Michal GREGOR*, Tomáš MICHULEK**

INTELLIGENT MANUFACTURING SYSTEMS – AUTOMATIC SPEECH RECOGNITION SYSTEM

Abstract

The University of Žilina and the Central European Institute of Competitiveness have been conducting research in development of Intelligent Manufacturing System. One of areas researched was Automatic Speech Recognition (ASR) and control of manufacturing processes through voice control. This paper presents chosen results from research done at the University of Zilina and at the Central European Institute of Technology.

1. INTRODUCTION

Global undertaking brings numerous troubles and difficulties to all manufacturers. Turbulent global markets, strong competition, growing costs, changing undertaking paradigms are only a few of what presses manufacturers to look for new approaches on how to effectively manage all manufacturing processes.

The future cannot exist without innovation of production processes and production systems as it cannot exist without the innovation of products. Production systems require redesign as well, new machines and devices, transport systems, control systems, work organisation etc. Such changes are introduced by teams of specialists, designers and planners.

To design whole factories is an extremely complex and difficult problem. Quality of the project determines the future long-term effectiveness of the factory. Digital models of factories (FMU – Factory Mock Up) make it possible to greatly enhance the communication among the design teams, to lower the risks evoked by making wrong decisions and to speed up innovation and increase the efficiency of the innovation process by improving the performance.

Intelligent Manufacturing Systems (IMS) represent the future of manufacturing. The main target of IMS is to increase manufacturing competitiveness and productivity.

Currently conducted Research and Development on this area is focused mainly on new directions in development of intelligent machine tools, intelligent tools and jigs, intelligent transportation system, intelligent control system, etc.

---

* University of Žilina, Institute of Competitiveness and Innovation, Univerzitná 1, 010 26 Žilina, Slovak Republic, tel.: +421 0948 044 505, e-mail: o.m.gregor@gmail.com.

** Central European Institute of Technology, Univerzitná 8413/6, 010 08 Žilina, Slovak Republic, tel.: +421 904 010 375, e-mail: tomas.michulek@ceit.eu.sk.

2. ASR SYSTEMS – CLASSIFICATION AND BASIC TERMS

Automatic Recognition Systems are designed to transform acoustic signal containing speech and certain maximum amount of noise into recognized units of speech (fig. 1). Recognized units are expressed in a computer-readable form so that they can undergo further processing (like syntactic and semantic analysis and similar), be used in control, displayed, etc. This paper specifically focuses on ASR systems used in remote control.

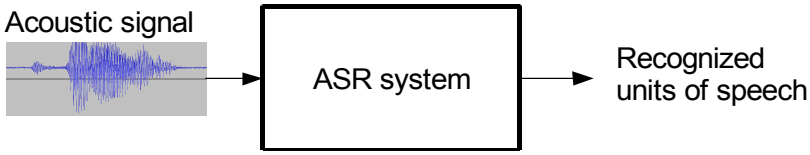

Fig. 1. General ASR system block diagram

Classification of ASR systems

ASR systems may be categorized from several perspectives [1]. The following section lists the most fundamental of these.

Mode of Operation

Considering the mode of operation, ASR systems may be divided into following groups:

- systems recognizing isolated units of speech,
- systems recognizing continuous speech (usually real-time).

Isolated Mode

Systems in the isolated mode classify recordings containing a single unit of speech (word, sentence, command, …), whereas systems in the continuous mode classify recordings containing multiple units of speech (a sequence of units).

The isolated mode is mainly used in remote control, using a small pre-defined vocabulary of commands. The whole speech signal is recorded first and then processed, which allows for application of certain concepts (like normalization), that are not feasible in real-time continuous systems. To start and stop the recording of a unit an external signal is thus required. Such signal may be provided manually (like pressing and releasing a button), or automatically (like volume-based trigger).

ASR systems in isolated mode using a small vocabulary tend to be more accurate for remote control problems, the external recording triggering requirement is a considerable disadvantage however, as well as its inability to provide for syntactically flexible commands.

Continuous Mode

Systems recognizing continuous speech recognize sequence of units instead of individual units as opposed to speech recognition in isolated mode. Such systems usually work in real time – complete signal is not yet available when the recognition process begins (a form of streamline interface is introduced).

