Language Identification from Speech

Sam Sucik

MInf Project (Part 1) Report

Master of Informatics
School of Informatics
University of Edinburgh

2019

Abstract

This is an example of `infthesis` style. The file `skeleton.tex` generates this document and can be used to get a "skeleton" for your thesis. The abstract should summarise your report and fit in the space on the first page. You may, of course, use any other software to write your report, as long as you follow the same style. That means: producing a title page as given here, and including a table of contents and bibliography.

Acknowledgements

Acknowledgements go here.

Table of Contents

Chapter 1

Introduction

The document structure should include:

- The title page in the format used above.
- An optional acknowledgements page.
- The table of contents.
- The report text divided into chapters as appropriate.
- The bibliography.

Commands for generating the title page appear in the skeleton file and are self explanatory. The file also includes commands to choose your report type (project report, thesis or dissertation) and degree. These will be placed in the appropriate place in the title page.

The default behaviour of the documentclass is to produce documents typeset in 12 point. Regardless of the formatting system you use, it is recommended that you submit your thesis printed (or copied) double sided.

The report should be printed single-spaced. It should be 30 to 60 pages long, and preferably no shorter than 20 pages. Appendices are in addition to this and you should place detail here which may be too much or not strictly necessary when reading the relevant section.

1.1 Using Sections

Divide your chapters into sub-parts as appropriate.

1.2 Citations

Note that citations (like **?** or **?**) can be generated using `BibTeX` or by using the

`thebibliography` environment. This makes sure that the table of contents includes an entry for the bibliography. Of course you may use any other method as well.

1.3 Options

There are various documentclass options, see the documentation. Here we are using an option (`bsc` or `minf`) to choose the degree type, plus:

- `frontabs` (recommended) to put the abstract on the front page;
- `twoside` (recommended) to format for two-sided printing, with each chapter starting on a right-hand page;
- `singlespacing` (required) for single-spaced formating; and
- `parskip` (a matter of taste) which alters the paragraph formatting so that paragraphs are separated by a vertical space, and there is no indentation at the start of each paragraph.

Chapter 2

The Real Thing

Of course you may want to use several chapters and much more text than here.

Chapter 3

Prosody for LID

3.1 Features used in literature

Various features used in the past in LID experiments:

1. Shifted delta cepstra (SDC) by Ferrer et al. (2016) and Sarma et al. (2018) (7D MFCC + 7-1-3-7 SDCs = 56D) and by Torres-Carrasquillo et al. (2002) (10-1-3-3),
2. 19 MFCCs + energy + 20 Δ + 20 $\Delta\Delta$ = 60D by Sarma et al. (2018)
3. 39D MFCC vectors combined with 4D pitch features (Song et al., 2013)
4. 39 PLP features (including Δ and $\Delta\Delta$) (Lopez-Moreno et al., 2014)
5. Lin and Wang (2005) do LID just from the pitch contour parametrised by Legendre polynomials
6. Ghahremani et al. (2014) show ASR improvements with a pitch-tracking algorithm that calculates pitch even for unvoiced frames

3.2 Experiments with Prosodic Features

MFCC fed into DNN SDC fed into DNN MFCC/SDC and KaldiPitch concatenatedfed and fed into X-vector DNN Two X-vector DNNs trained separately on MFCC/SDC and KaldiPitch - evaluated separately - the two X-vectors concatenated and classified and evaluated as a whole

Bibliography

Ferrer, L., Lei, Y., McLaren, M., and Scheffer, N. (2016). Study of senone-based deep neural network approaches for spoken language recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24:105–116.

Ghahremani, P., BabaAli, B., Povey, D., Riedhammer, K., Trmal, J., and Khudanpur, S. (2014). A pitch extraction algorithm tuned for automatic speech recognition. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2494–2498.

Lin, C.-Y. and Wang, H.-C. (2005). Language identification using pitch contour information. *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, 1:I/601–I/604 Vol. 1.

Lopez-Moreno, I., Gonzalez-Dominguez, J., Plchot, O., Martinez, D., González-Rodríguez, J., and Moreno, P. J. (2014). Automatic language identification using deep neural networks. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5337–5341.

Sarma, M., Sarma, K. K., and Goel, N. K. (2018). Language recognition using time delay deep neural network. *CoRR*, abs/1804.05000.

Song, Y., Jiang, B., Bao, Y., Wei, S., and Dai, L. (2013). I-vector representation based on bottleneck features for language identification. *Electronics Letters*, 49(24):1569–1570.

Torres-Carrasquillo, P. A., Singer, E., Kohler, M. A., Greene, R. J., Reynolds, D. A., and Deller, J. R. (2002). Approaches to language identification using gaussian mixture models and shifted delta cepstral features. In *INTERSPEECH*.