WOULD YOU LIKE TO MAKE YOUR PROGRAMS UNDERSTAND HUMAN VOICE?

Ing. Petr SCHWARZ, Doctoral Degree Programme (2)
Dept. of Computer Graphics, FIT, BUT
E-mail: schwarzp@fit.vutbr.cz

Supervised by: Dr. Ing. Jan Černocký

ABSTRACT

This paper presents a system for voice control ready to use in applications. The system is based on a keyword spotting algorithm which does not need previous speech segmentation. A general structure, application interface and this system usage are described. Some main properties are reported.

1 INTRODUCTION

During last year, we spent much time in keyword spotting algorithm investigation, in comparison of many existing approaches for keyword spotting and concurrently in improving our Hidden Markov Model (HMM) based Czech speech recognizer. Some results are published in [1][2]. In this project we wanted to use our knowledge and experience which is coming from research actually being performed at Faculty of Information Technology, BUT by Speech Processing Group [3] and offer some results to general public. The output is a dynamic linkable library (DLL) currently available for Windows operation systems, which encapsulates our keyword spotting system, routines for recording audio in real time, algorithms for speech parameterization and routines for manipulation with phonetic transcription and dictionary. Everything together composes a system for control by voice commands which can be very easy and immediately used in your application. The DLL can be downloaded from our web page [4].

2 SYSTEM DESCRIPTION

A block diagram is shown on figure 1. Speech is divided to frames 25 ms long, preemphasized, windowed, MFCC (Mel Frequency Cepstral Coefficients) are calculated and delta and acceleration coefficients are added. A vector at the output from parameterization has 39 dimensions (13 MFCC including C_0, Δ and $\Delta\Delta$). A confidence of each word from a dictionary is estimated every 10 *ms* (fig. 2). The HMM approach [5] is used. In our case