Continuous mode systems may implement a technique called word spotting, where the system listens to all units, but looks only for defined units (the entire signal is analysed – no

triggering needed; this also has a disadvantage – full-time recognition is usually more CPU-intensive than a relatively simple triggering mechanism).

Speaker Dependence

ASR systems may also be categorized as:

- speaker dependent,
- speaker independent.

Speaker dependent systems achieve optimal performance for a single speaker, or for a small group of speaker, for whom it is pre-trained. To adapt such system to other voice requires collection of a considerable amount of data from the particular speaker. Speaker dependent systems however usually maintain higher accuracy (for the particular speaker) than comparably complex speaker independent systems. These systems are as well suitable for implementation of some algorithms, such as energy-based cropping, which may contribute to considerable improvement of accuracy, and are not well applicable to real-time continuous systems.

Speaker independent systems achieve reasonable performance for new, unknown speakers (separate mechanisms are however usually required to account for mispronunciation and similar). This degree of independence is typically achieved by implementation of special adaptation algorithms, through collection of data from a large group of different individuals, or by mixing both techniques.

Speech Mode

ASR systems may be able to recognize:

- read speech (or other kind of speech prepared in advance),
- spontaneous speech.

Contemporary ASR systems typically require the speaker to adapt their speech – to speak fluently, make pauses of certain length etc. – these categorize as systems recognizing read speech, or other kind of speech prepared in advance.

The ability to recognize spontaneous speech as well is desirable, it is however a much more complex task, the rate of speech being dynamic and the melody of voice varying. The system also has to account for interjections such as hmmm, ehm, coughs and similar.

3. ANN-BASED AUTOMATIC SPEECH RECOGNITION

Artificial Neural Networks present several advantages over the more traditional method using HMM as the acoustic model. Most notably – many of the notorious assumptions of HMMs[1] don't apply, which makes ANNs the more fitting alternative. That the nature of ANNs allows for a straight-forward parallel implementation, thus helping to speed up the training phase considerably, is certainly a welcome advantage as well.

---

1 The false assumptions include the independence assumption – that is, the premise, that there is no correlation between the adjacent frames of input data; the quantization error for discrete HMMs and model mismatch for continuous and semi-continuous density models; and other. For further information see [2].

3.1. The Properties of ANNs used in ASR

ANN-based recognizer typically makes use of a layered or recurrent network architecture (fig. 2 and fig. 3) and may use an arbitrary activation function.

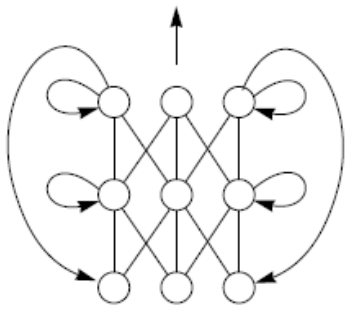

Fig. 2 Recurrent ANN [2]

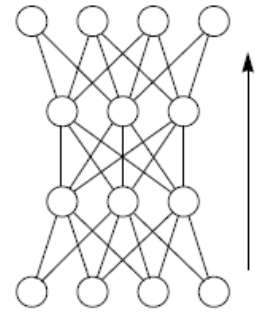

Fig. 3 Layered ANN [2]

The standard form of activation function (1) in combination with one of the well-known functions like sigmoid seems to be the most popular choice:

$$O = \Psi(u) \text{ where } u = \sum_{i=0}^{n} w_i x_i \ . \tag{1}$$

O stands for output of the Artificial Neuron, whereas $X = \{x_1, x_2, ..., x_n\}$ is a vector of inputs for that neuron and $W = \{w_1, w_2, ..., w_n\}$ are their corresponding weights. However, this is not the only possibility and some modified forms exist – Learning Vector Quantization for one, uses the following formula [2]:

$$O = \Psi(u) \text{ where } u = \sum_{i=0}^{n} (w_i - x_i)^2 \ . \tag{2}$$

The training method obviously depends on the chosen topology – layered networks are usually trained by some modification of back-propagation, while recurrent networks may use one of the methods listed in [3] or [4].

3.2. ANN-based ASR – the principle

The first important fact to note is that ANNs are as such only an option in isolated recognition, that is in recognition of isolated units of speech (words, sentences, phonemes, etc.) carried out once the audio signal of a complete unit has been captured[2].

