Dynamic Language Modeling for European Portuguese

***Author**: Ciro Martins*+ **Supervisors**: António Teixeira*, João Neto+*

[*]Department Electronics, Telecommunications & Informatics/IEETA – Aveiro University, Portugal
[+]L2F – Spoken Language Systems Lab – INESC-ID/IST, Lisbon, Portugal
Ciro.Martins@l2f.inesc-id.pt, ajst@det.ua.pt, Joao.Neto@inesc-id.pt

1. Committee

- Manuel Augusto Marques da Silva - Professor of Aveiro University (Portugal).
- Tanja Schultz - Professor of Karlsruhe University (Germany) and Assistant Research Professor of Language Technologies Institute (LTI) in School of Computer Science – CMU.
- Isabel Maria Martins Trancoso - Professor of Departamento de Engenharia Electrotécnica e de Computadores do Instituto Superior Técnico da Universidade Técnica Lisboa (Portugal).
- Francisco António Cardoso Vaz - Professor of Aveiro University (Portugal).
- João Paulo da Silva Neto - Professor of Departamento de Engenharia Electrotécnica e de Computadores do Instituto Superior Técnico da Universidade Técnica Lisboa (Portugal).
- António Joaquim da Silva Teixeira - Professor of Aveiro University (Portugal).

Qualification: Approved by unanimity

2. Abstract

Most of today methods for transcription and indexation of broadcast audio data are manual. Broadcasters process thousands hours of audio and video data on a daily basis, in order to transcribe that data, to extract semantic information, and to interpret and summarize the content of those documents. The development of automatic and efficient support for these manual tasks has been a great challenge and over the last decade there has been a growing interest in the usage of automatic speech recognition as a tool to provide automatic transcription and indexation of broadcast news and random and relevant access to large broadcast news databases. However, due to the common topic changing over time which characterizes this kind of tasks, the appearance of new events leads to high out-of-vocabulary (OOV) word rates and consequently to degradation of recognition performance. This is especially true for highly inflected languages like the European Portuguese language.

Several innovative techniques can be exploited to reduce those errors. The use of news shows specific information, such as topic-based lexicons, pivot working script, and other sources such as the online written news daily available in the Internet can be added to the information sources employed by the automatic speech recognizer. In this thesis we are exploring the use of additional sources of information for vocabulary optimization and language model adaptation of a European Portuguese broadcast news transcription system.

Hence, this thesis had 3 different main contributions: a novel approach for vocabulary selection using Part-Of-Speech (POS) tags to compensate for word usage differences across the various training corpora; language model adaptation frameworks performed on a daily basis for single-stage and multistage recognition approaches; a new method for inclusion of new words in the system vocabulary without the need of additional data or language model retraining.

3. Curriculum Vitae

3.1. Personal Details

Ciro Alexandre Domingues Martins
Rua dos Cabecinhos – Fráguas
3850-707 Ribeira de Fráguas
Phone: +351 965365208
E-mail: Ciro.Martins@gmail.com

3.2. Education

- Ph.D. in Informatics Engineering, Aveiro University, 2008.
- M. Sc. in Electrotecnic and Computers Engineering, Instituto Superior Técnico (IST), Technical University of Lisbon, 1998.
- Graduated in Mathematics and Informatics, Universidade da Beira Interior, 1993.

3.3. Teaching

- Instituto de Entre Douro e Vouga - Portugal, Assistant Professor, since 2006.
- Universidade Católica Portuguesa – Portugal, Assistant Professor, 1999-2007.

3.4. Publications

Martins, C., Teixeira, A. and Neto, J. (2008). "Automatic Estimation of Language Model parameters for unseen Words using Morpho-syntactic Contextual Information", in Proceedings of INTERSPEECH 2008, Brisbane, Australia, 2008.

Martins, C., Teixeira, A. and Neto, J. (2008). "Dynamic Language Modeling for the European Portuguese", in Proceedings of PROPOR 2008, Curia, Portugal, 2008.

Martins, C., Teixeira, A. and Neto, J. (2007). "Dynamic Vocabulary Adaptation for a daily Broadcast News Transcription System", in Proceedings of 2007 IEEE Automatic Speech Recognition and Understanding Workshop, Kyoto, Japan, 2007.

Martins, C., Teixeira, A. and Neto, J. (2007). "Vocabulary Selection for a Broadcast News Transcription System using a Morpho-syntactic Approach", in Proceedings of INTERSPEECH 2007, Antwerp, Belgium, 2007.

Martins, C., Teixeira, A. and Neto, J. (2006). "Dynamic Vocabulary Adaptation for a daily and real-time Broadcast News

Transcription System", in Proceedings of IEEE/ACL 2006 Workshop on Spoken Language Technology, Aruba, 2006.

Martins, C., Teixeira, A. and Neto, J. (2005). "Language Models in Automatic Speech Recognition", revista do Departamento de Electrónica e Telecomunicações da Universidade de Aveiro, 2005.

Martins, C., Neto, J. and Almeida, L. (1999). "Using Partial Morphological Analysis in Language Modeling Estimation for Large Vocabulary Portuguese Speech Recognition", in Proceedings of EUROSPEECH 99, Budapest, Hungary, 1999.

Martins, C. (1998). "Modelos de Linguagem no Reconhecimento de Fala Contínua", Master Thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisbon, Portugal, 1998.

Martins, C., Mascarenhas, M., Meinedo, H., Neto, J., Oliveira, L., Ribeiro, C., Trancoso, I. and Viana, M. (1998). "Spoken Language Corpora for Speech Recognition and Synthesis in European Portuguese", in Proceedings of RECPAD 98, Associação Portuguesa de Reconhecimento de Padrões, Lisbon, Portugal, 1998.

Neto, J., Martins, C. and Almeida, L. (1997). "The Development of a Speaker Independent Continuous Speech Recognizer for Portuguese", in Proceedings of EUROSPEECH 97, Rhodes, Greece, 1997.

Martins, C., Rodrigues, F. and Rodrigues, R. (1997). "An Isolated Letter Recognizer for Proper Name Identification Over the Telephone", in Proceedings of RECPAD 97, Associação Portuguesa de Reconhecimento de Padrões, Coimbra, Portugal, 1997.

Neto, J., Almeida, L., Hochberg, M., Martins, C., Nunes, L., Renals, S. and Robinson, T. (1995). "Speaker-Adaptation for Hybrid HMM-ANN Continuous Speech Recognition System", in Proceedings of EUROSPEECH 95, Madrid, Espanha, 1995.