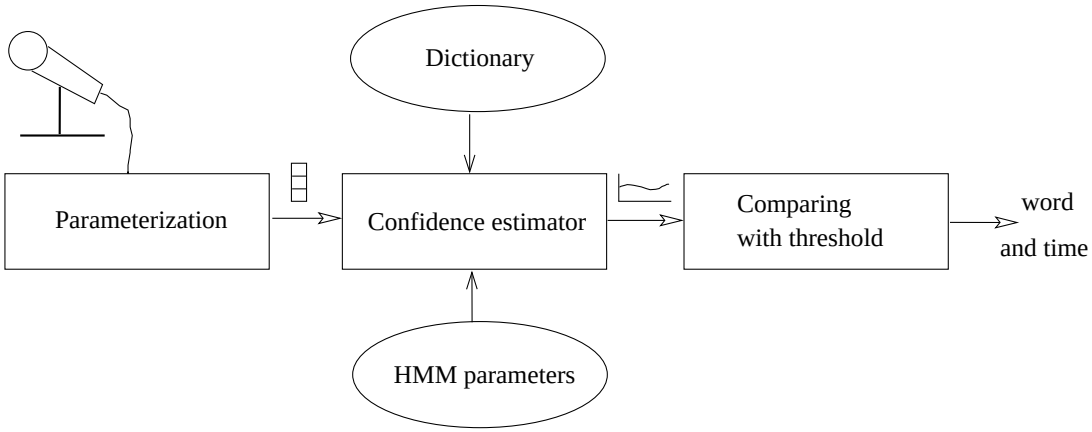

Figure 1: Block diagram of recognizer

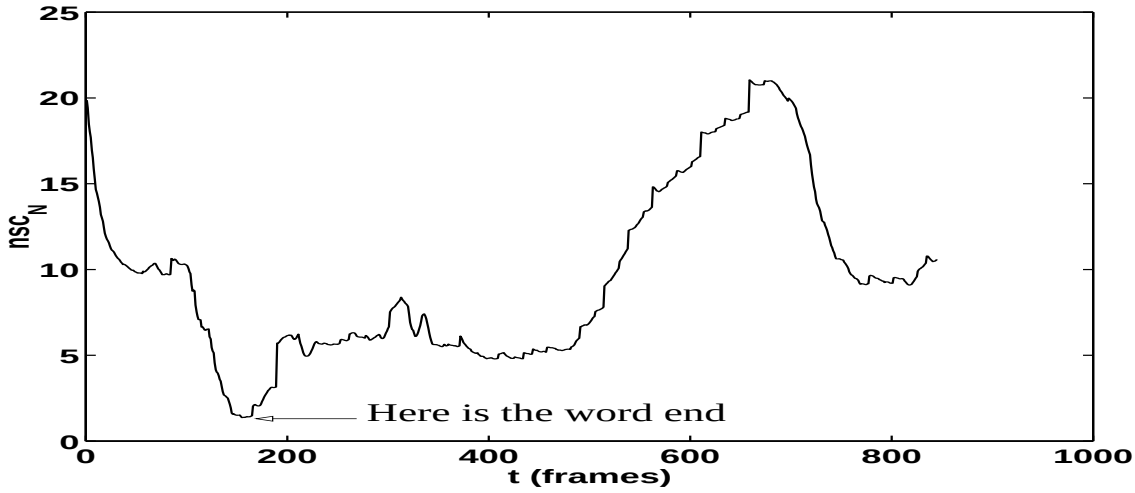

Figure 2: Confidence function of Czech word ’pokračovat’

we have triphone HMM with three states and 16 Gaussian functions trained on the Czech SpeechDat-E database [2][6]. A Modified Viterbi algorithm was chosen for decoding [7]. This algorithm has at least one big advantage. It does not need any pre-segmentation, therefore no voice activity detector is necessary and there are no additional errors caused by it. The confidence function is evaluated for each word separately. It is compared with a threshold and if the function is above a threshold, the word is accepted, else rejected. Accuracy of such system was evaluated in [1].

3 APPLICATION PROGRAMMING INTERFACE

An application programming interface (API) was set up for communication between client (user of the DLL) and the system (DLL). It was created as simple as possible not to confuse the user. It consists of four control functions and one callback function. Two control function are used for starting and stopping sound recording and consequently start-

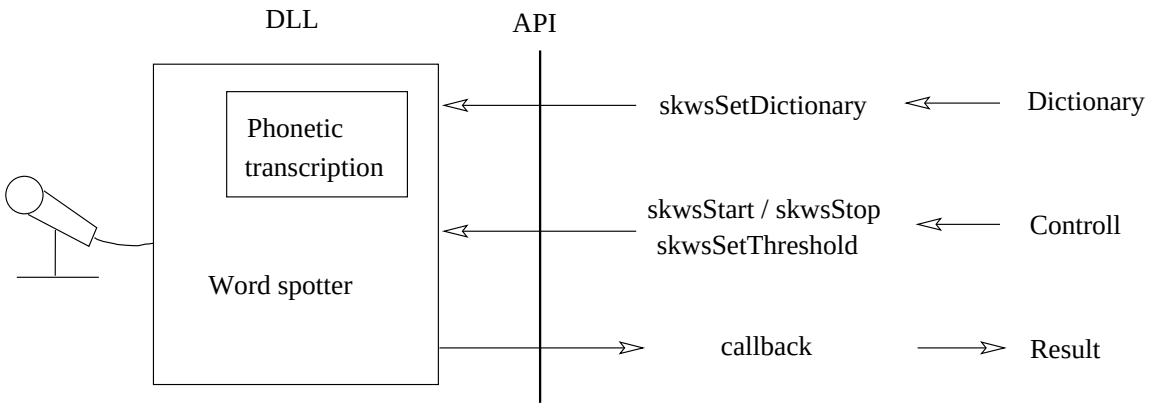

Figure 3: Application programing interface

ing and stopping the whole system. Next one is for manipulation with a dictionary. The dictionary is a list of words that are detected. A simple list can be entered or eventually a phonetic transcription for each word can be added. The phonetic transcription is mainly for words that do not have regular pronunciation. The last control function sets word acceptation/rejection threshold. There is a default one but the optimal value depends on surrounding conditions too and if the conditions – mainly noises – are entirely different, the user is allowed to adjust it. The callback function is used for handling of detection results.

When the client starts using this system, a dictionary has to be set. The system calculates confidences of each word as time goes. The confidence value is compared with the threshold and if a word is accepted, the callback function is raised. It depends on the user only how the information is processed then.