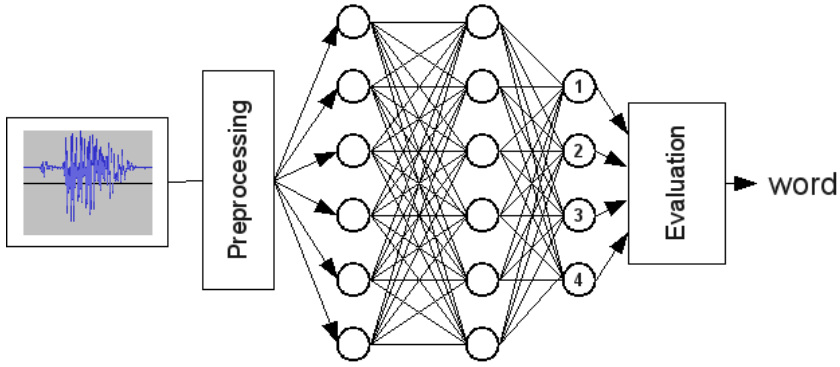

Fig. 4 The principle of an ANN-based recognizer

Figure 4 shows the principle of ANN-based speech recognition. The process can be subdivided into 3 major phases:

1) signal capturing and preprocessing,
2) pattern recognition,
3) evaluation of the results.

3.3. Preprocessing

The preprocessing phase does not differ much from that of an HMM-based recognizer. Virtually any type of coefficients can be used – be it spectral coefficients, cepstrum, MFCCs or other (often a combination of more types is chosen).

Fixed-size Inputs – Conventional Approaches

Still, there is a point to note: in HMM-based approaches, signal is generally cut into overlapping frames, for each of which the coefficients are calculated; an observation sequence is thus formed. The difference is that while an HMM will accept an arbitrary number of observation sequences, the input layer of an ANN has a defined number of input neurons which cannot be changed if the whole network is not to be retrained[3]. As signals – even samples of the very same word – will generally have various lengths, a method had to be devised to

---

[2] Although not able to perform ASR in the continuous mode by itself, ANN may present a viable option for continuous mode recognition if combined with HMMs – hence the ANN/HMM hybrids.

[3] There are specific training methods, which involve dynamic changes in structure like adding or removing neurons. However, this is an unrelated problem and the changes are generally confined to the hidden layer.

produce a fixed-size input sequence from these. There are two conventional approaches to this problem[4]:

- Use a static input buffer with a defined number of coefficients long enough for the longest spoken word. Pad briefer words with zeros (several such systems are listed in [2]).
- Select a defined number of windows in the signal (only certain parts of the signal will be processed, for more information see [5]).

Fixed-size Inputs – Average Coefficients Method

We propose yet another approach, which can be summarized as follows:

- Divide the signal into a given number of groups.
- Compute series of coefficients for each group on a per-window basis.
- Calculate average coefficients for each group.

This method combines positive features from both approaches – an input buffer of an arbitrary size[5] can be used – as opposed to restrictions set by method 1, while still taking the whole signal into consideration – as opposed to only selecting certain sections of signal as method 2 suggests.

Main feature thus include:

- fixed-size input,
- all data considered,
- reduced impact of random anomalies,
- final input size may be lower than with simple selection,
- improved duration scalability (vital for systems where duration of the longest and the shortest unit differs considerably).

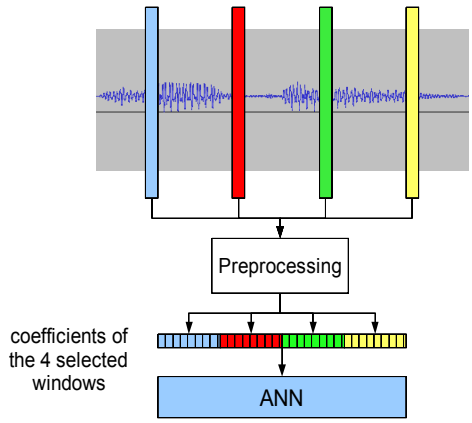

Fig. 5 Simple selection

[4] This problem can as well be solved by using an ANN/HMM hybrid system, where HMMs do the temporal modelling.

[5] With certain minimum size obviously, to ensure that there is at least one window of data for each group.

