In this paper we present our recent work in implementing Serbian spoken dialogue system for the bus information retrieval at the main Belgrade bus station. Dialogue is organized into several levels. At each level, system has to recognize a limited number of keywords in continuous speech of Serbian. The keywords were modeled by HMMs (Hidden Markov Models) in such a way that each syllable is three-state HMM. In order to obtain optimal thresholds for integral confidence measure in decoding procedure, we introduced the MSQ – measure of system's quality. The obtained results show that proposed procedure can be used in interactive man-machine dialogue services with the success.

1. INTRODUCTION

Successful application of speech technology need a careful dialogue design. Under the dialogue we consider 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 to the user. The focus of our research, which we will explain in this paper is to realize the Serbian dialogue system. It is the bus information retrieval system at the main Belgrade bus station, which should provide information about bus routes, departure dates, durations, costs and so on. It is wordspotting system based on statistical models (Hidden Markov models - HMMs).

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 [1], [2]. In Charter 3 we deal more with optimal step size and threshold determining for the confidence measure in the decoding phase. Charter 4 gives an overview of the system and more details about dialogue manager. Finally in Charter 5 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 negative logarithm of the keyword W a-posterior probability. When we apply the Bayes' rule and pass over to the frame level, we compute local confidence measure as in (1). The probability of the feature vectors )\( \log / (t)_{OP} \)) is calculated by taking all states of the HMM into account, as in (2).

\[
c(0_t / s_j) = -\log \frac{P(0_t / s_j)P(s_j)}{P(O_t)}
\]

\[
P(O_t) = \sum_k P(0_t / s_k)P(s_k)
\]

Each individual state \( s_k \) of the keyword’s HMMs now emits local confidence measure \( c(0_t / s_j) \) in
conventional HMM based Viterbi search. In the
decoding phase the authors in [1] suggest computing
of the integral confidence score $I_{Sc}$ as in (3), where $t_1$
and $t_2$ are to be supposed keyword boundaries. How
we determine the optimal step, which corresponds to
that time interval will be explained in the following
charter.

$$I_{Sc} = \sum_{t=t_1}^{t_2} c(O_t/s_j)$$  \hspace{1cm} (3)

Using confidence measure according to [1], we
made some modifications of the proposed algorithm,
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. Instead of the equation (4) we used
equation (5), where $k$ is a constant value,
experimentally obtained during the research.

$$p(\mathbf{Y}) = \frac{1}{\sqrt{(2\pi)^N|C|}} e^{-\frac{1}{2}(\mathbf{Y}-\mathbf{h})^T C^{-1}(\mathbf{Y}-\mathbf{h})}$$  \hspace{1cm} (4)

$$p(\mathbf{Y}) = \frac{1}{\sqrt{(2\pi)^N|C|}} e^{-k\frac{1}{2}(\mathbf{Y}-\mathbf{h})^T C^{-1}(\mathbf{Y}-\mathbf{h})}$$  \hspace{1cm} (5)

For $N$-dimensional vector $\mathbf{Y}$, $\mathbf{h}$ and $C$ are its mean
and covariance value respectively. Using equation (5),
we reduce the dynamics of the speech signal but
simultaneously it produced no effects on the
recognition scores.

### 3. OPTIMAL STEP SIZE DETERMINING

Three different speech databases were recorded via
standard telephone line and sound blaster card on the
standard PC. The sampling rate was 8 kHz. First
database SDB (sentence database) consists of 170
sentences with or without 60 keywords pronounced by
100 speakers. The keywords were names of the towns
in Yugoslavia and words: da (yes) and ne (no). The
second database KWDB (keyword database) consists of
the isolated pronounced keywords pronounced by
100 speakers. The third database TSDB (test sentence
database) consists of 500 sentences with or without
keywords, different from that in SDB database
pronounced by 100 speakers. That database has been
used for testing purposes.

According to (1) we computed confidence measure
for each sentence from the SDB for each time interval
shifting keyword’s HMM along the sentences. Each
HMM is obtained in the conventional training
procedure [5]. We assumed keyword’s model as
concatenation of the as many three-states HMMs as the
keyword has syllables.

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], [4]. We concern the overlapping
Hamming windowed signal portions of 32 ms length
with a frame period of 16 ms. Using only cepstral
coefficients (not $\Delta$ cepstral or/and $\Delta\Delta$ cepstral and
energy or some other parameters), our intention was
to obtain as better results as is possible with the
parameter vectors with as low dimension as is
possible.

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 (6) considering different steps and thresholds.
We introduced MSQ in our research as criteria how
good is our system.

$$MSQ = \frac{n_{g\_d\_kw}}{n_{kw}} \cdot \frac{(n_{\_nkw} - n_{g\_d\_nkw})}{n_{\_nkw}}$$  \hspace{1cm} (6)

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 falsely
detected keywords in the sentences which
did not 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.

4. DIALOGUE MANAGER

Our dialog manager is realized in MS Visual C++. The block diagram of the whole system is shown on the Figure 1. The dialogue manager has to coordinate the communication between wordspotting system and sound blaster. At the moment, the messages are generating as recorded messages, but later we will include our TTS system for Serbian language.

Figure 1. System block diagram

Dialogue manager works with dialogue nodes. One dialogue node consists of six fields, as Figure 2 shows. Dialogues level is one system’s question and one user’s answer. Each level has his own node and appropriate number of keywords, which have to be recognized on that dialogue’s level. In the field number 3 of the dialogues node there are the pointers to keywords models. In the field number 4 are confidence measure thresholds for keywords.

Each dialogues node has its own systems messages, which have to be in field number 5, on the Figure 2. System has to know what’s going on at every moment in the system. Field number 6 is for users message. It’s some kind of information which is sent via sound blaster and telephone line to the user.

5. FUTURE ACTIVITIES

We have to test our system in real situations. The quality of the dialogue system could be proofed answering on the following questions:

- Does the user retrieve the desired information in the sufficiently short time?
- Is the dialogue system conformable for the user?

The answers on these questions will be the proof that our dialogue system is with good performances and will be the topic of one of our next papers.

References


