2000 Neural Networks for Signal Processing X (V, A)

NEURAL NETWORKS FOR SIGNAL PROCESSING X

PROCEEDINGS
OF THE 2000 IEEE
SIGNAL PROCESSING
SOCIETY WORKSHOP

Tenth in a Series of Workshops
Organized by the IEEE Signal Processing Society
Neural Networks Technical Committee

Edited by

Bernard Widrow
Ling Guan
Kuldip Paliwa
Tulay Adali
Jan Larsen
Elizabeth Wilson
Scott Douglas

Published under the sponsorship of the IEEE Signal Processing Society
in cooperation with the IEEE Neural Networks Council
and the Motorola Australian Research Centre

The Institute of Electrical and Electronic Engineers, Inc.
New York, New York

ℓeural Networks for Signal Processing X—Proceedings of the 2000 IEEE ignal Processing Society Workshop

'opyright and Reprints Permission

IEEE Catalog Number	00TH8501
ISBN	0-7803-6278-0
ISSN	1089-3555

lditional copies of this publication are available from

EE Operations Center
O. Box 1331
5 Hoes Lane
›cataway, NJ 08855-1331 USA

800 678 IEEE
732 981 1393
732 981 0600
732 981 9667 (FAX)
ail: customer.service@ieee.org

PREFACE

This book contains refereed papers presented at the Tenth IEEE Workshop on Neural Networks for Signal Processing (NNSP2000) held at University of Sydney, Sydney, Australia, December 11-13, 2000.

The Neural Networks for Signal Processing Technical Committee of the IEEE Signal Processing Society sponsored the NNSP2000 Workshop with co-operation from the IEEE Neural Network Council.

In recent years, the field of neural networks has matured considerably in both methodology and real-world application domains, going far beyond the "standard" backpropagation network and academic examples. Contemporary neural networks for signal processing research shares many ideas of adaptive signal/image processing, machine learning, and advanced statistics. We consider the field of neural networks for signal processing as an open-minded playground that combines many approaches for solving complex real-world signal processing problems.

Continuing the tradition of paperless and easy communication, many of the details of the NNSP2000 Workshop were handled electronically through a re-designed workshop webpage (*http://eivind.imm.dtu.dk/nnsp2000*) that, among other features, included web-based submission, review, and registration.

This year, the workshop features a special session organized by Shigeru Katagiri on Intelligent Multimedia Processing, an increasingly important application area for neural networks.

Our sincere thanks goes to Professor Tamas Roska of the Hungary Academy of Science, Dr. David Fogel of Natural Selection, and Dr. Brian Ferguson of Defence Science and Technology Organisation, Australia for accepting to give plenary speeches at the workshop. Further, we would like to thank Professor Sun-Yuan Kung of Princeton University for agreeing to present a joint keynote speech on "Adaptive Techniques for Intelligent Internet Multimedia Communication" to the attendees of both the NNSP2000 Workshop and the 1st IEEE Pacific-Rim Conference on Multimedia. Also, we would like to thank the Technical Committee for taking the time to provide quality reviews. Special thanks go to the folks at University of Sydney Signal and Multimedia Processing Laboratory for the careful handling of workshop registration and local arrangements. Above all, thanks to all of the authors who made this workshop possible and successful.

Bernard Widrow, *Stanford University*
Ling Guan, *University of Sydney*
Kuldip Paliwa, *Griffith University*
Tulay Adali, *University of Maryland, Baltimore County*
Jan Larsen, *Technical University of Denmark*
Elizabeth Wilson, *Raytheon Company*
Scott Douglas, *Southern Methodist University*