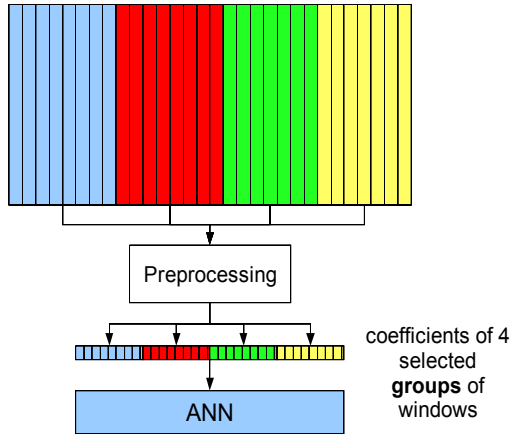

Fig. 6 Coefficient grouping

The main disadvantage is obvious – coefficients for more windows need to be calculated, which results in longer preprocessing times. However, although more time is spent in preprocessing, considerable amounts of time may be saved in the training itself, as – in comparison to simple selection – a lower number of average coefficients is required to achieve comparable accuracy. The number of input neurons may therefore decrease, reducing the complexity of training.

To alleviate time losses within the preprocessing phase itself, the methods may combine (e.g. only every third window in a group will be processed).

3.4. Non-Real-time and Isolated Mode Preprocessing

Several notes might be added considering the non-realtime mode and recognition of isolated words. It is obvious that certain concepts are problematic to implement, or unachievable when in real-time and continuous mode. These include normalization to a defined average, which requires tow-pass processing – all the data to be analysed first to find the peaks and bottoms (first pass) and then the normalization itself can be applied (second pass). That means the whole signal has to be recorded and analysed before it may leave the block of preprocessing.

Another of these methods – energy-based cropping – is only useful in isolated mode as well. Let us now list some more details about this method as well about related experiments.

Energy-based Cropping

As mentioned in [5], accurate end-point detection (and subsequent cropping) increases the accuracy of an ASR system. There are 2 main approaches to detection of start-point and end-point of the useful signal – a method based on energy of the signal and a method based on the rate of zero crossings. Combining both methods leads to good results. However, the processing time increases.

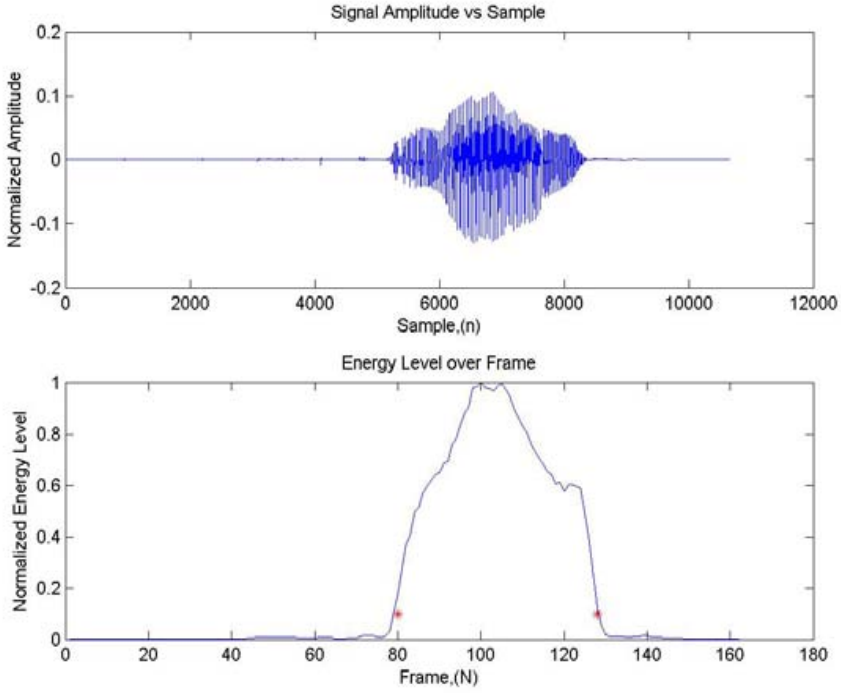

Fig. 7 Energy-based cropping [5]

If only the energy cropping is introduced, the processing time tends to decrease considerably, as the resulting signal is usually much shorter. Figure 7 shows the principle of the algorithm. Signal is divided into sections – frames. For each frame, the energy level is computed. Frames at the beginning and at the end of the signal, with the energy lower than a predefined level are cropped.

