Analysis and Evaluation of Continuous Speech by Simultaneous Spotting of Phonemes, Words and Phrases

Yasuo ARIKI and Mitsuyasu SAKURAI
Department of Electronics and Informatics,
Faculty of Science and Engineering,
Ryukoku University, Seta, Otsu, Shiga, Japan 520-21
e-mail: ariki@rins.ryukoku.ac.jp

1. Introduction

In human speech recognition, we can discriminate known words and unknown words contained in natural speech. This indicates following two points. Firstly we can easily locate or spot important known words and then understand the sentence meaning by skipping unknown words. Secondly we can also learn phoneme sequence of unknown words from their utterances and sometimes learn their meaning from the context.

This word spotting function seems to be effective even in a man-machine dialogue system, because human speech is not so rigid nor complete as it can be analyzed computationally by natural language grammar. The purpose of this paper is to clarify techniques to discriminate known words from unknown words in natural speech without using any language grammar. Here, four word spotting methods are theoretically and experimentally compared; forward-backward, heuristic Viterbi and their simplified methods.

2. Theoretical comparison

2.1 Probability ratio method

Known words have the constraints on a sequence of phonemes included in them. If x is utterance of a known word w, the probability $P(x|p^*_{seq})$, where p_{seq} is any phoneme sequence and p^*_{seq} is the best fit phoneme sequence to the utterance x, will be almost same as the probability $P(x|w)$, because the phoneme sequence of the known word w will be almost same as the best fit phoneme sequence p^*_{seq}. Then probability ratio $P(x|w)/\ P(x|p^*_{seq})$ almost equals unit.

On the other hand, if the utterance x is unknown word, any probabilities $P(x|w)$ will be far from that of the best fit phoneme sequence. Therefore the probability ratio $P(x|w)/P(x|p^*_{seq})$ will be less than some threshold θ for unknown words. We call this method *probability ratio method.*

2.2 Forward-backward probability method

Probability ratio can be expressed as follows under the assumption of $P(w) = P(p_{seq})$;

$$\begin{aligned} \frac{P(x|w)}{P(x|p^*_{seq})} &= \frac{P(x|w)P(w)}{P(x|p^*_{seq})P(p^*_{seq})} \\ &= \frac{P(x,w)}{P(x,p^*_{seq})} \end{aligned}$$

$$= \frac{P(x,w)}{\max_{p_{seq}} P(x,p_{seq})}$$

If we sum the phoneme sequence probabilities, instead of selecting the best fit phoneme sequence, the probability ratio can be regarded as a posteriori probability as shown in the following expression [1];

$$\begin{aligned} \frac{P(x,w)}{\max_{p_{seq}} P(x,p_{seq})} &= \frac{P(x,w)}{\sum_{p_{seq}} P(x,p_{seq})} \\ &= P(w|x) \end{aligned}$$

If the input utterance x is continuous speech, *a posteriori probability* of each $word_i$ is computed at every time t as follows;

$$\begin{aligned} \gamma_t & \ (word_i\_last\_state) \\ = & \ P(s_t = word_i\_last\_state|x) \\ = & \ \frac{\alpha_t(word_i\_last\_state)\beta_t(word_i\_last\_state)}{\sum_{alls} \alpha(s)\beta(s)} \end{aligned}$$

where $\alpha_t(s)$ and $\beta_t(s)$ are the forward and backward probabilities respectively. If a posteriori probability shows local peaks, known words are located at the peaks. We call this method forward-backward method. (Fig.1) shows the topology of the word and phoneme occurrence in a forward-backward method.

2.3 Viterbi method

This approach can be formalized using Viterbi decoding in 2.2, instead of forward-backward computation. Namely, the best path ending with $word_i$ at time t is searched by Viterbi decoding and its accumulated probability is computed. This probability corresponds to $\alpha_t(word_i\_last\_state)$ in expression(3). In general, the best path just before $word_i$ will be a phoneme sequence as shown in (Fig.2), because phoneme sequence has heigher probability than word sequence.

In the same way, after $word_i$, the best path will be a phoneme sequence. This accumulated probability of this best path corresponds to $\beta_t(word_i\_last\_state)$ in expression(3). Therefore, the topology shown in (Fig.3) is obtained from Fig.1.

The Viterbi decoding contributes to find the best ending time of $word_i$.[2] This approach basically spots one word by Viterbi decoding. To catch more than one word on utterance, N best words are searched by changing the $word_i$ in the topology. We call this method Viterbi method.

This method computes a likelihood $f(w, t_1, t_2)$ that a word w locates between time t_1 and t_2 in spoken sentence as follows

$$f(w, t_1, t_2) = g(w, t_1, t_2) + h(w, t_1, t_2)$$

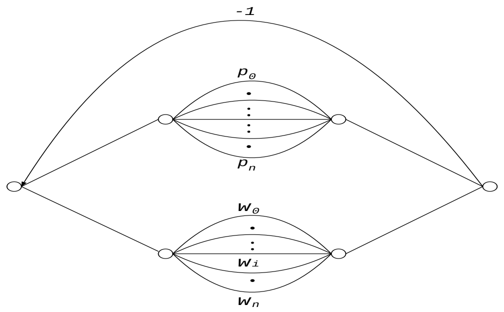

Fig.1: Topology for forward-backward method

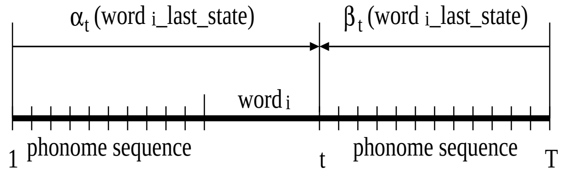

Fig 2: Interpretation of forward-backward probability in Viterbi method

where $g(w, t_1, t_2)$ is the word score and $h(w, t_1, t_2)$ is the language heuristic score for the time section 1 to t_1 and t_2 to T. The language heuristic score is further separated into two parts as follows:

