Available online at www.sciencedirect.com
SCIENCE DIRECT
Computer Speech and Language 20 (2006) 125–127
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

Editorial

1. Introduction

Welcome to the November 2005 special issue of *CSL*. Readers of this journal are well aware of progress in speech recognition, but may not be too familiar with similar progress in speaker and language recognition and their incorporation of speech recognition methods. We invite you to explore this progress through this special issue of *CSL* featuring eight outstanding expanded works from the *Odyssey 2004: The Speaker and Language Recognition Workshop*. This ISCA Tutorial and Research Workshop was held 31 May–3 June 2004 at the Hotel Beatriz in Toledo, Spain. We hope you find these papers interesting and consider attending the upcoming *IEEE Odyssey 2006: The Speaker and Language Recognition Workshop*, 28–30 June 2006 at the Ritz Carlton Hotel, Spa & Casino in San Juan, Puerto Rico (http://www.speakerodyssey.com) and participating in the NIST Evaluations (http://www.nist.gov/speech).

The need for secure and convenient recognition of speakers (verification, identification, segmentation, and clustering) and languages, including dialect and accent, is of growing importance for commercial, forensic, and government applications. The goal is to accomplish this recognition process in a fast, efficient, accurate, and robust way. The primary aim of *Odyssey-04* is to disseminate and further the research needed to achieve these goals. The Odyssey workshop fosters interactions among researchers and experts in speaker and language recognition.

Odyssey-04 is the latest in a series of speaker recognition workshops that started with the *Workshop on Automatic Speaker Recognition, Identification and Verification*, Martigny, Switzerland in April 1994; followed by *Speaker Recognition and Its Commercial and Forensic Applications (RLA2C)*, Avignon, France in April 1998; and by *2001: A Speaker Odyssey, The Speaker Recognition Workshop*, Crete, Greece in June 2001. *Odyssey-04* had 147 attendees, which is an increase of 100% over previous workshops in the series. Likewise, the number of accepted papers presented increased, as did the rejection rate. Sixty-one papers were presented and three expert keynote speakers completed the *Odyssey-04* technical program. These are all healthy signs of scientific growth in speaker and language recognition and interest in *Odyssey*.

Further signs of scientific advancement in automatic speaker and language recognition can be seen in the series of Speaker Recognition Evaluations (SRE) and Language Recognition Evaluations (LRE) coordinated by the National Institute of Standards and Technology (NIST). In these evaluations, participants submit blind results of their systems under primarily text-independent conversational telephone speech conditions. The LRE is intended to establish performance capability for automatic language and dialect recognition and, likewise, for the SRE in automatic speaker recognition. The core SRE speaker detection task trains on one side of one 5-min telephone conversation and tests on another conversation side. Recent extensions to the SRE include working with combined conversation sides (both speakers are mixed together in one file), extended data training conditions (e.g., training on eight 5-min telephone conversations and testing on one), bilingual talkers (e.g., training on a speaker's Spanish call and testing on his/her English call), and multiple microphones (using

0885-2308/$ - see front matter
doi:10.1016/j.csl.2005.09.001

desktop, handsfree cellular, dictation, and distant microphones, in addition to conventional wireline and cellular telephones).[1]

The Odyssey Program Committee deliberated over the eight invitations to authors comprising this special issue from among many fine *Odyssey-04* papers. The primary considerations were the reviewers' ratings of the papers and the quality of the presentations at the *Odyssey-04* workshop. Secondary considerations included spanning the workshop's topics and balancing the authors' affiliations and regions. Some of the papers invited were merges of *Odyssey-04* papers and the authors were invited to collaborate. All 8 papers are significantly expanded from *Odyssey-04* and underwent a rigorous review and revision process by 50 reviewers.

2. In this special issue

The major themes running through the papers in this special issue are speaker and language recognition, speaker segmentation and diarization, evaluation, forensic applications, support vector machines, signal quality, robustness, and performance improvements.

We begin with David van Leeuwen, et al.'s contribution, the ''NIST and NFI-TNO Evaluations of Automatic Speaker Recognition'' on text-independent systems. NIST and NFI-TNO describe their corpora and tasks; compare participants' performance; investigate the effects of speech duration, training handsets, transmission type, language, and accent; and analyze trends across participating sites' systems. NIST's speaker recognition evaluation is the internationally accepted standard for comparing, benchmarking, and measuring progress in speaker recognition technologies using conversational-style telephone speech. The new NFI-TNO evaluation, reported here and patterned after the NIST evaluation, used forensic speech data, which gives us new insights into the performance of automatic speaker recognition technology on this very challenging and important application.