Energy of a discrete signal is defined as follows:

$$E = \sum_{i=0}^{N} x_i^2 \ , \tag{3}$$

where x_i is *i*-th sample of the signal and N is the total number of samples.

An experiment has been carried out using an ANN-based recognizer implemented by an author of this paper, which shows that, using no noise-reduction algorithms and simple MFCCs the recognition accuracy may improve by up to 7% using an appropriate cropping level.

3.5. Pattern Recognition

As mentioned, virtually any method of supervised learning can be used to train the network and so the only question is as to what the desired output data for each pattern should be. It is a general practice to structure the network so that the number of outputs is equal to the number of different units that the ASR system is supposed to recognize.

Be the unit a single word, then every word is assigned its own output neuron, the output of which represents the probability[6] that the input pattern has been produced by a sample of that word. The desired output pattern O_i for word i is then obviously:

$$O_i = \begin{pmatrix} o_1 = 0 \\ o_2 = 0 \\ \dots \\ o_{i-1} = 0 \\ o_i = 1 \\ o_{i+1} = 0 \\ \dots \end{pmatrix}, \tag{4}$$

where o_i is the output of $i-th$ neuron.

3.6. Evaluation

If the ANN is trained as stated in the previous section, the evaluation function only needs to find i such that o_i is the maximum element of O. However, it is usually necessary to do some additional checks. For an instance – should value of several outputs be very close to o_i, or should $o_i \ll 1$ this may indicate, that an unknown word has been spoken. The recognizer is generally required to be able to detect unknown words. Such behaviour is especially desirable in the sphere of voice control in automation and similar areas, where false positives or incorrectly recognized commands could trigger potentially dangerous behaviour.

4. ANN/HMM HYBRID SYSTEMS

ANN/HMM hybrid systems combine advantages of both ANNs – as listed above – and HMMs – like the ability to model temporal aspects of speech [6] and perhaps most notably the ability to perform continuous mode recognition.

---

[6] The sum of probabilities is though usually not normalized to 1 as there seems to be no practical reason to do so.

4.1. Types of ANN/HMM Hybrids

The following methods of integrating ANNs and HMMs have been devised [2]:

1) Reimplement various parts of an HMM using ANNs

The main advantage of this approach is that it allows for parallelism. Several instances of this approach are known, like the Viterbi net (introduced by Lippmann and Gold in 1987) and the AlphaNet introduced by Bridle in 1990 [2]. This approach however does not increase accuracy of the system and is as such not especially popular.

2) Frame level training

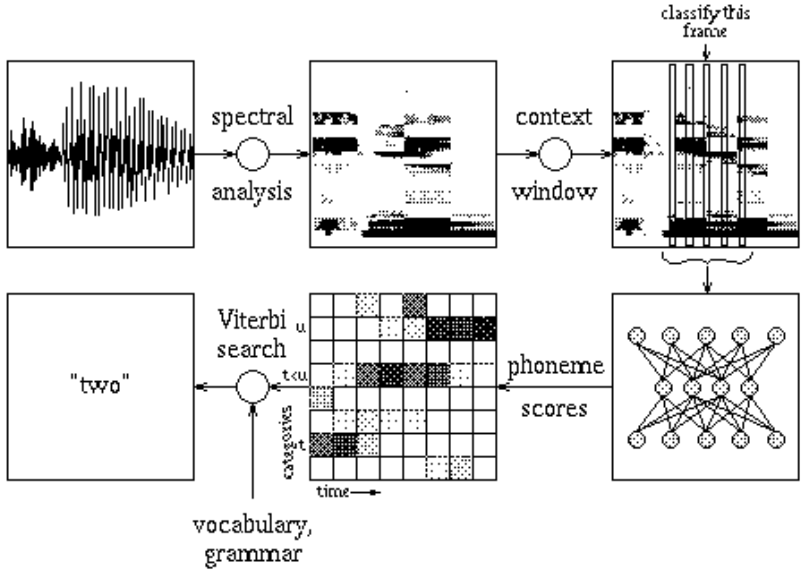

Fig. 8 Overview of the CSLU ASR system [7]

The more typical approach is to capitalize on the strengths of both systems by using an ANN as the acoustic model, while letting an HMM to perform the temporal modelling.

