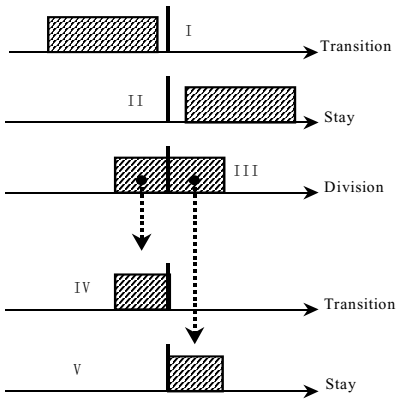

Figure 1. Rules for state decoding

Here ▨ denotes $(STATE\ (t) * R_{\min}(t),\ STATE\ (t) * R_{\max}(t)]$ and | denotes $FVS\ (t+1)$

Case 2:

$$FVS(t+1) \leq STATE(t) * R_{\min}(t).$$

From this equation and Corollary 4, we get

$$RANGE(t+1) = RANGE(t)$$

and

$$STATE(t+1) = STATE(t).$$

Then the *(t+1)*-th frame will stay at the current state. This case is corresponding to (*II)* in *Figure 1*.

Case 3:

$$STATE\ (t) * R_{\min}(t) < FVS\ (t+1)$$

and

$$FVS\ (t+1) \leq STATE\ (t) * R_{\max}(t).$$

In this case, we must divide $RANGE(t)$ into

$$RANGE(t) = RANGE\_L(t) \bigcup RANGE(t).$$

This case is illustrated by (*III)* in *Figure 1*.

Here

$$RANGE\_L(t) = (R_{\min}(t),\ FVS(t+1) / STATE(t)]$$

and

$$RANGE\_R(t) = (FVS\,(t+1) / STATE\,(t),\ R_{\max}(t)].$$

Then $RANGE\_L(t)$ and $RANGE\_R(t)$ are used to control the state dwell and transition of the *(t+1)*-th frame, respectively. A state transition is given by $RANGE\_L(t)$ which is corresponding to (*IV)* and a state dwell is given by $RANGE\_R(t)$ which is corresponding to (*V)* in *Figure 1*.

Case 4:

$$N * R_{\min}(t) < FVS(t+1) \le N * R_{\max}(t).$$

Then the *(t+2)*-th frame is the last frame of the SRU.

Case 5:

$$FVS(t+1) > N * R_{\max}(t).$$

Then the searching fails, which indicates the duration of the first state is too long.

Case 6:

$$FVS(L-2) \le N * R_{\min}(t).$$

Then the searching fails, which indicates the duration of the first state is too short.

3.3. Simplified algorithm

When the boundaries of the unit are known, a simplified algorithm is deduced which can generate the state sequence of the SRU feature vector sequence without dynamic searching during the state decoding.

Let T be the length of the SRU feature vector sequence. It is easy to obtain Δ_{state} by

$$\Delta_{state} = \frac{1}{N}\sum_{t=0}^{T-1}\Delta_t .$$

Let the number of feature vectors in the first n states be L_n, where $1 \le n \le N-1$. If there exists k such that

$$\sum_{t=0}^{k-1}\Delta_t < n * \Delta_{state} \le \sum_{t=0}^{k}\Delta_t ,$$

then $L_n = k$ which indicates that L_n is the boundary between state n and $n+1$. Apparently, $L_N = T-1$ is satisfied. So the processing of searching boundary between two states is changed to searching the number of feature vectors in the first n states.

4. EXPERIMENTAL RESULTS AND CONCLUSIONS

The experiment is made across a continuous speech word recognition system with 24,667 basic words and many user words. The system acoustic model is CDCPM[8] based on syllable units with 6 states and 16 mixed density distributions in each state. The test data contains 204 Chinese words with 2-4 syllables uttered by 5 untrained men.

Table 1 shows that when the number of searching paths increase, the system is made slower by 26.3% for EFVS method and 64.9% for FSS method. Compared to FSS, EFVS reduces the error rate by 36.8% and speed up the system by 65.6%.

EFVS algorithm has following advantages:

(1) the state deconding speed is fast and the hit rate is high;

(2) the state dwell distribution fits for the speech;

(3) the last state of SRU is not a trap for the state transition;

(4) the state decoding results reflect the rules of the speech varying;

(5) the variable speech speed can be adapted to;

(6) the wizened utterances can be dealt with.

5. REFERENCES

[1] Rabiner L R. (1989) A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proc. of the IEEE* Vol. 77, No.2, Feb. 1989 pp. 257~285

[2] Lee C H, Rabiner L R. (1989) A frame synchronous network search algorithm for connected word recognition. *IEEE Trans. ASSP*, Nov. 1989, 37(11): 1649-1658

[3] Russell M J, Moore R K. (1985) Explicit Modeling of state occupancy in Hidden Markov Models for Automatic Speech Recognition. *Proc. ICASSP'85* pp. 5~8, Mar. 1985

[4] Levinson S E. (1985) Structural Methods in automatic speech recognition. *Proc. IEEE,* Vol. 73, No. 11, pp. 1551~1558, Nov. 1985

[5] Levinson S E. (1986) Continuously variable duration hidden Markov models for automatic speech recognition. *Computer Speech and Language,* Vol.1 No.1 pp.29~45 Mar. 1986

[6] Wang Z Y, (1993) Inhomogeneous HMM for Speech Recognition and THED Recognition and Understanding System. *Telecommunication Science,* Vol. 9, No.4, July 1993, pp. 31-36 (in Chinese)

[7] Zheng F, Song Z J, Xu M X, et. al. (1999) Research and implement on technologies of Chinese Speech Dictation Machine. *Journal of Software* , Apr. 1999. (in Chinese)

[8] Zheng F, Chin H X, Shi Z J, Wu W H, Fang D T. (1997) A real-world speech recognition system based on CDCPMs. *In '97 Int. Conf. Computer Processing of Oriental Languages (ICCPOL'97)*, Apr. 2, 1997, Hong Kong, 1: 204-207.

Table 1. The word recognition rate of different state decoding algorithms

Algorithm	Path Num.	Speed(s/w)	word recognition rate(%)							
FSS	800	3.964	65.7	67.2	68.1	68.1	70.1	70.1	70.1	70.1
	1200	6.432	65.2	66.7	67.7	67.7	69.1	69.6	69.6	69.6
EFVS	800	1.363	82.4	85.8	87.3	89.2	89.7	89.7	89.7	90.2
	1200	1.722	82.4	86.8	88.2	89.2	89.7	90.2	90.7	91.2