Processing of Acoustic Signal for Speech Recognition

Radoslav Bučko, Technical University of Košice

Abstract
This paper describes recognition of spoken speech and problems with recognition, basic parts of recognizer and especially with signal processing.

Voice and speech recognition could be a new way of car’s equipment control such a GPS, radio etc.

1. Introduction
Communication using verbal speech is the most basic, most natural and most important form of information transfer between people. If computer or other device has to be commanded by voice, acoustic signal from speaker, voice recognition and understanding the information has to be technically and algorithmically solved.

Voice recognition software consist of 2 main parts. First part is the signal processing, which results in sequence of observations (mostly vectors). Second part are the classifiers, which assign the most suitable word from the dictionary to each sequence.

2. Embedded system
An embedded system is a computer system designed to perform one or more dedicated functions often with real-time computing constraints. It is embedded as part of a complete device often including hardware and mechanical parts. Embedded systems are controlled by one or more main processing cores that are typically either microcontrollers or digital signal processors (DSP). Since the embedded system is dedicated to specific tasks, design engineers can optimize it to reduce the size and cost of the product and increase the reliability and performance. Some embedded systems are mass-produced, benefiting from economies of scale. The program instructions written for embedded systems are referred to as firmware, and are stored in read-only memory or Flash memory chips. They run with limited computer hardware resources: little memory and operating output. Because of this, it is important to optimize the acoustic signal processing used by embedded systems for speech recognition. [1]

3. Processing of acoustic signal
Choice of representation which will choose for speech processing will be affected by its final purpose. The criterias which we must consider are:
- complexity,
- speed of information transfer,
- flexibility.

a) Complexity – is the difficulty of processing which means the amount of mathematical operations with which we get the chosen representation. This is often considered by the cost of technical equipment.
b) Speed of information transfer is connected with excessivity of chosen representation. Low transfer speed means that the encoded representation can be transferred by low capacity channel or effectively remembered. Lowering the speed of transfer under certain threshold can lead to significant reduction of transferred content which can lead to worse classification abilities.
c) Flexibility is the relative simplicity with which the acoustic representation can be further processed. [2]

The basic principle of most methods for acoustic signal processing is the assumption that its properties are changing slowly. This assumption leads to application of methods called short-term analysis, where the segments of speech signal are separated and processed like separate short sounds. These segment are micro segments which are represented by the time segment of 10 to 30 ms. Because these micro segments are connected or can overlap each other we will get the sequence of numbers which describes the speech. Methods of short-term analysis require the input information gained by digitalization, which means by coding of the wave shape.

3.1 Coding the shape of the wave
Speech signal is recorded mostly by microphone, so the analog signal is recorded. Analog cycles are digitalized, that the continuous signal is represented by sequence of numbers. This process is called pulse
code modulation. Pulse code modulation consists of two operations:
- sampling in time,
- quantization.

**Sampling in time** – Samples are taken from continuous signal in periodic moments \( t_n = n \cdot T \) which size corresponds to immediate values of continuous signal in sampling time \( t_n \). \( T \) is the sampling period and \( n=0,1,\ldots, \infty \). [3]

According to Shannon’s sampling theorem the frequency of sampling \( f_s \) must be twice as the maximum frequency of analog signal \( f_m \):

\[
f_s \geq 2f_m
\]  

(1)

**Quantization** is the operation which allows the change of signal with continuous variable to signal with finite number of values. Principle of quantization is shown on figure 2.

3.2 Processing by time

Most methods of short term analysis in time can be described by the following equation:

\[
Q_n = \sum_{k=-\infty}^{\infty} r(s(k)) w(n-k),
\]  

(2)

where \( Q_n \) is the short time characteristic, \( s(k) \) is the sample of acoustical signal get by pulse code modulation in time \( k \), \( r(s(k)) \) is the transformation function a \( w(n) \) is the weight sequence (or window) which chose the samples \( s(k) \).

Hamming’s windows are used when processing in time. Rectangular window means that the same weight is applied to every sample of micro segment.

Rectangular window is defined as(fig.3):
- \( w(n)=1 \) for \( 0 \leq n \leq N-1 \),
- \( w(n)=0 \) for other \( n \).

Hamming’s window is defined as(fig.4):
- \( w(n)=0,54-0,46\cos(2\pi n / (N-1)) \) for \( 0 \leq n \leq N-1 \),
- \( w(n)=0 \) for other \( n \).

3.3 Processing by frequency

Spoken speech can be in area of frequency represented like composition of spectral envelope, which characterizes the properties of speech apparatus and fine structures which characterizes the aroise. The most used methods are based upon short term Fourier analysis and strip filtration.

3.4 Mel-scale frequency cepstral coefficient

Homomorphic analysis belongs to the group of methods of nonlinear signal processing which are based on using the generalized principle of superposition. These methods are suitable for analysis of signals which were created by convolution or multiplication of arousing function (periodic sequence of pulses or random noise generator) and pulse response of speech apparatus, so this method is suitable for speech processing. By experiments, Biswas determined, that from methods of signal processing using the homomorphic analysis the most effective are mel-scale frequency cepstral coefficients [4]

Mel-scale frequency cepstral coefficients are trying to compensate the nonlinear perception of frequency using the bank of triangular strip filters with linear spread of frequencies in so called mel-scale frequency. This scale is defined. This scale is defined by equation:

\[
f_m = 2595\log_{10}\left(1 + \frac{f}{700}\right),
\]  

(11)

where \( f [\text{Hz}] \) is the frequency in linear scale and \( f_m [\text{mel}] \) is corresponding frequency in nonlinear melov’s scale. Algorithm of this filtration is realized using the bank triangular strip filters along the frequency axis with scale in mel-scale (fig.7). It is wise to choose the number of the strip according to number and placement of critical strips and
respecting the sampling frequency $f_s$/[Hz] and total width of transferred strip $B_w$/[Hz] or $B_{mw}$/[mel].

For example, when sampling frequency is 16 kHz, then we will have only 20 coefficients, when the frequency is 8 kHz, we will have only 15 coefficients.

**Conclusion**

First and very important step when recognizing speech is the signal processing. It creates output for classifiers. In order of fastest classifying (real-time if possible) these information must be reduced to lowest possible rate with insignificant loss of information content. This is very important especially for embedded systems in cars which have less memory and operating output than PC. The most effective are mel-scale cepstral coefficients.

**Bibliography**


**Author:**

Ing. Radoslav Bučko
KTEEM FEI TU Kosice
Park Komenského 3, 04200 Kosice, Slovakia
tel. +421 55 602 2706
e-mail: radoslav.bucko@tuke.sk

**ACKNOWLEDGMENT**

We support research activities in Slovakia / Project is co-financed from EU funds. This paper was developed within the Project "Centrum excelentnosti integrovaného výskumu a využití progresívnych materiálov a technológií v oblasti automobilovej elektroniky", ITMS 26220120055