The principle is outlined in figure 8, which shows the CSLU ASR system [7]. An ANN is used to classify each frame of signal (composed of selected windows) providing the emission probabilities for the HMM. Viterbi search is then performed and the most likely unit of speech is found (or series of units, if operating in the continuous mode).

3) Segment level training

The segment level training is based on a similar principle except that whole segments of speech – whole phonemes or even whole words (instead of single frames) – are classified by the ANN, the rest of the process being nearly identical. The drawback of this method is that the speech has first to be divided into segments for the ANN to classify them.

4.2. Training the ANN/HMM Hybrid

The training of an ANN/HMM hybrid system tends to be a little more cumbersome than that of a purely ANN-based one. The training vocabulary typically consists of whole pre-recorded sentences for each of which a transcript is provided (as is the case in HMM-based systems).

It is however as well necessary to provide time-aligned phonetic labels – that is, speech must be – manually or automatically – phonetically segmented, generating markers defining the start and end point of each phoneme [8].

As said, the alignment can be done manually, yet for a larger vocabulary that becomes costly, inconvenient and possibly error-prone. Multiple methods of automatic segmentation exist. For an instance – an existing HMM-based recognizer if provided with a transcript as well as with a recording is able to produce phonetic labels – as the correct state sequence is known, the output of Viterbi search is reduced to the time when transitions between the states happen.

The transcription of speech is an interesting problem in itself. As mentioned in [8], using a standardized international method of transcription can be very helpful, as it might facilitate sharing the acoustic models of phonemes internationally, which is of great interest as creating the training corpus is problematic, expensive and thus not very profitable for smaller countries. This approach would as well be convenient in multi-language systems as each phoneme would be modelled just once for all languages that use it and not for every language separately.

5. NOISE IN ASR SYSTEMS

Let us divide approaches to noise robustness into the following groups:

- inherently robust speech parameters,
- clean speech estimation,
- multi-band ASR.

A brief information on each of these will be listed in the next section (for a more complete description and references see [9]).

5.1. Inherently Robust Speech Parameters

These techniques include the well known Cepstral Mean Normalization, which involves calculation of cepstral features and subtraction of mean value from each cepstral coefficient [9].

Other methods include RASTA-PLP, Short-term Modified Coherence, Linear Discriminant Analysis, Generalised Cepstral Analysis and other.

5.2. Clean Speech Estimation

The clean speech estimation approach makes use of more traditional noise-removal algorithms. The idea is to remove as much noise as possible, before other processing takes place. These methods include Spectral Subtraction, perhaps the simplest technique for clean speech estimation, Probabilistic Optimal Filtering, Code-book Dependent Cepstral Normalisation and other [9].

6. MULTI-BAND ASR

As stated in [10], in noisy speech, automatic recognition can often be improved by simply ignoring the parts of the spectral signal most affected by noise. However, this approach has several drawbacks. Data mismatch might not be detected correctly; useful data might be ignored, etc. Using ANN-based methods, it may as well be problematic to actually ignore a certain band of signal as that would again lead to fixed-size input problem mentioned earlier. Also, the performance for clean speech tends to be unacceptably low when using this method [10].

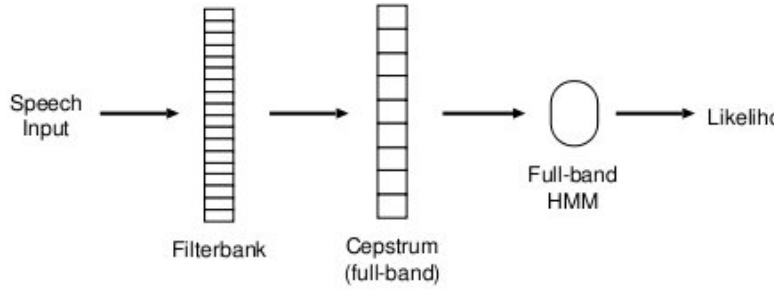

Fig. 9 Full-band recognition [11]

A related principle, called product of errors rule, seems to be more useful. It states that (under certain conditions) in human perception the full-band error rate is equal to the product of sub-band error rates [10]. This leads to the multi-band approach, in which the [11] signal is divided into several parts using a banks of band filters. Each of these bands is then recognized separately and the results are combined.