LIST OF FUNCTIONS

void WINAPI skwsStart(SKWSCALLBACK func, LPARAM lParam)

Starts recording and sets a callback function where the detected words will being sent. The *lParam* parameter is passed to the callback function as one of its arguments.

void WINAPI skwsStop()

Stops recognition.

void WINAPI skwsSetThreshold(float threshold, float delta)

Sets a threshold for word acceptation / rejection.

void WINAPI skwsSetDictionary(LPCSTR words)

Sets a dictionary. Words are separated by a new line. Each word can alternatively have its phonetic transcription in a Czech Sampa phonetic alphabet [6]. This transcription has to be closed in composed brackets and follows the word.

void CALLBACK OnWord(char *word, float score, LPARAM lParam)

This is definition of a callback function. User has to have implemented one in his code. All detected words are sent there.

4 EXAMPLE

```
#include <stdio.h>
#include “skws.h”

void CALLBACK OnWord(const char *word, float score,
                     LPARAM lParam)
{
    printf(“%s\n”, word);
}

int main()
{
    skwsSetDictionary(“nula\njedna\ndva”);
    skwsSetThreshold(3.0f, 0.0f);
    skwsStart(&OnWord, 0);
    getchar();
    skwsStop();
    return 0;
}
```

5 CONCLUSION

A dynamic likable library with system for voice control was developed. This library was tested under Windows 2000 and Windows XP and is currently available for download at our web page. During tests two development environment – Microsoft Visual C++ and C++ Builder – were used. For each environment a sample program was written. It can now help everyone in easier and faster using of this library. Detection accuracy for this task (through microphone) was evaluated subjectively only. It depends mainly on quality of microphone and on degree of background noise. Generally using of a close-talk microphone is much better. It attenuates both background noise and other speakers’ voice. A response of a system on an other speaker’s voice is usually negative too. If a table-top microphone is used, reaction on the other’s speaker voice can be particularly eliminated by word acceptation/rejection threshold lowering. Quite high influence of detection accuracy on word length was seen. If the word is long (7 phonemes), the accuracy is very good and if the word is short (3 phonemes), the accuracy is poor. It is much better to use longer word if we have two synonyms. One negative property was found if we have both word and its sub-word in our dictionary. Each is meaningful word and therefore both are detected.

6 FUTURE WORK

We would like to perform more experiments and improve this library to be more user-friendly. We will work on increasing of detection accuracy, on increasing robustness against noise and on removing some negative aspects as detection of sub-words. We hope that experiments on such on-line real-condition keyword spotting system enable us to better understand speech recognition and problems that are still there, show as something more we have never seen from experiments on speech databases and helps as in research. Inventions and good ideas gained from the research should be implemented in this or similar system again and should show both us and others that they are good and ready for practical applications. In current time, there is a student's year project running with the goal to implement this library to control of Microsoft Windows and to prove usability of this library in such difficult task.

ACKNOWLEDGEMENTS

This research has been partially supported by Grant Agency of Czech Republic under project No. 102/02/0124 and by EC project Multi-modal meeting manager (M4), No. IST-2001-34485. Jan Černocký is supported by post-doctoral grant of Grant Agency of Czech Republic No. GA102/02/D108.

REFERENCES

[1] P. Schwarz and J. Černocký, *Modifications of Viterbi algorithms for keyword detection*, in Proc. EEICT 2002, Brno, Czech Republic

[2] M. Karafiát and J. Černocký, *Context dependent Hidden Markov models in recognition of Czech*, in Proc. Radioelektronika 2002, Bratislava, Slovak Republic

[3] *Speech Processing Group at Faculty of information technology, BUT*, url: http://www.fit.vutbr.cz/research/groups/speech/

[4] *On-line keyword spotting system with a microphone input – DLL*, url: http://www.fit.vutbr.cz/research/groups/speech/sw/skws_dll.html

[5] S. Young, J. Jansen, J. Odell, D. Ollason and P. Woodland, *The HTK book*, Entropics Cambridge Research Lab., 1996

[6] *SpeechDat-E homepage*, url: http://www.fee.vutbr.cz/SPEECHDAT-E/

[7] J. Junkawitsch, L. Neubauer, H. Höge, G. Ruske, *A new keyword spotting algorithm with pre-calculated optimal thresholds*, Proc. CSLP, 1996