Prosodic features in state-of-the-art spoken language identification

Sam Sucik

MInf Project (Part 1) Report
Master of Informatics
School of Informatics
University of Edinburgh

2019

Abstract

TO-DO

Acknowledgements

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

Table of Contents

Chapter 1

Introduction

1.1 Motivation

LID is useful in ASR, in voice assistants, emergency call routing, etc. Traditionally, acoustic features are used (influence of ASR on LID and SID). Prosodic LID is much rarer, although results show that prosodic information can help identify language (Lin and Wang, 2005), and that both LID and ASR can benefit from using acoustic *and* prosodic features (González et al., 2013; Ghahremani et al., 2014). Just last year, a novel architecture for LID was proposed by Snyder et al. (2018), dramatically improving the state-of-the-art results. Although the authors find that using bottleneck features from an ASR DNN yields better results than using the standard acoustic MFCC features, even the ASR DNN was trained just using MFCCs. Thus the work ignores the potential of speech information other than that captured by MFCCs.

1.2 Aims

In this work, I aim to reproduce the state-of-the-art LID system and explore the use of prosodic features in addition to acoustic ones. Because the system uses a relatively novel architecture, in which a TDNN aggregates information across a speech segment, I also compare two types of acoustic features, one which has such aggregation over time encoded (SDC) and one that only containes information about single frames (vanilla MFCC).

1.3 Contributions

what exactly I did

1.4 Overview

structure of the report

Chapter 2

Background

Definitions (interespersed, not having a separate section) LID, identification vs verification acoustic (MFCC, SDC) vs prosodic (intonation, stress, rhythm) LID other definitions added later as needed

2.1 Spoken language identification

2.2 Traditional approaches

GMMs and i-vectors

2.3 State of the art

TDNNs and x-vectors

2.4 Features used for LID

Acoustic MFCCs (vanilla and with delta terms) SDCs Prosodic F0 (intonation), and KaldiPitch Energy (stress) Duration (rhythm)

Chapter 3

Data

Intro: Popular datasets

3.1 GlobalPhone

3.2 Data partitioning

explanation of the use of training/enrollment/evaluation/testing data

3.3 Data preprocessing

SHN to WAV splitting long utterances into shorter ones

Chapter 4

Implementation

4.1 The Kaldi toolkit

4.2 Adapted implementations

Snyder et al.'s LID x-vector paper GlobalPhone recipe SRE16 recipe

4.3 Final architecture

Description of x-vector+GP architecture (highlighting own contributions) how the whole pipeline works: go into much more detail than in the Related work section mention possibility of direct classification with TDNN and that I focus on using independent classifier because it provides extra flexibility

4.4 Computing environment

cluster, Slurm, GPUs, parallelisation, rough runtimes of the different stages

4.5 Hyperparameters

tuned on the baseline (see next section) decided: TDNN layers, activation function, learning algorithm (TO-DO: understand shrinking), log-regression parameters empirically established: number of TDNN training epochs (also mention training time)

Chapter 5

Experiments

Intro: I will compare acoustic and prosodic features, leaving formants and BNF for later. Evaluation metric: $C_primary$

5.1 Baseline

vanilla MFCCs (expand on MFCC configuration) 30s enrollment segments, 10s eval/test segments

5.2 SDC vectors

wanna compare with MFCCs

5.3 KaldiPitch+Energy vectors

5.4 MFCC/SDC + KaldiPitch+Energy

feature vectors concatenated, wanna see if they are complementary

5.5 Possibly fusion of MFCC/SDC and KaldiPitch+Energy scores

fucion of results instead of feature vectors using evaluation data for training the fusion logistic regression

Chapter 6

Results

Reported: overall $C_primary$, language-specific score, accuracy (for illustrative purposes), confusion matrix (to see which language pairs are confusing) Focus on Slavic languages (Czech, Croatian, Polish, Russian, Bulgarian)

Chapter 7

Discussion

Chapter 8

Future work

8.1 Plans for Part 2 of the MInf project

8.2 Other future ideas

Chapter 9

Conclusions

Bibliography

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.

González, D. M., Lleida, E., Ortega, A., and Miguel, A. (2013). Prosodic features and formant modeling for an ivector-based language recognition system. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6847–6851.

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.

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