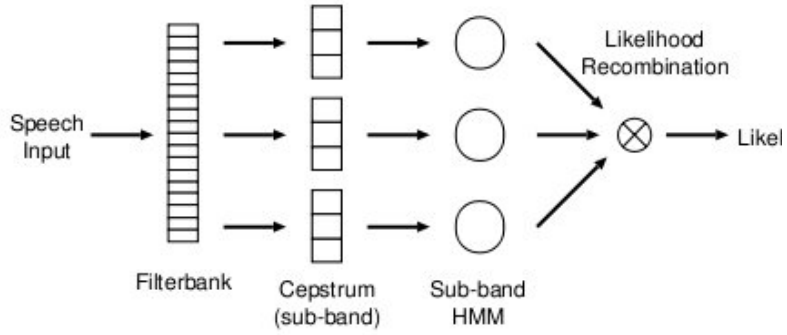

Fig. 10 Multi-band recognition [11]

Both [10] and [11] include further improvements of the method.

7. CONCLUSION

Intelligent Manufacturing Systems (IMS) represent the future of manufacturing. The main target of IMS is to increase manufacturing competitiveness and productivity. Innovative methods in control and organisation of production processes may contribute to the creation of an effective system.

Voice control – and speech recognition in general – is one of such methods. This paper provides an overview of chosen areas from the sphere of speech recognition. The distinction between isolated and continuous recognition mode is considered together with advantages and disadvantages of each approach having voice control in mind.

The basics of ANN-based speech recognition and advantages in comparison to HMM-based solutions are covered together with an alternative solution for the finite-input problem. A brief information about ANN/HMM hybrid systems is provided as well.

The area of robustness and reliability in noisy conditions is of special importance (as it is often vital for successful application of speech recognition in voice control) and is yet to be deeply investigated. A brief list of the best known methods is provided.

Acknowledgements

This paper was supported by the Agency for Support of Research and Development as a part of the project APVV-0597-07.

References

[1] JUHÁR J.: *Spracovanie signálov v systémoch automatického rozpoznávania reči*. Technická univerzita v Košiciach, Habilitačná práca 1999.

[2] TEBELSKIS J.: *Speech Recognition using Neural Networks*. Carnegie Mellon University, Thesis 1995.

[3] TRENTIN E., GORI M.: *A survey of hybrid ANN/HMM models for automatic speech recognition*. In: Neurocomputing, 37(1/4), 2001.

[4] VAN DER SMAGT P., KRÖSE B.: *An Introduction to Neural Networks*. 1996. http://www.avaye.com/files/articles/nnintro/nn_intro.pdf

[5] TAN C. L., JANTAN A.: *Digit Recognition Using Neural Networks*. In: Malaysian Journal of Computer Science, Vol. 17 No. 2, 2004.

[6] GEMELLO R., MANA F., ALBESANO D.: *Hybrid HMM/Neural Network basedSpeech Recognition in Loquendo ASR*. http://www.loquendo.com/en/brochure/Speech_Recognition_ASR.pdf

[7] HOSOM J. P., COLE R., FANTY M.: *Speech Recognition Using Neural Networks*. http://www.cslu.ogi.edu/tutordemos/nnet_recog/recog.html

[8] IVANECKÝ J.: *Automatická transkripcia a segmentácia reči*. Technická univerzita v Košiciach, Fakulta elektrotechniky a informatiky, Dizertačná práca 2003.

[9] GALES M. J. F.: *Model-based Techniques for Noise Robust Speech Recognition*. Gonville and Caius College, University of Cambridge, Dissertation 1995. svr-www.eng.cam.ac.uk/~mjfg/thesis.pdf

[10] MORRIS A. C., HAGEN A., BOURLARD H.: *The Full Combination Sub-Bands Approach To Noise Robust HMM/ANN-based ASR*. In: Proc. Eur. Conf. Speech Commun. Technol., pp. 599-602, 1999.

[11] OKAWA S., BOCCHIERI E., POTAMIANOS A.: *Multi-band Speech Recognition in Noisy Environments*. In: Proc. IEEE Intl. Conf. Acoust., Speech, SignalProcessing, pp. 641–644, 1998.