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

Thanks to Paul Moore for productive collaboration while building the baseline system, to Steve Renals for his guidance, feedback and optimism, and to David Snyder for his advice.

Table of Contents

1 Introduction **7**

2 Building the System **9**

2.1 Choice of classifier 9

2.2 Setting Hyperparameters 9

3 Prosody for LID **11**

3.1 Features used in literature 11

3.2 Experiments with Prosodic Features 11

Bibliography **13**

Chapter 1

Introduction

Chapter 2

Building the System

System was built in Kaldi (Povey et al., 2011). Reproducing the architecture described by Snyder et al. in LID setting (Snyder et al., 2018a). The authors adapted the architecture from the speaker recognition setting (Snyder et al., 2018b), which is accessible as the SRE16 recipe in Kaldi. We used that recipe and the GlobalPhone Kaldi recipe.

2.1 Choice of classifier

The SRE16 recipe uses PLDA because it does not do identification, but verification. For our purposes, we needed a proper classifier. Current popular and well-performing classifiers are various flavours of GMM and logistic regression, with no clear winner. Snyder et al. (2018a), for instance, used GMM trained using MMI – based on McCree (2014). We decided to re-use a model which is already implemented in the LRE07/v2 recipe – logistic regression. Our decision was also consulted with Snyder (r 24).

2.2 Setting Hyperparameters

1. MFCC extraction parameters
2. DNN architecture: Layers, activation functions, choice of extraction layer, ...
3. DNN training: Learning algorithm and its parameters, number of epochs, ...
4.

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) – outperformed by SDCs, but note that they trained an ASR TDNN for generating i-vectors
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.

McCree, A. (2014). Multiclass discriminative training of i-vector language recognition. In *Proc. Odyssey*, pages 166–172.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

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

Snyder, D. (2018, December 24). Language identification with x-vectors: Choice of classifier [online forum comment]. `https://groups.google.com/forum/#!topic/kaldi-help/v6Uh7avv-cY`.

Snyder, D., Garcia-Romero, D., McCree, A., Sell, G., Povey, D., and Khudanpur, S. (2018a). Spoken language recognition using x-vectors.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (2018b). X-vectors: Robust dnn embeddings for speaker recognition. *2018 IEEE International*

Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5329–5333.

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*.