$$h(w, t_1, t_2) = h_l(w, 1, t_1) + h_r(w, t_2, T)$$

The likelihood $f(w, t_1, t_2)$ can be computed at first by computing forward heuristics $h_l(w, 1, t_1)$ and backward heuristics $h_r(w, t_2, T)$. The word Viterbi score $g(w, t_1, t_2)$ is computed starting from time t_1 and the accumulated score from time 1 to t_2 is added to the backward heuristics.

2.4 Alpha/sum method

This method is the simplification of the forward-backward method shown by the following equation.

$$P_t(word_i\_last\_state) = \frac{\alpha_t(word_i\_last\_state)}{\sum_{alls} \alpha_t(s)}$$

The backward is not required and then it can spot the keywords time synchronously by finding local peaks.

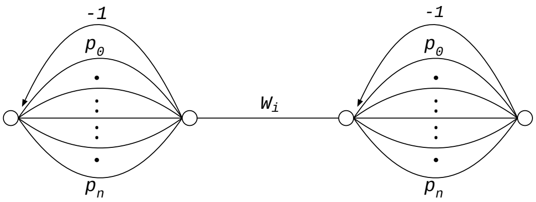

Fig. 3: Topology for Viterbi method

2.5 Normalized Viterbi method

This method is simplification of Viterbi method described in 2.3 and we call it as normalized Viterbi method. The backward heuristics is omitted and it is expressed as follows:

$$\begin{aligned} P(w,t_1,t_2) &= h_l(w,1,t_1)+g(w,t_1,t_2) \\ &\quad -maxP(1,t_2) \end{aligned}$$

This method is completely same as the probablistic ratio method.

3. Experimental comparison

3.1 Experimental condition

Word spotting experiments were carried out on Japanese continuous speech produced by Acoustic Society of Japan. The number of words to be spotted was 63 and their total occurrence was 161. The total duration time of the keywords was 0.0642 hour. The acoustic models were constructed by concatenating phoneme HMMs according to a dictionary. The number of phoneme HMMs was 40. They were trained using 416 sentences included in a,g,h,s,w,z sets of ASJ continuous speech database. The selected speaker was sha0001. The continuous speech, on which the words were to be spotted, was 77 sentences in o set, spoken by the same person. The phoneme recognition rate of the HMMs was 70.0% by the evaluation without phoneme label.

Table1 shows the experimental conditions for acoustic analysis (AA), vector quantization (VQ) and HMM.

The experiment was carried out by using phonemes, syllables and words for P_i of the syntax shown in Fig.1 and Fig.3. The number of words was all words included in 77 sentences to be evaluated, including keywords.

3.2 Experimental result

Table2 shows the spotting result by four techniques. In the table, A/W is the ratio of the number of correctly extracted keywords to 161 extracted words with high probability. An fa is the number of falsely extracted words (false alarm) per keyword per hour (fa/kw/h). From the table, it can be said that spotting with word shows the heighest extraction rate (A/W). Next is that with syllables and the worst is that with phonemes.

Table 1: Acoustic Analysis(AA),V Q and HMM

AA	Sampling frequency	12KHz
	High-pass filter	$1 - 0.97z^{-1}$
	Feature parameter	LPC cepstrum(16th)
	Frame length	20ms
	Frame shift	5ms
	Window type	Hamming window
VQ	Codebook size	256 codewords
	Distance measure	Euclidean distance
HMM	Number of states	5 states 3 loops
	Learning method	Concatenated training
	Type	Left to right discrete HMM

Table 2: Result of keyword spotting

	Phoneme		Syllable		Word	
	A/W	fa	A/W	fa	A/W	fa
Forward-backward	79.50	8.15	82.61	6.92	92.55	2.97
Heuristic Viterbi	79.50	8.15	81.99	7.17	93.79	2.47
Alpha/sum	78.26	8.65	80.75	7.66	90.06	3.95
Normalized Viterbi	80.12	7.91	80.75	7.66	(88.82)	(44.23)

Forward-backward method and Viterbi method showed almost same results. The Alpha/sum which is the simplification of the forward-backward method showed the good result.

4. Conclusion

In this paper, the word spotting methods are theoretically compared and four techniques are implemented and applied to continuous speech. We obtained the result that forward-backward method are equal to the Viterbi method and their simplifyed techniques showed good results.

References

[1] P.Jeanrenaud, M.Siu, K.Ng, R.Rohlicek, H.Gish,: " Phonetic-based Word Spotter " ,Proc. ESCA Eurospeech, vol.II, pp.1057-1060, 1993.

[2] T.Munetsugu, T.Kawahara, M.Araki, S.Doshita: " Keyword Spotting for Spontaneous Speech Understanding " , Technical Report of IEICE, SP92-116, pp.7-14, 1993. (Japanese)

Keywords: spotting, forward-backward, Viterbi, a posteriori probability, probability ratio