FACTA UNIVERSITATIS (NIŠ)
Series: Electronics and Energetics vol. 12, No. 3 (1999), 95-101

UDC 621.396

SERBIAN KEYWORD SPOTTING SYSTEM

Ljiljana Stanimirović and Milan D. Savić

Abstract. In this paper we present our recent work in implementing a keyword spotting system for detecting a limited number of keywords in continuous speech of Serbian. The keywords are detecting without modeling the non-keyword parts of the sentence using confidence measure and HMMs (Hidden Markov Models). Only keywords have to be model by HMMs in the way which we propose in this paper, that each syllable is three-state HMM. In this paper we also introduce MSQ - measure of system's quality in order to determine optimal step and optimal threshold for the confidence measure in the decoding phase. The obtained results show that proposed procedure can be used in interactive man-machine dialogue services.

1. Introduction

Despite the fact that speech recognition technology has advanced substantially in recent years in the world, its use is still not wide spread for some languages. The Serbian is one of them. There are a little, if any publications in Journals concerning word spotting systems for Serbian language. Successful applications of speech technology need a careful dialogue design. The dialogue means the system's ability to recognize one of the selected keywords in continuously spoken language and to produce some action, for example, to give some information.

The focus of our research, which we will explain in this paper, was to implement Serbian word spotting system based on statistical models (Hidden Markov models - HMMs), taking into account a fact that we have great experiences with Hidden Markov Models in implementing of Serbian isolated word recognition system [2],[3],[7] and the growing need for interactive speech

Manuscript received Nov. 11, 1998, revised Aug. 5 1999.

M.Sc.EE Lj. Stanimirović is with Mihajlo Pupin Institute, Volgina 15, Belgrade, Serbia, e-mail: ljiljana@kondor.imp.bg.ac.yu. Prof. dr M. Savić is with School of Electrical Engineering, University of Belgrade, 11120 Belgrade, Serbia, e-mail: esavic@ubbg.etf.bg.ac.yu .

technologies as well. Using confidence measure according to [1], we made some modifications of the proposed algorithm. Our goal was to show that even though we did not have big speech database on our disposal, we could realize word spotting system for Serbian language with good performances. In some countries and for some languages there are even Institutes, which the main or the only task is to record the speech material for researching needs.

In Charter 2 of this paper an overview of keyword spotting in continuous speech is given. The stress is put on using statistical methods i.e. Hidden Markov Models and confidence measure [4].

Due to inaccurate computations of the Gaussian distribution, because of the limitations in double floating format caused by the substantial dynamics of the speech signal, we suggested some modifications. Instead of the equation (1) we used equation (2), where k is a constant value, experimentally obtained during the research

$$p(\vec{x}) = \frac{1}{\sqrt{(2\pi)^N}\,|\underline{C}|}\exp(-\frac{1}{2}(\vec{x}-\vec{m})'\underline{C}^{-1}(\vec{x}-\vec{m})) \tag{1}$$

and

$$p(\vec{x}) = \frac{1}{\sqrt{(2\pi)^N}\,|\underline{C}|}\exp(-k\frac{1}{2}(\vec{x}-\vec{m})'\underline{C}^{-1}(\vec{x}-\vec{m})) \tag{1}$$

In (1) and (2) for the N-dimensional vector $\vec{x}$, $\vec{m}$ and $\underline{C}$ are its mean and covariance value respectively, as is shown in (3)

$$\begin{aligned} \vec{m} &= E\left\{\vec{x}\right\} = \frac{1}{N}\sum\nolimits_{k=1}^{N}\vec{x}_k \\ \underline{C} &= E\left\{(\vec{x}-\vec{m})(\vec{x}-\vec{m})'\right\} \\ |\underline{C}| &= \det\underline{C} \end{aligned} \tag{3}$$

Using (2) we reduce the dynamics of the speech signal but simultaneously it produced no effects on the recognition scores.

In Charter 3 we deal more with optimal step size and threshold determining for the confidence measure in the decoding phase. In Charter 4 the experimental results are given. Finally Charter 5 presents conclusions. We outline the future research that should be done.

