Models and Analysis of Vocal
Emissions for Biomedical Applications
(MAVEBA 1999)
Firenze, Italy, September 1-3, 1999

THE JOINT SPEECH/VIDEO SIGNAL PROCESSING FOR PERSONS WITH LIMITED PHYSICAL POSSIBILITIES.

E.I.Bovbel, P.D.Kukharchik, I.E.Kheidorov
Dept. of Radiophysics, Belarusian State University,
F.Scoriny av., 4, 220050, Minsk, Belarus
e-mail: ikheidorov@poboxes.com

Abstract. The task of joint speech/video processing is considered. The approach based on two sets of autoregressive hidden Markov models (audio and video models) and neural network is proposed in order to improve the speech recognition performance. The data from each word is processed in two separate channels, and we have two sets of M aposteriory probabilities as the outputs. To combine these results in order to improve the processing accuracy we introduce the direct links neural network. Such technique can be especially useful for persons with limited physical possibilities.

INTRODUCTION

During last years the appearance of multimedia provides wide possibilities for the "man-computer" communication tools improvement. Now there are many different software and hardware tools to increase the possibilities and extend the application area of personal computers. But, as a rule, these tools are for ordinary users and require, firstly, the certain work skills, and secondly, the tools are rather complex and use eyes, ears and hands simultaneously. From the other side only multimedia allows the information conversion from one form to another mostly suitable for the exact user. The design of specialized multimedia systems oriented for the usage of one feeling body at the same time only (eye, ear, hands, etc.) will make the global process of computerization available for men with limited physical possibilities.

Human speech is bimodal both in production and perception. It is produced by the vibration of the vocal cord and the configuration of the vocal tract that is composed of articulatory organs, including: the nasal cavity, the tongue, teeth, velum, and lips. Using these articulatory organs, together with the muscles that generate facial expressions, a speaker produced speech. To perceive speech, an observer listens to the acoustical speech and looks at visible articulatory organs and facial expressions. It has been shown that human perception of acoustical speech can be affected by visual cues from lip movements [1]. That is why it is reasonable to use both speech and visual image of lip movements in order to increase speech recognition accuracy. The joint audio-video processing provides us the possibility to transform information from one form to another.

JOINT AUDIO/VIDEO PROCESSING

The problem of joint systems design can be successfully decided using last achievements at the area of image recognition and synthesis. The image is both speech and picture. High level of modern technologies, in the first instance at the area of high-effective digital signal processing hardware design, allows the decision of the image recognition and synthesis task at the real time scale and with high accuracy [2], [3]. Thus now there are the following trends of multimedia technology for the men with limited physical abilities:

- the usage of speech recognition to reduce the labour-intensity operation for complex way throughout different levels of computer interface (for blind persons, persons with problems of move-support system), and also the usage of speech input of instructions if it is impossible to use alternative communication control methods (for person with problems of move-support system);

-"speech-to-text" transformation (for deaf persons);

-"text-to-speech" transformation (speech synthesis for blind persons, persons with problems of speech producing system);

- speech recognition using the image of lips motions (for persons with problems of speech producing system and move-support system);

- the usage of speech recognition as the most convenient communication tool for the computer distance learning systems.

As a result of large amount of theoretical and experimental researches held during last time it was established that the most effective method of real image recognition systems design is the method based on the hidden Markov models (HMM) and neural networks [4]. Wide application of hidden Markov models can be explain by their powerful and flexible mathematical structure which allows to easily adopts them for different specific tasks of speech and picture recognition. Neural networks let to extend the possibilities and application area of hidden Markov models. Multilayer neural networks can form any decision areas and that is why do not require assumptions for the input parameters distribution form. Great number of calculation units linked by many connections makes neural networks robust for errors and lets to realize parallel signal processing algorithms.

The following system was proposed for joint speech/video processing. For M belarussian spoken words there were trained two sets of autoregressive hidden Markov models [5]. Autoregressive audio hidden Markov models λ_m^{audio}, $m=1,...,M$ were trained based on speech spectrum parameters such as cepstrum. Lip movements for each word pronunciation were chosen as video parameters for autoregressive video hidden Markov models was designed, λ_m^{video}, $m=1,...,M$. The data from each word is processed in two separate channels, and we have two sets of M aposteriory probabilities as the outputs. To combine these results in order to improve the processing accuracy we introduce the direct links neural network. This network uses the $2M$-size vector of audio/video models aposteriory probabilities as an input and classifies it to one of M spoken words.

EXPERIMENTS AND CONCLUSION

Training and recognition algorithms for such joint audio/video processing technique were developed. For isolated belarussian words recognition two sets of HMMs were implemented using samples from specially created data base. The experiments show that the joint adio/video processing system provides much better recognition performance than the separate usage of audio and video HMM. For the task of 200 isolated words recognition such system provides the recognition performance over 98.9% instead 96.5% for separate audio autoregressive hidden Markov model. Once we break down the boundary between speech research and image research we can invent a large number of new techniques and applications including bimodal person identification and verification, security access, image coding, etc. Such technique can be especially useful for persons with limited physical possibilities.

REFERENCES.

1. H.McGurk, and J. MacDonald, "Hearing lips and seeing voices", Nature, pp.746-748, December, 1976.

2. V.G.Atlas, D.G.Atlas, E.I.Bovbel, 2-D arithmetic Fourier transform using Bruns method //IEEE Trans. on Circuits and Systems, vol.44, №6, pp.546-552, 1997.

3. E.I.Bovbel, I.E.Kheidorov, Speech Parameters Vector Based on Arithmetic Fourier Transform //Proc. of XI European Signal Processing Conference EUSIPCO 98. pp.713-717, 1998

4. L. R. Rabiner and B. H. Juang, "Fundamentals of Speech Recognition", PTR Prentice Hall inc., NJ, 1993

5. E.I.Bovbel, P.P.Tkacheva, I.E.Kheidorov, "Autoregressive Hidden Markov Models for Isolated Words Recognition", Proc. of the 2nd IMACS International Conference on: Circuits, Systems and Computers (IMACS-CSC'98). vol.1, pp.453-456, 1998

This paper is partialy supported by the International Science and Technology Center, project B-95.