In ''Technical Forensic Speaker Recognition: Evaluation, Types and Testing of Evidence,'' Phil Rose continues in the forensic vein by sharing his intimate first-hand knowledge of forensic aspects of speaker recognition, focusing on evidence, likelihood ratios, evaluating the strength of evidence, and challenges in accurate estimation. The so-called Daubert requirement of testability, closely related to the US Federal Rules of Evidence 702, is also discussed. This paper is an extension of Prof. Rose's and Dr. Meuwly's colorful keynote presentation, ''Forensic Speaker Recognition: An Evidence Odyssey.''

In ''Using Quality Measures for Multilevel Speaker Recognition,'' Daniel García-Romero, et al. report on their use of signal quality measures in their scoring and multilevel fusion stages to improve speaker recognition robustness and performance. The signal quality measures used are SNR, F_0 deviations, and the ITU P.563 objective speech quality assessment. Scores are fused using a multilevel Support Vector Machine (SVM) adapted to include signal quality information. Experimental results are provided using the Switchboard-I corpus.

William Campbell, et al. present MIT Lincoln Laboratory's ''Support Vector Machines for Speaker and Language Recognition.'' SVMs are applied to both speaker and language recognition and shown to achieve state-of-the-art performance through the introduction of a kernel that compares sequences of feature vectors and produces a measure of similarity based upon generalized linear discriminants. Furthermore, SVMs are shown to be complementary to Gaussian mixture model (GMM) methods by fusing the two together to improve performance. This SVM method and its fusion with GMMs are demonstrated in the NIST speaker and language recognition evaluations.

Niko Brümmer and Johan du Preez present their work on ''Application-Independent Evaluation of Speaker Detection.'' They propose an alternative to the traditional error- or cost-based evaluation metrics to assess speaker detection performance based on an information-theoretic metric. An expected cost or a total error-rate is formed over a range of different application types. The method is demonstrated by evaluating three speaker detection systems submitted to the NIST 2004 Speaker Recognition Evaluation.

[1] J. P. Campbell, H. Nakasone, C. Cieri, D. Miller, K. Walker, A. F. Martin and M. A. Przybocki, *The MMSR Bilingual and Crosschannel Corpora for Speaker Recognition Research and Evaluation*, *Odyssey: The Speaker and Language Recognition Workshop*, ISCA, Toledo, Spain, 2004, pp. 29–32. Available from: <http://www.isca-speech.org/archive/odyssey_04/ody4_029.html>.

In "A Syllable-Scale Framework for Language Identification," Terrence Martin, et al. investigate a syllable-like unit automatic language recognition system. Phone triplets are used to approximate syllable-length sub-word segmental units. The authors examine the contributions made by acoustic, phonotactic, and prosodic information sources in accordance with the NIST 1996 LID protocol. They find extended phonotactic information can be used to complement and improve the performance of the popular Parallel Phone Recognition Language Modelling (PPRLM) technique.

In "Step-by-Step and Integrated Approaches in Broadcast News Speaker Segmentation," Sylvain Meignier, et al. shifts from conversational telephone speech to broadcast news program audio and from speaker/language detection to speaker diarization (segmenting a conversation into homogeneous segments which are then grouped into speaker classes). The authors compare and describe two approaches for speaker diarization and investigate various strategies for the fusion of diarization results. Results obtained since 2002 on speaker diarization for various corpora are presented.

Concluding our special issue with "Robust Estimation, Interpretation and Assessment of Likelihood Ratios in Forensic Speaker Recognition," Joaquín González-Rodríguez, et al. present the Bayesian framework for interpretation of evidence when applied to forensic speaker recognition. Using voice as evidence in court is addressed, as well the forensic expert's use of the likelihood ratio to express the strength of the evidence. Robust methods of likelihood ratio estimation are proposed. These algorithms were assessed on the Switchboard corpora and in the NFI-TNO 2003 Forensic SRE and NIST 2004 SRE.

3. Looking ahead

Speaker and language recognition technology is currently viable in many applications. With future improvements, this technology will be practical for more unconstrained and uncontrolled situations. One of our greatest challenges is the frailty of our recognizers with respect to variable channels (including microphones/handsets). By increasing robustness to channel and speaker variabilities (e.g., by incorporating higher level information and by improving fusion and scoring), we will allow even more successful applications of automatic speaker and language recognition technologies.

We are especially grateful to the 30 authors and 50 reviewers who contributed to the eight papers in this special issue. Their tireless community effort has produced this outstanding issue in record time and before the next Odyssey workshop, which is planned for June 2006 and where we hope to see you!

¡gracias, amigos!

Guest Editor
Joseph P. Campbell
MIT Lincoln Laboratory
E-mail address: j.campbell@ieee.org

Guest Editor
John Mason
University of Swansea
E-mail address: J.S.D.Mason@swansea.ac.uk

Guest Editor & Chair Odyssey '04
Javier Ortega-Garcia
Universidad Autónoma de Madrid
E-mail address: javier.ortega@uam.es

Available online 3 October 2005