2. Confidence Measure

It is very important to eliminate modeling of non-keyword speech outside the keyword boundaries. It can be achieved by modeling only keywords with HMM and by computing confidence measure on the whole pronounced sentence in the time interval corresponding to keyword boundaries. The keyword detection is achieved comparing the accumulated confidence measure in the mentioned interval with the determined threshold for each keyword. According to [1] confidence measure is computing as in (4) as negative logarithm of the keyword W a-posterior probability

$$C = -\log \Pr(W/O) \tag{4}$$

When we apply the Bayes' rule and pass over to the frame level, we compute local confidence measure as in (5). The probability of the feature vectors $\Pr(O_t)$ is calculated by taking all states of the HMM into account, as in (6)

$$c(O_t/s_j) = -\log \frac{\Pr(O_t/s_j)\Pr(s_j)}{\Pr(O_t)} \tag{5}$$

$$\Pr(O_t) = \sum_k \Pr(O_t/s_k)\Pr(s_k) \tag{6}$$

Each individual state of the keyword's HMMs now emits local confidence measure in conventional HMM based Viterbi search [2]. In the decoding phase the authors in [1] suggest computing of the integral confidence score ISc as in (7), where t_1 and t_2 are to be supposed keyword boundaries. But, they didn't say how they determine these boundaries. How we determine the optimal step, which corresponds to that time interval will be explained in the following charter

$$IS_c(O) = \sum_{t=t_1}^{t_2} c(O_t/s_j) \tag{7}$$

3. Optimal Step Size Determining

We recorded three speech databases for this research. Each one was recorded via standard microphone with sound blaster on the standard PC in the office environment. The sampling rate was 8 kHz. First database SDB (the sentence database) consists of 60 sentences with or without 4 keywords pronounced by 20 speakers. The keywords were *Beograd*, *Beopetrol*, *krstaši* and *pobednik*. The second database KWDB (the keyword database) consists of the isolated pronounced keywords pronounced by 20 speakers. The third database TSDB (test sentence database) consists of 100 sentences with or without keywords, different from that in SDB database pronounced by 20 speakers. That database has been used for testing purposes.

According to (5) we computed confidence measure for each sentence from the SDB for each time interval moving keyword's HMM through the sentence. Each HMM is obtained in the conventional training procedure [7]. We assumed keyword's model as concenation of the as many three-states HMMs as the keyword has syllables. Each syllable has been modeled by three-state HMM as Figure 1. shows.

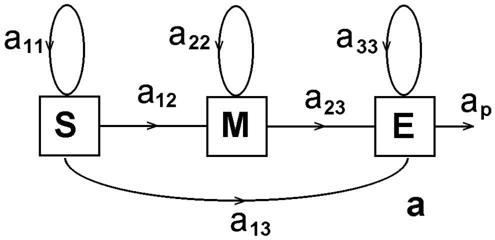

Fig. 1. HMM model for syllable.
S-start, M-midlle, E-end state

The front-end processing used 12 cepstral parameters computed along a MEL frequency scale in the telephone band. A 0.95 pre-emphasis factor was adopted with 8 kHz sampling frequency. MEL frequency grouping was carried out on FFT 256 samples [3]. We concern the overlapping Hamming windowed signal portions of 32 ms length with a frame period of 16 ms. Using only cepstral coefficients (not Δ cepstral or/and $\Delta\Delta$ cepstral and energy E, or some other parameters), our intention was to prove the word spotting algorithm with the parameter vector with as low dimension as is possible. In [6] has been shown that parameter vector with only cepstral coefficients can be used to obtain satisfied recognition results, although it's clear that the better results could be achieved with combination of Δ and $\Delta\Delta$ coefficients. In the first phase of our research, we wanted to reduce the computation efforts in order to achieve, as fast testing procedure of the word spotting algorithm as is possible.

According to (7) we computed integral confidence measure for each time interval in the following way. In the SDB database we determined possible keyword duration, i.e. step boundaries for each keyword. During that interval the keyword has been pronounced for different speakers. For each possible step, we computed integral score according to (7) assuming the step as time interval from t_1 to t_2. For example, for the keyword Beograd, the possible keyword duration, i.e. step in the database SDB is from 30 to 50. The minimum value of the integral confidence measure for each sentence in the SDB for each step is determined in order to find the optimal step and threshold. While we have known which sentences had keywords and which had not, we could investigate how to improve *measure-of-system's quality* - MSQ, as in (8) considering different steps and thresholds. We introduced MSQ in our research as criteria how good is our system

$$MSQ = \frac{n\_g\_d\_kw}{n\_kw} \times \frac{n\_g\_d\_kw}{n\_nkw} \qquad (8)$$

where are:

- $n\_g\_d\_kw$ is the number of correctly detected keywords in the database,
- n_kw is the total number of keywords,
- $n\_g\_d\_nkw$ is the number of sentences in which the system didn't detect keywords (and they didn't have keywords),
- n_nkw is the number of the sentences in the database without keywords.