NNSP 2000 Technical Committee

Amir Assadi	University of Wisconsin at Madison, USA
Yianni Attikiouzel	University of Western Australia, Australia
John Asenstorfer	University of South Australia, Australia
Andrew Back	Katestone Scientific, Brisbane, Australia
Geoff Barton	University of Sydney, Australia
Hervé Bourlard	IDIAP, Switzerland
Andy Chalmers	Manukau Polytechnic, New Zealand
Zheru Chi	Hong Kong Polytechnic University, Hong Kong
Andrzej Cichocki	RIKEN, Japan
Tharam Dillon	La Trobe University, Australia
Tom Downs	University of Queensland, Australia
Hsin Chia Fu	National Chiao-Tung University, Taiwan, ROC
Suresh Hangenahally	Griffith University, Australia
Marwan Jabri	University of Sydney, Australia
Haosong Kong	Motorola Australian Research Centre, Australia
Shigeru Katagiri	NTT and ATR, Japan
Anthony Kuh	University of Hawaii at Manoa, USA
Yi Liu	CSIRO Minerals Division, Australia
Fa-Long Luo	ReSound Corporation, USA
David Miller	Pennsylvania State University, USA
Christophe Molina	The Sanger Centre, Cambridge, UK
M Mohammadian	University of Canberra, Australia
Erkki Oja	Helsinki University of Technology, Finland
Soo-Chang Pei	National Taiwan University, Taiwan, ROC
Jose Principe	University of Florida, USA
Ponnuthurai Suganthan	Nanyang Technological University, Singapore
Ah Chung Tsoi	University of Wollongong, Australia
Marc Van Hulle	K.U. Leuven, Belgium
A.N. Venetsanopoulos	University of Toronto, Canada
Yue Wang,	The Catholic University of America, USA
Wilson Wen	Telstra Research Laboratories, Australia

CONTENTS

INTELLIGENT MULTIMEDIA PROCESSING

THEORY AND ALGORITHMS

APPLICATIONS

THEORY AND ALGORITHMS

AUDIO-VISUAL SPEECH RECOGNITION USING MINIMUM CLASSIFICATION ERROR TRAINING

Chiyomi Miyajima, Keiichi Tokuda, and Tadashi Kitamura
Department of Computer Science, Nagoya Institute of Technology
Gokiso-cho, Showa-ku, Nagoya 466-8555, Japan
Phone: +81 52 735 5404 Fax: +81 52 735 5477
E-mail: {chiyomi, tokuda, kitamura}@ics.nitech.ac.jp

Abstract. This paper presents a framework for designing a hidden Markov model (HMM)-based audio-visual automatic speech recognition system based on minimum classification error (MCE) training. Audio/visual HMMs are optimized with MCE training based on the generalized probabilistic descent (GPD) method, and their likelihoods are combined using model-dependent stream weights which are also estimated with the GPD method. Experimental results of speaker independent isolated word recognition show that the GPD optimization of the audio/visual HMMs and the use of GPD-based model-dependent stream weights provide significant improvement in system performance leading to 47% – 81% error reduction over a conventional system which consists of HMMs trained based on the maximum likelihood criterion and globally-tied stream weights estimated with the GPD method.

INTRODUCTION

Human speech perception is enhanced by seeing speaker's lip movements as well as hearing speaker's voice. Consequently, an audio-visual (bimodal) aspect of automatic speech recognition (ASR) has been widely studied and many works have shown that the robustness and accuracy of ASR can be improved by the use of labial information in addition to the acoustic speech signal, especially in noisy conditions [1, 2, 3, 4].

In audio-visual as well as auditory speech recognition, hidden Markov models (HMMs) [5] are most currently used for modeling acoustic and labial features. Such systems consists of HMM-based audio/visual classifiers and integration modules in which the outputs of the HMMs are combined using stream weights. Most widely used HMM training algorithms are based on the maximum likelihood (ML) estimation principle, which aims at approximating the assumed source distribution of each class. The ML estimation, however, is inappropriate for the training of stream weights. Instead, minimum classification error (MCE) training based on the generalized probabilistic descent (GPD) method [6] has been successfully applied to the stream weight estimation [4]. Furthermore, the utility of GPD is not

0-7803-6278-0/00$10.00 (C) IEEE

limited to stream weight estimation. The GPD training aims at maximizing the separation between classes and has been successfully applied to classifier designs in the wide range of pattern recognition domains [7], including HMM parameter training for speech recognition [8].

