Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 41 (2017) 88–115

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Parametric representation of excitation source information for language identification

Dipanjan Nandi [a,*], Debadatta Pati [b], K. Sreenivasa Rao [a]

[a] School of Information Technology, Indian Institute of Technology, Kharagpur 721302, West Bengal, India
[b] Department of Electronics and Communication Engineering, National Institute of Technology, Nagaland 797103, India

Received 11 August 2015; received in revised form 5 May 2016; accepted 12 May 2016
Available online 16 June 2016

Abstract

In this work, the linear prediction (LP) residual signal has been parameterized to capture the excitation source information for language identification (LID) study. LP residual signal has been processed at three different levels: sub-segmental, segmental and supra-segmental levels to demonstrate different aspects of language-specific excitation source information. Proposed excitation source features have been evaluated on 27 Indian languages from Indian Institute of Technology Kharagpur-Multi Lingual Indian Language Speech Corpus (IITKGP-MLILSC), Oregon Graduate Institute Multi-Language Telephone-based Speech (OGI-MLTS) and National Institute of Standards and Technology Language Recognition Evaluation (NIST LRE) 2011 corpora. LID systems were developed using Gaussian mixture model (GMM) and *i*-vector based approaches. Experimental results have shown that segmental level parametric features provide better identification accuracy (62%), compared to sub-segmental (40%) and supra-segmental level (34%) features. Excitation source features obtained from three levels show distinct language-specific evidence. Therefore, the scores from all three levels are combined to obtain the complete excitation source information for the LID task. LID performances achieved from both the excitation source and vocal tract system are compared. Finally, the scores obtained by processing the vocal tract and excitation source features are combined to achieve better improvement in LID accuracy. The best recognition accuracies obtained from stage-IV integrated LID systems I, II and III are 69%, 70% and 72% respectively.

Keywords: Excitation source information; Language identification (LID); LP residual; GFD parameters; RMFCC; MPDSS; Pitch contour; Epoch strength contour; MFCC; *i*-Vector; IITKGP-MLILSC; OGI-MLTS; NIST LRE; C_{avg}

1. Introduction

Human speech is intended to convey messages. Speech signal carries not only the message information but also the information about the speaker, language and emotion. The primary objective of a language identification (LID) task is to determine the identity of language from the uttered speech. Due to several real-life applications of automatic LID systems such as speech to speech translation systems, information retrieval from multilingual audio databases and multilingual speech recognition systems, it has become an active research problem. Indian languages belong to several language groups and sub-groups. The major two language groups are the Indo-Aryan languages spoken by 76.86% of Indian citizens and the Dravidian languages spoken by 20.82% Indians (Vanishree, 2011). Most of the

* Corresponding author at: School of Information Technology, Indian Institute of Technology, Kharagpur 721302, West Bengal, India. Tel.: +91-3222-282336; fax: +91-3222-282206.
E-mail address: dipanjanconnect.08@gmail.com (D. Nandi).

http://dx.doi.org/10.1016/j.csl.2016.05.001
0885-2308/

languages in India have a common set of phonemes and also follow similar grammatical structure. To develop a language identification system, it is necessary to derive non-overlapping language-specific information for each language. Therefore, building an automatic LID system in the Indian context is quite a challenging task.

Human speech production system has two major components: vocal tract system and source of excitation. The constriction caused by expiration of air acts as an excitation source during the production of speech. The quasi-periodic air pulses generated by the vocal folds vibration act as the primary source of excitation to vocal tract resonator during voiced speech production. During the production of unvoiced speech, the expiration of air is constrained either completely (e.g. unvoiced stops) or partially (e.g. fricatives). In speech production, the majority of excitation takes place during the production of voiced speech. This excitation source information can be captured by passing the speech signal through the inverse filter (Makhoul, 1975). We have used 10th order linear prediction (LP) analysis followed by inverse filtering the speech signal (sampled at 8 kHz) for estimating LP residual signal. Both the vocal tract system and excitation source have significant contribution in the production of voiced speech. Most of the contemporary works exploited the vocal tract system characteristics to determine the language-specific cues from the speech signal. The dynamics of vocal tract system are captured by spectral analysis of speech, which can be represented by Mel-frequency cepstral coefficients (MFCCs) (Balleda et al., 2000), linear prediction coefficients (LPCs) and linear prediction cepstral coefficients (LPCCs) (Sugiyama, 1991). Language related prosodic features extracted from syllable, word and sentence levels have also been explored in recent works (Reddy et al., 2013). However, the source characteristics have not been investigated to obtain language-specific information present in a speech signal.

In this work, the excitation source information has been studied for developing automatic LID system. Excitation source features are extracted from the linear prediction residual (LPR) signal (Makhoul, 1975). In the present study, LP residual signal has been processed at three different levels: within a glottal cycle known as sub-segmental (*sub*) level information, within 2–3 glottal cycles known as segmental (*seg*) level information and across 50 glottal cycles known as supra-segmental (*supra*) level information. The glottal flow derivative (GFD) parameters are extracted from LP residual signal to capture the *sub* level excitation source information. The energy and periodicity of excitation source are captured by parameterizing the LPR signal at *seg* level. The pitch and epoch strength contour information are obtained by processing the LPR signal at *supra* level. The excitation source parameters derived from different levels may capture some non-overlapping language-specific information. Therefore, scores from different levels are combined to obtain the complete parametric excitation source information. The LID accuracies achieved by proposed excitation source features are also compared with the LID performance obtained by processing vocal tract information represented by MFCCs. The vocal tract system and source for exciting the vocal tract are two different components of speech production system. Hence, non-overlapping language-specific information may be present in these two components. Therefore, the combination of these two information sources may provide improved LID accuracy. For developing the systems GMM and *i*-vector based approaches are explored. It is observed that *i*-vector based systems work better, compared to GMM based system. The significance of excitation source features is also examined by investigating five different dialects of Hindi language. Excitation source features are also evaluated on OGI-MLTS and NIST LRE 2011 corpora.

The rest of the paper is organized as follows: In Section 2, earlier works on LID are discussed. Section 3 presents the motivation for the present work. A brief description of the language databases used in this work has been provided in Section 4. Proposed parametric representation of excitation source is described in Section 5. In Section 6, experimental setup and methodology have been explained. Description of LID systems developed using the proposed features is laid out in Section 7. Evaluation of LID systems using proposed excitation source features is given in Section 8. Section 9 concludes the present work.

2. Previous works

This section presents a brief overview of existing works on language identification systems. Sugiyama (Sugiyama, 1991) has explored linear prediction coefficients (LPCs) and cepstral coefficients (LPCCs) for language recognition. Morgan et al. (1992) and Zissman (Zissman, 1993) have proposed the Gaussian mixture models (GMMs) (Reynolds and Rose, 1995) for language identification study. In the Indian context, Balleda et al. (2000) have first attempted to identify Indian languages automatically using speech signal. In their work, vector quantization (VQ) has been used for classification purpose and MFCC features have been used to represent the language-specific information. Vector quantization is used to represent large number of multi-dimensional feature vectors into few representative vectors