Our goal was to maximize MSQ in the way that system has to recognize maximum number of the keywords in sentences which include them and at the same time system does not have to recognize the keywords in as many sentences without keywords as is possible. We examined the minimum value of the integral score for the sentences in the SDB with keywords and we used that value to determine the threshold. For each possible step (from 30 to 50 for keyword Beograd), we computed threshold as the minimum value of all minimum values those sentences.

4. Experimental Results

For the test purposes we used TSDB database. The obtained recognition results are given in the Table 1 [5]. It can be seen that the system recognizes each keyword very well, i.e. in each of ten sentences with keywords, ten keywords were recognized for each keyword. System made some errors in recognizing the keywords in the sentences without keywords (for example, for the keyword Beograd, system false recognizes 3 from 90 sentences).

It is worth to mention that disputes the fact that those three keywords: Beograd, Beopetrol and pobednik are confusable (they sound similarly), the system shows good recognition results. It is well known that the choice of suitable keywords is a critical parameter for the good performances of the recognition system. Because of that fact our results are of greater importance.

Table 1. Word spotting recognition results

keyword	$\frac{n\_g\_d\_kw}{n\_kw}$	$\frac{n\_g\_d\_nkw}{n\_nkw}$	MSQ
Beograd	10/10	87/90	96 %
Beopetrol	10/10	90/90	100 %
pobednik	10/10	81/90	90 %
krstaši	10/10	84/90	93.33 %

5. Conclusion

Our goal was to show that we obtained good results in Serbian word spotting system, although confusable keywords have been chosen and we did not have big database on disposal for model's training. It means that our keyword's models could be better with the larger database. Also the recognition results could be better if we include Δ and $\Delta\Delta$ cepstral coefficients in the parameter vector.

We introduced some modifications of the formula for Gaussian distribution, because of the limitations in double floating format for the equation (1), caused by substantial dynamics of the speech signal. Instead of equation (1), we used equation (2) where k is experimentally obtained value.

Our HMM keyword's models are obtained by modeling each syllable with one three-state HMM. The next step in our research would be to replace each phoneme in context (i.e. triphone) with one three-state HMM. Also, it would be interesting to show how this system works when larger number of keywords is concerned.

6. Acknowledgment

This study has been supported by the research grant S.1.04.10.0137 from the Ministry for Science and Technology of Serbia, Belgrade.

REFERENCES

1. J. Junkawitsch, G. Ruske, H. Hoege: *Efficient methods for detecting keywords in continuous speech.* Proceedings of the IEEE ICASSP'96, Vol. II, Munich, Germany, 1996.
2. L. Rabiner, B-H. Juang: *Fundamentals of speech recognition.* Prentice Hall, 1993.
3. Lj. Stanimirović, Z. Ćirović, M. Savić: *Isolated Serbian word recognition system.* Proceedings of the International Conference of Signal Processing and Communication - ICSPC'98, Las Palmas, Spain, 1998.

4. Lj. Stanimirović, N. Stanković: *Word spotting in continuously spoken Serbian.* (In Serbian). Proceedings of the ETRAN'98, vol. II, pp. 399-401, Vrnjačka Banja, 1998.

5. Lj. Stanimirović, Z. Ćirović: *Keyword spotting system for Serbian language.* Proceedings of the ICT' 99, Korea, 1999.

6. Lj. Stanimirović: *Optimal speech parameter vector in speech recognition systems based on HMMs.* (in Serbian). Journal TEHNIKA, num. 5, 1998.

7. Z. Ćirović, Lj. Stanimirović: *Man-Machine Communication: An Isolated Word Recognition System Based On Hidden Markov Models.* Proceedings of the DMMS'97, pp. 111-117, Budapest, Hungary, 1997.