In this study, audio/visual HMM parameters as well as stream weights are estimated using the GPD method, and their likelihoods are combined using GPD-trained model-dependent stream weights [9]. To evaluate this framework, speaker-independent isolated word recognition experiments are conducted using the M2VTS database [10]. The GPD optimization of HMM parameters provides higher recognition abilities in audio- and visual-only ASR, yielding significant improvement in audio-visual ASR performance. A further significant gain is obtained by using different values of stream weights for each word class.

HMM-BASED AUDIO-VISUAL ASR SYSTEM

There are essentially two approaches to integrating audio and visual information: early integration, in which acoustic and labial feature vectors are combined by synchronizing their frame rates using a frame interpolation technique before being processed in a classifier, and late integration, in which separate classifiers are used for audio and visual data streams and a recognition score is computed by combining their outputs [2]. Our system is based on the latter approach.

As illustrated in Figure 1, each word w ($w = 1, 2, \ldots, W$) is modeled by two separate HMMs $\{\lambda_{\mathrm{A}w}, \lambda_{\mathrm{V}w}\}$, and their likelihoods are combined being weighted by their respective stream weights $\{\gamma_{\mathrm{A}w}, \gamma_{\mathrm{V}w}\}$. Let each spoken word be represented by a set of audio and visual feature observation sequences:

$$\boldsymbol{X} = (\boldsymbol{O}_{\mathrm{A}}, \boldsymbol{O}_{\mathrm{V}}), \tag{1}$$

$$\boldsymbol{O}_{\mathrm{A}} = (\boldsymbol{o}_{\mathrm{A}1}, \boldsymbol{o}_{\mathrm{A}2}, \ldots, \boldsymbol{o}_{\mathrm{A}T_{\mathrm{A}}}), \tag{2}$$

$$\boldsymbol{O}_{\mathrm{V}} = (\boldsymbol{o}_{\mathrm{V}1}, \boldsymbol{o}_{\mathrm{V}2}, \ldots, \boldsymbol{o}_{\mathrm{V}T_{\mathrm{V}}}), \tag{3}$$

with T_{A} and T_{V} being the number of frames in the audio and visual sequences, respectively. We define the audio-visual recognition score of the w-th word model as

$$S_w(\boldsymbol{X}) = P(\boldsymbol{O}_{\mathrm{A}}, \boldsymbol{q}_{\mathrm{A}} \mid \lambda_{\mathrm{A}w})^{\gamma_{\mathrm{A}w}} P(\boldsymbol{O}_{\mathrm{V}}, \boldsymbol{q}_{\mathrm{V}} \mid \lambda_{\mathrm{V}w})^{\gamma_{\mathrm{V}w}}, \tag{4}$$

where $\boldsymbol{q}_{\mathrm{A}}$ and $\boldsymbol{q}_{\mathrm{V}}$ denote the optimal state sequences (Viterbi paths) of $\boldsymbol{O}_{\mathrm{A}}$ and $\boldsymbol{O}_{\mathrm{V}}$ in $\lambda_{\mathrm{A}w}$ and $\lambda_{\mathrm{V}w}$, respectively:

$$\boldsymbol{q}_{\mathrm{A}} = (q_{\mathrm{A}0}, q_{\mathrm{A}1}, \ldots, q_{\mathrm{A}T_{\mathrm{A}}}), \tag{5}$$

$$\boldsymbol{q}_{\mathrm{V}} = (q_{\mathrm{V}0}, q_{\mathrm{V}1}, \ldots, q_{\mathrm{V}T_{\mathrm{V}}}), \tag{6}$$

and $P(\boldsymbol{O}, \boldsymbol{q} \mid \lambda)$ is the likelihood of $\boldsymbol{O}$ along a Viterbi path $\boldsymbol{q}$ in an HMM λ:

$$P(\boldsymbol{O}, \boldsymbol{q} \mid \lambda) = \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(\boldsymbol{o}_t), \tag{7}$$